

Christian Homburg
Martin Klarmann
Arnd Vomberg
Editors

Handbook of Market Research

Handbook of Market Research

Christian Homburg • Martin Klarmann •
Arnd Vomberg
Editors

Handbook of Market Research

With 211 Figures and 130 Tables

 Springer

Editors

Christian Homburg
Department of Business-to-Business
Marketing, Sales, and Pricing
University of Mannheim
Mannheim, Germany

Martin Klarmann
Department of Marketing & Sales Research
Group
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

Arnd Vomberg
Marketing & Sales Department
University of Mannheim
Mannheim, Germany

ISBN 978-3-319-57411-0 ISBN 978-3-319-57413-4 (eBook)
ISBN 978-3-319-57412-7 (print and electronic bundle)
<https://doi.org/10.1007/978-3-319-57413-4>

© Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Already in 2015, the *Wall Street Journal* claimed that companies sit on a treasure trove of market data. They have an ever-increasing amount of data at their disposal. However, it is not only about access to data. Companies need to develop strong empirical and analytical skills to turn their data into a competitive advantage. Traditional market research firms and hundreds of new startup companies specializing in “Data Science” and analytics support companies in building and maintaining customer relationships, developing strategies to increase customer satisfaction, improving sales strategies, personalizing the marketing mix, and automating marketing processes in real time. The *Handbook of Market Research* seeks to provide material for both, firms specialized in data analysis and firms hiring those firms. On the one hand, it seeks to provide in-depth coverage of established and new marketing research methods. On the other hand, by giving examples throughout, it aims to be as accessible as possible.

The *Handbook of Market Research* helps readers apply advanced market research methods in their projects and provides them with a valuable overview of various analytical techniques. It targets three groups: graduate students, scholars, and data science practitioners. Graduate students obtain an introduction to diverse market research topics. Scholars can use the handbook as a reference, supporting their research and teaching. Practitioners receive a state-of-the-art overview of scientific practices.

What is special about the *Handbook of Market Research*?

- Chapters in this handbook are not purely technical but also offer an intuitive account of the discussed methodologies.
- Many chapters provide data and software code to replicate the analyses. Readers can find such supplementary material on the handbook’s online site (<https://link.springer.com/referencework/10.1007/978-3-319-05542-8>).
- Nearly all chapters in this handbook have gone through a friendly review process. The friendly reviewers helped to improve all chapters of this handbook further.
- We publish the handbook dynamically. Novel chapters will appear continuously on the handbook’s online site. Moreover, authors have the opportunity to update existing chapters online to respond to emerging trends and new methods.

The handbook has three parts: Data, Methods, and Applications. The Data part supports readers in collecting and handling different types of data. The Method part outlines how readers can analyze structured and unstructured data. The Application part equips readers with knowledge on how they can use data analytics in specific contexts.

Our special thanks go to the authors of the chapters for their willingness to share their knowledge and experience with the readers. Furthermore, we would like to take this opportunity to thank the friendly reviewers who have helped further to increase the high quality of the individual contributions. We want to thank Dr. Prashanth Mahagaonkar, Veronika Mang, and Barbara Wolf from Springer Verlag for their excellent cooperation.

Mannheim, Germany
Karlsruhe, Germany
Germany
November 2021

Christian Homburg
Martin Klarmann
Arnd Vomberg

List of Reviewers

Last Name	First Name	Title	Description	Institution
Artz	Martin	Prof. Dr.	Professor for Management Accounting and Control	University of Münster
Atalay	Selin	Prof. Dr.	Professor of Marketing	Frankfurt School of Finance and Management
Becker	Jan-Michael	Dr.	Associate Professor at the Marketing Department	BI Norwegian Business School
Bhattacharya	Abhimanyu	Ph.D.	Assistant Professor at the Marketing Department	University of Alabama, Tuscaloosa
Bruno	Hernán	Prof. Dr.	Professor of Marketing and Digital Environment	University of Cologne
Colicev	Anatoli	Ph.D.	Assistant Professor at the Marketing Department	Bocconi University
De Vries	Thom	Dr.	Assistant Professor at the Faculty of Economics and Business	University of Groningen
Dehmamy	Keyvan	Dr.	Post-Doctoral Researcher at the Marketing Department	Goethe University Frankfurt
Delre	Sebastiano	Dr.	Associate Professor at the Marketing, Sales and Communication Department	Montpellier Business School
Dew	Ryan	Ph.D.	Assistant Professor at the Marketing Department	Wharton School of the University of Pennsylvania
Dinner	Isaac	Ph.D.	Director of Econometric Modeling	Indeed
Draganska	Michaela	Ph.D.	Associate Professor at the Marketing Department	Drexel University
Entrop	Oliver	Prof. Dr.	Professor of Finance and Banking, Chair of Finance and Banking	University of Passau

(continued)

Last Name	First Name	Title	Description	Institution
Fuchs	Christoph	Prof. Dr.	Professor of Marketing and Chair of Marketing	University of Vienna
Fürst	Andreas	Prof. Dr.	Chair of Business Administration (Marketing)	Friedrich-Alexander-Universität Erlangen-Nürnberg
Gensler	Sonja	Prof. Dr.	Extraordinary Professor at the Chair for Value-Based Marketing	University of Münster
Gijzenberg	Maarten J.	Prof. Dr.	Full Professor at the Marketing Department	University of Groningen
Groening	Christopher	Ph.D.	Associate Professor at the Marketing Department	Kent State University
Haans	Hans	Dr.	Marketing Department, Director Econasium	Tilburg University
Hahn	Carsten	Prof. Dr.	Professor für Innovation und Entrepreneurship	Karlsruhe University of Applied Sciences
Hartmann	Jochen	Dr.	Post-doctoral Researcher at the Chair Marketing and Branding	University of Hamburg
Hattula	Stefan	Dr.	Market analyst	Robert Bosch GmbH
Henseler	Jörg	Prof. Dr.	Chair of Product-Market Relations	University of Twente
Hohenberg	Sebastian	Dr.	Assistant Professor at the Marketing Department	University of Texas at Austin
Junc	Vanessa	Dr.	Senior CRM Analyst	Douglas GmbH
Kamleitner	Bernadette	Prof. Dr.	Marketing Department	WU Vienna
Klarmann	Martin	Prof. Dr.	Professor of Marketing	Karlsruhe Institute of Technology
Klein	Kristina	Prof. Dr.	Professor of Marketing	University of Bremen
Landwehr	Jan	Prof. Dr.	Professor of Marketing and Chair for Product Management and Marketing Communications	Goethe University Frankfurt
Lanz	Andreas	Dr.	Assistant Professor at the Marketing Department	HEC Paris
Lemmens	Aurélie	Dr.	Associate Professor of Marketing	Rotterdam School of Management, Erasmus University
Ludwig	Stephan	Dr.	Associate Professor at the Department of Management and Marketing	University of Melbourne
Mayer	Stefan	Prof. Dr.	Assistant Professor of Marketing Analytics	University of Tübingen
Miller	Klaus	Dr.	Assistant Professor at the Marketing Department	HEC Paris

(continued)

Last Name	First Name	Title	Description	Institution
Mooi	Erik	Dr.	Associate Professor at the Department of Management and Marketing	University of Melbourne
Nitzan	Irit	Dr.	Assistant Professor of Marketing	Coller School of Management
Osinga	Ernst Christiaan	Ph.D.	Associate Professor of Marketing	Singapore Management University
Otter	Thomas	Prof. Dr.	Professor of Marketing	Goethe University Frankfurt
Papies	Dominik	Prof. Dr.	Professor of Marketing	University of Tübingen
Roelen-Blasberg	Tobias		Co-Founder	MARA
Sarstedt	Marko	Prof. Dr.	Chair of Marketing	Otto von Guericke University Magdeburg
Schlereth	Christian	Prof. Dr.	Chair of Digital Marketing	WHU – Otto Beisheim School of Management
Schulze	Christian	Prof. Dr.	Associate Professor of Marketing	Frankfurt School of Finance and Management
Sichtmann	Christina	Prof. Dr.	Research Associate for the Chair of International Marketing	University of Vienna
Stahl	Florian	Prof. Dr.	Professor of Marketing at the Department of Business Administration	University of Mannheim
Totzek	Dirk	Prof. Dr.	Chair of Marketing and Services	University of Passau
Van Heerde	Harald	Ph.D.	S.H.A.R.P. Research Professor of Marketing	University of South Wales
Vomberg	Arnd	Prof. Dr.	Professor of Marketing	University of Mannheim
Weeth	Alexander	Dr.	Engagement Manager	McKinsey & Company
Weijters	Bert	Ph.D.	Associate Professor in the Department of Work, Organization and Society	Ghent University
Wentzel	Daniel	Prof. Dr.	Chair of Marketing	RWTH Aachen University
Yildirim	Gokham	Dr.	Associate Professor of Marketing	Imperial College London

Contents

Volume 1

Part I Data	1
Experiments in Market Research	3
Torsten Bornemann and Stefan Hattula	
Field Experiments	37
Veronica Valli, Florian Stahl, and Elea McDonnell Feit	
Crafting Survey Research: A Systematic Process for Conducting Survey Research	67
Arnd Vomberg and Martin Klarmann	
Challenges in Conducting International Market Research	121
Andreas Engelen, Monika Engelen, and C. Samuel Craig	
Fusion Modeling	147
Elea McDonnell Feit and Eric T. Bradlow	
Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers	181
P. Ebbes, D. Papies, and H. J. van Heerde	
Part II Methods	219
Cluster Analysis in Marketing Research	221
Thomas Reutterer and Daniel Dan	
Finite Mixture Models	251
Sonja Gensler	
Analysis of Variance	265
Jan R. Landwehr	
Regression Analysis	299
Bernd Skiera, Jochen Reiner, and Sönke Albers	

Logistic Regression and Discriminant Analysis	329
Sebastian Tillmanns and Manfred Krafft	
Multilevel Modeling	369
Till Haumann, Roland Kassemeier, and Jan Wieseke	
Panel Data Analysis: A Non-technical Introduction for Marketing Researchers	411
Arnd Vomberg and Simone Wies	
Applied Time-Series Analysis in Marketing	469
Wanxin Wang and Gokhan Yildirim	
Modeling Marketing Dynamics Using Vector Autoregressive (VAR) Models	515
Shuba Srinivasan	
 Volume 2	
Structural Equation Modeling	549
Hans Baumgartner and Bert Weijters	
Partial Least Squares Structural Equation Modeling	587
Marko Sarstedt, Christian M. Ringle, and Joseph F. Hair	
Automated Text Analysis	633
Ashlee Humphreys	
Image Analytics in Marketing	665
Daria Dzyabura, Siham El Kihal, and Renana Peres	
Social Network Analysis	693
Hans Risselada and Jeroen van den Ochtend	
Bayesian Models	719
Thomas Otter	
Choice-Based Conjoint Analysis	781
Felix Eggert, Henrik Sattler, Thorsten Teichert, and Franziska Völckner	
Exploiting Data from Field Experiments	821
Martin Artz and Hannes Doering	
Mediation Analysis in Experimental Research	857
Nicole Koschate-Fischer and Elisabeth Schwill	
Part III Applications	907
Measuring Customer Satisfaction and Customer Loyalty	909
Sebastian Hohenberg and Wayne Taylor	

Market Segmentation	939
Tobias Schlager and Markus Christen	
Willingness to Pay	969
Wiebke Klingemann, Ju-Young Kim, and Kai Dominik Füller	
Modeling Customer Lifetime Value, Retention, and Churn	1001
Herbert Castéran, Lars Meyer-Waarden, and Werner Reinartz	
Assessing the Financial Impact of Brand Equity with Short Time-Series Data	1035
Natalie Mizik and Eugene Pavlov	
Measuring Sales Promotion Effectiveness	1055
Karen Gedenk	
Return on Media Models	1073
Dominique M. Hanssens	
Index	1097

About the Editors



Prof. Christian Homburg holds the Chair of Business-to-Business Marketing, Sales and Pricing at the University of Mannheim. He is also Distinguished Professorial Fellow of the University of Manchester (UK) and Director of the Institute for Market-Oriented Management (IMU) at the University of Mannheim. He specializes in market-oriented management, customer relationship management, and sales management. Professor Homburg has published numerous books and articles at the national and international levels and has thus established a research portfolio that places him as one of the leading German management professors and most productive scholars in the marketing discipline. In 2019 and 2020, *WirtschaftsWoche* honored Professor Homburg for his Lifetime Achievement as the leading management professor in Germany, Austria, and Switzerland.

He is currently a member of the editorial boards of five scientific journals in the United States and Europe. Since April 2011, he works as the first German area editor for the *Journal of Marketing*. Professor Homburg received several awards for his scientific research from the American Marketing Association, the world's leading scientific association in the area of marketing, and is the first European university professor to be honored as an AMA Fellow for his lifetime achievement in marketing research. In 2021, Professor Homburg ranked fourth in the American Marketing Association's global ranking, which is based on the number of publications in the most important marketing journals.

Prior to his academic career, Professor Homburg was Director of marketing, controlling, and strategic planning in an industrial company that operates globally.

In addition to his academic position, he is Chairman of the scientific advisory committee of Homburg & Partner, an international management consultancy.



Prof. Martin Klarmann is Professor of Marketing at the Karlsruhe Institute of Technology (KIT), Germany. Professor Klarmann's research is centered around three core themes: marketing using new technologies, marketing methods, and B2B sales management. His research has been published in several leading journals of the field, including the *Journal of Marketing*, the *Journal of Marketing Research*, the *Journal of the Academy of Marketing Science*, and the *International Journal of Research in Marketing*. Professor Klarmann has received several awards for his research, including an overall best paper award at the American Marketing Association's Winter Educators' Conference.



Prof. Arnd Vomberg is Professor of Digital Marketing and Marketing Transformation at the University of Mannheim, Germany. Professor Vomberg has also been an Associate Professor (with tenure) at the Marketing Department of the University of Groningen, The Netherlands. Professor Vomberg's research focuses on digital marketing and marketing transformation. He studies omnichannel strategies, online pricing, marketing automation, agile transformation, marketing technology, and marketing's impact on employees. His research has been published in several leading journals of the field, including *Journal of Marketing*, *Journal of Marketing Research*, *Strategic Management Journal*, *Journal of the Academy of Marketing Science*, and *International Journal of Research in Marketing*. Professor Vomberg has received several awards for his research, including the Ralph Alexander Best Dissertation Award from the Academy of Management.

Contributors

Sönke Albers Kuehne Logistics University, Hamburg, Germany

Martin Artz School of Business and Economics, University of Münster, Münster, Germany

Hans Baumgartner Smeal College of Business, The Pennsylvania State University, State College, PA, USA

Torsten Bornemann Department of Marketing, Goethe University Frankfurt, Frankfurt, Germany

Eric T. Bradlow The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Herbert Castéran Humanis Institute, EM Strasbourg Business School, Strasbourg, France

Markus Christen Faculty of Business and Economics (HEC) University of Lausanne, Lausanne, Switzerland

Daniel Dan Department of New Media, Modul University Vienna, Vienna, Austria

Hannes Doering School of Business and Economics, University of Münster, Münster, Germany

Daria Dzyabura New Economic School and Moscow School of Management SKOLKOVO, Moscow, Russia

P. Ebbes HEC Paris, Jouy-en-Josas, France

Felix Eggers University of Groningen, Groningen, The Netherlands

Siham El Kihal Frankfurt School of Finance and Management, Frankfurt, Germany

Andreas Engelen TU Dortmund University, Dortmund, Germany

Monika Engelen TH Köln, Cologne University of Applied Science, Köln, Germany

Elea McDonnell Feit LeBow College of Business, Drexel University, Philadelphia, PA, USA

Kai Dominik Füller Karlsruhe Institute of Technology, Institute for Information Systems and Marketing – Services Marketing, Karlsruhe, Germany

Karen Gedenk University of Hamburg, Hamburg, Germany

Sonja Gensler Marketing Center Münster – Institute for Value-based Marketing, University of Münster, Münster, Germany

Joseph F. Hair University of South Alabama, Mobile, AL, USA

Dominique M. Hanssens UCLA Anderson School of Management, Los Angeles, CA, USA

Stefan Hattula Department of Marketing, Goethe University Frankfurt, Frankfurt, Germany

Till Haumann South Westphalia University of Applied Sciences, Soest, Germany

Sebastian Hohenberg McCombs School of Business, The University of Texas, Austin, TX, USA

Ashlee Humphreys Integrated Marketing Communications, Medill School of Journalism, Media, and Integrated Marketing Communications, Northwestern University, Evanston, IL, USA

Roland Kassemeyer Marketing Group, Warwick Business School, University of Warwick, Coventry, UK

Ju-Young Kim Goethe University Frankfurt, Department of Marketing, Frankfurt, Germany

Martin Klarmann Department of Marketing & Sales Research Group, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Wiebke Klingemann Karlsruhe Institute of Technology, Institute for Information Systems and Marketing – Services Marketing, Karlsruhe, Germany

Nicole Koschate-Fischer University of Erlangen-Nuremberg, Nuremberg, Germany

Manfred Krafft Institute of Marketing, Westfälische Wilhelms-Universität Münster, Muenster, Germany

Jan R. Landwehr Marketing Department, Goethe University Frankfurt, Frankfurt, Germany

Lars Meyer-Waarden School of Management, CRM CNRS University Toulouse 1 Capitole, IAE Toulouse, Toulouse, France

Natalie Mizik Foster School of Business, University of Washington, Seattle, WA, USA

Thomas Otter Goethe University Frankfurt am Main, Frankfurt am Main, Germany

D. Papiés School of Business and Economics, University of Tübingen, Tübingen, Germany

Eugene Pavlov Foster School of Business, University of Washington, Seattle, WA, USA

Renana Peres School of Business Administration, Hebrew University of Jerusalem, Jerusalem, Israel

Werner Reinartz University of Cologne, Köln, Germany

Jochen Reiner Goethe University Frankfurt, Frankfurt, Germany

Thomas Reutterer Department of Marketing, WU Vienna University of Economics and Business, Vienna, Austria

Christian M. Ringle Hamburg University of Technology (TUHH), Hamburg, Germany

Faculty of Business and Law, University of Newcastle, Callaghan, NSW, Australia

Hans Risselada University of Groningen, Groningen, The Netherlands

C. Samuel Craig New York University, Stern School of Business, New York, NY, USA

Marko Sarstedt Otto-von-Guericke University, Magdeburg, Germany

Faculty of Business and Law, University of Newcastle, Callaghan, NSW, Australia

Henrik Sattler University of Hamburg, Hamburg, Germany

Tobias Schlager Faculty of Business and Economics (HEC) University of Lausanne, Lausanne, Switzerland

Elisabeth Schwille University of Erlangen-Nuremberg, Nuremberg, Germany

Bernd Skiera Goethe University Frankfurt, Frankfurt, Germany

Shuba Srinivasan Boston University Questrom School of Business, Boston, MA, USA

Florian Stahl University of Mannheim, Mannheim, Germany

Wayne Taylor Cox School of Business, Southern Methodist University, Dallas, TX, USA

Thorsten Teichert University of Hamburg, Hamburg, Germany

Sebastian Tillmanns Westfälische Wilhelms-Universität Münster, Muenster, Germany

Veronica Valli University of Mannheim, Mannheim, Germany

Franziska Völckner Department of Marketing and Brand Management, University of Cologne, Köln, Germany

Jeroen van den Ochtend University of Zürich, Zürich, Switzerland

H. J. van Heerde School of Communication, Journalism and Marketing, Massey University, Auckland, New Zealand

Arnd Vomberg Marketing Department, University of Groningen, Groningen, The Netherlands

Wanxin Wang Imperial College Business School, Imperial College London, London, UK

Bert Weijters Faculty of Psychology and Educational Sciences, Department of Work, Organization and Society, Ghent University, Ghent, Belgium

Simone Wies Goethe University Frankfurt, Frankfurt, Germany

Jan Wieseke Sales Management Department, University of Bochum, Bochum, Germany

Gokhan Yildirim Imperial College Business School, Imperial College London, London, UK

Part I

Data



Experiments in Market Research

Torsten Bornemann and Stefan Hattula

Contents

Introduction	4
Experimentation and Causality	5
Experimental Design	6
Definition of the Research Question	7
Determination and Operationalization of the Sources of Variation	7
Definition and Operationalization of the Measured Response-Variables	14
Decision About the Environmental Setting	17
Determination of the Experimental Units and Assignment to Treatments	21
Preliminary Testing	27
Exemplary Experimental Study	28
Ethical Issues in Experimental Research	31
Conclusion	32
References	33

Abstract

The question of how a certain activity (e.g., the intensity of communication activities during the launch of a new product) influences important outcomes (e.g., sales, preferences) is one of the key questions in applied (as well as academic) research in marketing. While such questions may be answered based on observed values of activities and the respective outcomes using survey and/or archival data, it is often not possible to claim that the particular activity has actually caused the observed changes in the outcomes. To demonstrate cause-effect relationships, experiments take a different route. Instead of observing activities, experimentation involves the systematic variation of an independent variable (factor) and the observation of the outcome only. The goal of this chapter

T. Bornemann (✉) · S. Hattula

Department of Marketing, Goethe University Frankfurt, Frankfurt, Germany

e-mail: torsten.bornemann@wiwi.uni-frankfurt.de; stefan.hattula@wiwi.uni-frankfurt.de

is to discuss the parameters relevant to the proper execution of experimental studies. Among others, this involves decisions regarding the number of factors to be manipulated, the measurement of the outcome variable, the environment in which to conduct the experiment, and the recruitment of participants.

Keywords

Experimental design · Laboratory experiment · Data collection · Cause-effect relationship · Manipulation · Experimental units

Introduction

Former US-president Obama's election campaign in 2008 made him the president with more total votes than any other US-president before him. One of the challenges of Obama's team members – among them former Google manager Dan Siroker and the social psychologist Todd Rogers – was to increase the chance that a visitor of the campaign's website would provide her or his e-mail address to become a donor or volunteer. For instance, would visitors be more likely to sign-up when the respective button asked them to “Learn More,” “Join Us Now,” or “Sign Up Now”? And which accompanying picture of Obama would be more suitable? In order to identify the website design that would generate the highest sign-up rates, visitors were randomly exposed to different button/image combinations and the respective sign-up rates were tracked. Ultimately, the best performing combination was chosen for the campaign – and this most effective design led to 140 percent more donors than the least performing combination (Nisbett 2015).

This is actually an example of a type of experiment that is as effective as it is simple, often referred to as A/B testing. A/B testing is particularly common in online environments and heralded by companies such as Google and Amazon. In A/B testing, a fraction of users is exposed to a modified version of an existing website and their behavior is then compared to that of visitors of the standard website. If the modifications lead to superior results (e.g., conversion rates), they are adopted (Christian 2012). Later in this chapter, we will refer to such designs as between-subjects designs with one factor being manipulated at different levels.

The question of how a certain activity (e.g., the intensity of communication activities during the launch of a new product) influences important outcomes (e.g., sales, preferences) is one of the key questions in applied (as well as academic) research in marketing. While such questions may be answered based on *observed* values of activities and the respective outcomes using survey and/or archival data, it is often not possible to claim that the particular activity has actually *caused* the observed changes in the outcomes. For instance, the higher levels of communication intensity may have been made possible by the initial market success of the product, or some unobserved factor may have influenced both communication intensity and sales (see also chapter ► “[Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)” by Papies, Ebbes, and van Heerde in this volume). To

demonstrate cause-effect relationships, experiments take a different route. Instead of observing activities, experimentation involves the systematic *variation* of activities (factors) and the observation of the outcome only.

The goal of this chapter is to discuss the parameters relevant to the proper execution of experimental studies. Among others, this involves decisions regarding the number of factors to be manipulated, the measurement of the outcome variable, the environment in which to conduct the experiment, and the recruitment of participants. For information on statistical techniques to analyze the resulting experimental data, the reader may refer to chapter ▶ “[Analysis of Variance](#)” by Landwehr in this volume or to one of the various works on the statistics of experimental design (e.g., Maxwell and Delaney 2004).

Experimentation and Causality

Experiments aim at probing cause-effect relationships. Referring to the example from the introductory section of this chapter, a question that Obama’s team might have asked was “What happens to conversion rates if we change the label caption from ‘Learn More’ to ‘Join Us Now?’” Thus, the goal was to examine how an independent variable (the cause; here: label caption) influences a dependent variable (the effect; here conversion rate). The conditions required to actually demonstrate that a certain cause creates a certain effect has been subject to substantial philosophical debate, but generally accepted requirements are that (1) the cause temporally precedes the effect, that (2) variations in the cause are related to variations in the effect, and that (3) rival explanations for variations in the effect can be ruled out (Shadish et al. 2002).

The characteristics of experiments are in line with these considerations (Thye 2014). The requirement of temporal order is ensured because we first actively manipulate a presumed cause (such as the label caption), thereby exposing participants to different possible realizations of the cause, and then observe the effect (i.e., what happens to the dependent variable). Also, we can directly assess how far variations in the cause are associated to variations in the effect (since we can directly track differences in conversion rates). Finally, experiments rely on random assignment of participants to the different treatments (e.g., each website visitor has the same probability of being exposed to either the “Learn More” or the “Join Us Now” label) to ensure that the different groups are equally composed in terms of factors that may have an influence on the dependent variable (e.g., physical factors such as gender, weight, age, or personality factors such as preferences, values). Simply speaking, randomization mathematically equates the groups on any known and unknown factors, which is why measured differences in the dependent variable between groups can be ascribed to the manipulated cause. Experimental settings where randomization is not possible due to practical or ethical concerns are referred to as *quasi-experimental designs*. In such settings, approaches such as propensity score matching may be used to account for factors that may differ between participants in the experimental groups (e.g., Stuart and Rubin 2007).

Experimental Design

Experimental design refers to the process of planning an experimental study that meets the objectives specified at the outset. A concise planning ensures appropriate data quality and quantity to answer the underlying research question with the required precision (Meyvis and Van Osselaer 2018). In the following, we discuss the steps involved in designing an experiment (see Fig. 1) and subsequently illustrate this process with an example.

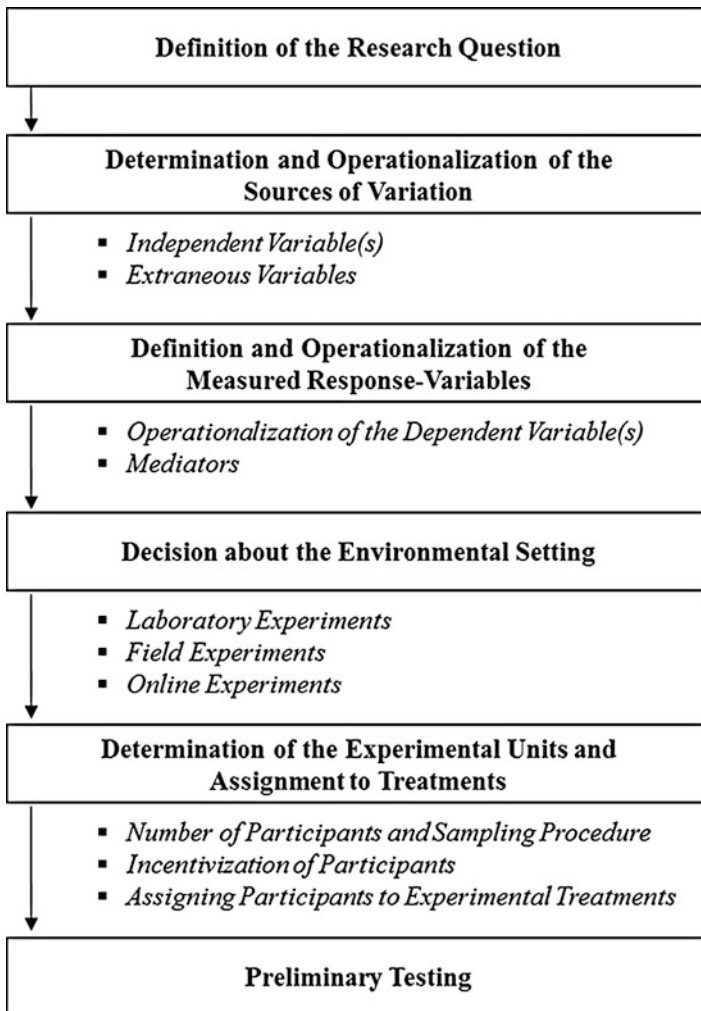


Fig. 1 Steps involved in the design of an experiment

Definition of the Research Question

The actual planning of a study starts with a precise formulation of the research problem(s) and, resulting from that, a clear definition of the objectives of the experiment. The problem statement typically lists the essential issues that will be tested in the experiment (Dean et al. 2017). It should be formulated as specific as possible to reduce complexity while at the same time avoiding so-called Type III errors, which refer to the problem of leaving out important factors from the experimental design and/or choosing insufficient measures of process performance. This also includes the determination of the response variable(s) and the population to be studied. For instance, a problem statement could ask which price promotion causes more sales – a 10% off or a 10€ off coupon.

The objectives of the experiment then determine the questions to be addressed (Morales et al. 2017). Typically, experimenters also propose answers to the questions being investigated by formulating hypotheses. This includes a null hypothesis suggesting no difference between the experimental treatments and an alternative hypothesis that argues for such a difference. For instance, the null hypothesis in the example above would suggest that the 10% off coupon and the 10€ off coupon show the same results, while the alternative hypothesis claims differences in resulting sales. As an important remark, experimenters should be careful to not formulate untestable hypotheses – i.e., those that cannot be easily interpreted, are unethical, or even empirically untestable. For instance, oftentimes it is more feasible to divide a very extensive hypothesized claim into several subhypotheses.

Determination and Operationalization of the Sources of Variation

In line with the idea of experimentation to systematically vary an independent variable and then observe changes in a dependent variable of interest, the second step in designing an experimental study entails an identification of the sources of variation. Sources of variation comprise the manipulated independent variable(s), whose effect on the outcome variable(s) is of primary interest, and extraneous variables, which may challenge the inference of causality and thus lead to wrong interpretations of the results.

Independent Variable(s)

In experimental research, the manipulated independent variable is referred to as a factor. The experimenter manipulates the factor by creating different levels of it. Thus, participants of the experiment are randomly assigned to different levels of a factor or, in other words, they are “treated” with different values of the factor, which is why the different factor levels are also referred to as treatments. In most instances, one level of a factor serves as a control in which no experimental treatment is administered. This control group makes it possible to determine whether the independent variable had any effect on the dependent variable. Central decisions in this step of the experimental design process thus involve the determination of the

number of factors to be varied and their respective levels (which emanate from the research question) as well as the operationalization of the different treatments.

Factors and their levels. Single-factor designs are the most basic form of comparative studies, where one independent variable (factor) is manipulated at multiple, qualitatively distinct levels (the treatments). The idea of such a design is to study the causal main effect of one independent variable on one or multiple dependent variable(s). Single factor designs are often used to get an initial feeling of how a treatment works. For instance, experimenters may want to test the impact of ambient scent (cool vs. warm) on ad preference and product purchase intention. To determine whether scenting the environment has any effect at all on the dependent variables, a control group may be included where participants also indicate their ad preference and product purchase intention, but without being exposed to any scent. The result thus is a single-factor design involving three levels of ambient scent: cool [e.g., peppermint] vs. warm [e.g., cinnamon] vs. control [no scent] (Madzharov et al. 2015).

In contrast, multifactor designs systematically vary the levels of two or more independent variables (i.e., treatment factors) (Collins et al. 2009). These factors are crossed such that combinations of the respective levels are formed. The classic case of a multifactor design is the $2 \times 2 = 4$ experimental conditions case. It describes the combination of two independent variables, which are manipulated at two distinctive levels each. Such multifactor designs enable an examination of interactions between factors. For instance, the previously described effect of ambient scent (cool vs. warm) on product purchase intention may depend on the salesperson's behavior (rude vs. polite) (Madzharov et al. 2015). Such a multifactor design where all combinations of the respective levels are formed is referred to as a full factorial design.

Despite the advantages of multifactor designs, one should consider that an increasing number of independent variables may result in a large number of experimental conditions and thus an increased complexity of the experimental design (Koschate-Fischer and Schandelmeier 2014). For instance, the interpretation of interactions between two (two-way interactions) or three (three-way interactions) variables seems plausible. However, the more combinations of variables exist, the less meaningful are their interpretations. Some of those combinations may even be nonsensical, logistically impractical, or otherwise undesirable (Collins et al. 2009). Therefore, to make settings with a large number of variables feasible, reduced/fractional factorial designs are commonly applied (Cox 1992).

Fractional factorial designs describe designs where only a fraction of a full factorial experiment is used. Common is the so-called half fractional factorial design, where only half of the experimental treatments are used. For instance, experimenters may investigate a design with three factors and two levels each. This would involve $2^3 = 8$ experimental runs in a full factorial design. In Table 1, we illustrate this case for three factors (A, B, and C) that are each varied at two levels (high and low). The small letters denote treatment combinations, while capital letters refer to factors and effects. "I" denotes the identity column that only contains "+" signs (any other column multiplied by itself yields this column), while (1) describes the combination

Table 1 Example of a fractional design

Treatment combination	Identity column	Factor			Two-factor interaction			Three-factor interaction
	I	A	B	C	AB	AC	BC	ABC
(1)	+	-	-	-	+	+	+	-
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
ab	+	+	+	-	+	-	-	-
c	+	-	-	+	+	-	-	+
ac	+	+	-	+	-	+	-	-
bc	+	-	+	+	-	-	+	-
abc	+	+	+	+	+	+	+	+

of treatments where all factors are at their low levels. Moreover, a small letter in the treatment combination symbol indicates that the respective factor is at its high level; if the factor does not appear, it is at its low level (e.g., the treatment combination symbol ac indicates that factors A and C are at their high level, while factor B is at its low level). The “+” and “-” signs in the respective factor columns indicate that the factor is at its high or low level. Signs for the interaction columns are the result of taking the product of the signs of the involved factor columns.

If we were now restricted in our resources and only able to run four experimental conditions, we would have to resort to a fractional factorial design, which in this case would be written as 2^{3-1} . The central question is which four out of the eight conditions to choose. A first step is to select an interaction that can be neglected. Typically, experimenters choose to sacrifice the highest order interaction for this purpose since it is the most difficult to interpret interaction. In our case, we select ABC as the highest order interaction. Hence, ABC is used to generate the fraction of conditions to be tested. In order to identify those conditions, we select the conditions from those rows where the ABC column indicates a “+”. So in our case, the four conditions to be run are a, b, c, and abc (see the shaded rows in Table 1).

Subsequently, we can determine which effects are aliases (i.e., are confounded). In our case of three factors, we would take the symbols for the highest order interaction (ABC) and attach them to every effect. Whenever the same factor appears twice in such a combination, it is erased – what remains is the alias of the original effect. For instance, if we take effect A, we have A(ABC). When we delete the A’s, we are left with BC. That means the main effect of factor A is confounded with the BC interaction, or in other words, the BC interaction and factor A are aliases. In the current example, also B and AC as well as C and AB are aliases (for more details, see Montgomery 2009).

The specific number of levels that are manipulated for each factor is an important decision on its own. For instance, previous research indicates that attribute importance measures can be sensitive to the number of levels. The more levels exist for a particular attribute, the more attention and thus subjective importance is given to this attribute (Verlegh et al. 2002).

Decisions about the number of levels are easy if the independent variables are nominal-scaled (Koschate-Fischer and Schandelmeier 2014). They result directly from the research question. However, if the underlying scale is metric or quasi-metric, more than two levels might be necessary. This especially holds for research questions where nonmonotonic or nonlinear relationships between the independent variable and the dependent variable are likely (Cox 1992). For instance, previous research has examined the relationship between repeated advertising exposure and consumers' affective response in experimental designs with four levels (zero, low, medium, high) of the advertising repetition manipulation (Nordhielm 2002). Results suggest an inverted U-shaped relationship, where consumers' affective judgment was most positive for the moderate frequency of advertising exposure.

However, an increasing number of factor levels not only increases the experimental sensitivity, but also the number of treatments. Therefore, more than three levels may only be considered in single-factor designs in order to identify the most promising levels for further investigation. In multiple-factor designs, each individual variable should be limited to two or three manipulated levels each (Eriksson et al. 2008). If more than three levels are considered in multiple-factor designs, a fractional design may be advisable. As mentioned earlier, those designs consider not all possible treatments but only investigate logically useful ones to keep the complexity at a manageable and interpretable level (Cox 1992).

Operationalization of the treatments. A key challenge when translating a theoretically defined independent variable of interest into an operational treatment to be administered to participants is that the respective treatment might not incorporate all characteristics of the variable of interest (underrepresentation) or that it might contain content that is not part of the theoretical variable (thereby creating a confound). A number of aspects can lead to such mismatches (Shadish et al. 2002). For instance, a design that compares a treatment group with a control group may fail to detect an effect of the treatment on the dependent variable simply because the treatment is implemented at a too low level. Thus, in most instances, it is advisable to manipulate the independent variable at several levels. Appropriate and realistic levels of the independent variable can be identified based on a pretest (see section “[Preliminary Testing](#)”).

The manipulation of the independent variable is often implemented within stimulus material containing text (a scenario in which the study is embedded, the instructions to the participants, and measures assessing the dependent variable), pictures, and sometimes video material. Rashotte et al. (2005) describe the creation of this material as “creating an alternate reality that volunteer participants will inhabit for a time” (p. 153). The design of such stimulus material requires great care, and it is indispensable to conduct preliminary tests before the experiment to ensure that participants interpret the material as intended. Generally, material and instructions should be as standardized as possible, avoiding complexity and ambiguity in the language used. A predetermined script conveys all necessary information to the participants and avoids any distractions. As a rule of thumb, all crucial instructions are repeated three times using slightly different wording to ensure reception of the information (Kuipers and Hysom 2014).

To assess whether the manipulation has worked successfully (and not created a confound) and how far participants have really attended to the information containing the manipulation of the independent variable (Oppenheimer et al. 2009), most (academic) research includes manipulation checks in the experimental material (Kuipers and Hysom 2014; Perdue and Summers 1986). Three aspects should be considered when designing a manipulation check in experiments.

First, the experimenter should be able to show that the manipulation indeed changes the theoretical variable that it was intended to alter (Perdue and Summers 1986). To check whether the manipulation meets this criterion, one has to find good indicators that assess the respective theoretical construct (Koschate-Fischer and Schandelmeyer 2014). This is a rather easy task if the manipulated variable pertains to a simple construct such as price level. If the researcher has manipulated the price level (e.g., low/medium/high) of a given product to examine how different prices affect purchase intention, a simple single-item scale may serve as a manipulation check by asking participants whether they perceived the price as low vs. high. For a successful manipulation, significant differences in the single-item measure should be observed between the different treatment groups.

Second, experimenters should also detect and account for satisficing behavior of participants (Oppenheimer et al. 2009). Satisficing means that participants might wish to minimize cognitive effort in survey responses and therefore often do not properly read texts, questions, or instructions in the respective studies. Reception of such information, however, is necessary to produce useful data, which is why satisficing behavior can cause noise and reduce experimental power. This particularly holds for scenario-based experimental designs, where textual descriptions are used to create realistic experimental environments. To detect satisficing behavior, instructional manipulation checks have been developed (Oppenheimer et al. 2009). These are directly embedded in the experimental material and consist of questions that are similar to other questions in length and response format. If participants follow these instructions, experimenters get confirmation that texts, questions, and instructions have been read carefully and participants have spent the cognitive effort necessary to understand the experimental material. Figure 2 provides an example of such an instructional manipulation check. Here, participants are instructed to ignore the private transportation usage question but to proceed with the survey by directly clicking the continue button.

Third, experimenters have to decide about the point of time in the experiment when the manipulation check has to be answered. This also includes the question where to place the items that check the effectiveness of the manipulation – whether they appear before or after the measurement of the dependent variable (Perdue and Summers 1986). In this respect, it has long been argued that the items should be placed after measurement of the dependent variable in order to avoid demand effects (Wetzel 1977). However, this approach has been criticized since answering the dependent variable before might bias the subsequent manipulation check (particularly in self-report measures of the manipulation check). Moreover, the effect of the manipulation might weaken substantially over time (Kuipers and Hysom 2014; Perdue and Summers 1986). Placing the manipulation check before the assessment

Virtual QR Code Stores

It is 8am on a Monday morning and, just like every morning, you take the tram to get to work. However, when arriving at the platform, you notice that there are no longer posters with advertisements for local retailers, events, etc. Instead, the posters show pictures of supermarket shelves listing numerous products, which are typically offered in supermarkets.

Each single product on these shelves is listed together with an own QR code. By scanning this code, you can put the respective product in an online shopping cart and you have the possibility to initiate a same-day home delivery afterwards. We are interested in unbiased and meaningful results. Therefore, it is important that you have read and understood the Virtual QR Code Store scenario. To demonstrate this, please ignore the following question on your private transportation usage by not checking any of the six answer boxes. Instead, please directly click the continue button. Thank you very much!

Which of the following transportation do you typically use?
(multiple checks are allowed)

<input type="checkbox"/> Own car	<input type="checkbox"/> Car sharing	<input type="checkbox"/> Bus
<input type="checkbox"/> Tram	<input type="checkbox"/> Train	<input type="checkbox"/> Bike

Fig. 2 Example of instructional manipulation check

of the dependent variable, however, may cause interference with the effect of the manipulation or may even be interpreted as a necessary condition for the observed effect (Trafimow and Rice 2009).

Given these issues, researchers today call for careful evaluation of whether the benefits of a manipulation check truly outweigh the costs (Sawyer et al. 1995; Trafimow and Rice 2009). Many situations exist where manipulation checks add only little informational value to theory testing. This includes variables that are isomorphic with their operationalization such as black and white vs. colored pictures in advertising (Sawyer et al. 1995). Moreover, manipulation checks are not necessarily required in situations where well-known manipulations of variables are used and therefore confidence in the manipulation is backed by existing literature, or where the manipulations have been extensively pretested (see section “Preliminary Testing”). Verifying the appropriateness of a manipulation in the course of a pretest separately from the main experiment may therefore be advisable to circumvent the previously discussed problems associated with the placement of the manipulation check items.

Besides employing manipulation checks, also the use of more than one operationalization of a given theoretical variable may reduce the threat of false inferences. While it is relatively easy to use multiple measures to assess an outcome variable (e.g., different multi-item scales to capture variables such as attitude or preference), employing different manipulations for the independent variable is more challenging

since it increases the required sample size. For instance, by manipulating psychological distance based on both temporal and social aspects, Bornemann and Homberg (2011) show that psychological distance differently impacts price perception depending on the respective operationalization of psychological distance.

Extraneous Variables

Different types of extraneous variables may challenge the inference of causality in experimental research. Most problematic are so-called confounding variables, which describe variables that are related to two variables of interest (i.e., factor and observed response variable) and thus may explain their relationship (Harris et al. 2006). A confounding variable may vary with the levels of the manipulated variable and at the same time be associated with the dependent variable. As a consequence, the observed causal relationship between the independent and the outcome variable can be the result of the confounding variable's impact.

Generally, randomization should solve the issues resulting from confounding effects by equally distributing both known and unknown confounding variables between experimental conditions. However, in some situations, randomization may be hard to apply (e.g., in quasi-experimental designs) or ineffective to achieve constant terms between experimental conditions (e.g., due to small sample sizes). Moreover, some manipulations may naturally cause more than just the intended effects. For instance, one may create different levels of crowding in stores to manipulate the extent of potential social contacts. Such contacts should be valued by older consumers as they constitute a compensation for the age-related loss of companionship (Myers and Lumbers 2008). However, there is also evidence that crowding increases perceived stress in a shopping environment, which may prevent older consumers from shopping at all (Albrecht et al. 2017). As such, not necessarily the level of social contact but instead the perceived stress may relate to older consumers' affective and behavioral states. To overcome the issues arising from confounding effects, researchers may include confound checks in experimental designs to show that the manipulation does not change related but different constructs (Perdue and Summers 1986). Similar to a manipulation checks, this involves the administration of indicators that assess the respective theoretical constructs (Sawyer et al. 1995).

The timing of measurement of confounding variables depends on their potential interrelationships with the manipulated independent variable, the manipulation check, and the outcome variable (Koschate-Fischer and Schandelmeier 2014). If the extraneous variable is affected by the manipulation (check), it should be measured after provision of the treatments to the participants (the manipulation check). The same order holds true if effects of the dependent variable on the confounding variables are likely.

In addition to confounding variables, so-called suppressor variables may affect the magnitude of the relationship between the independent and the dependent variable (Cohen et al. 2003). Suppressor variables are correlated with the independent variable but not with the outcome and therefore add irrelevant variance, which may hide or suppress the real relationship between the latter two variables. As

a consequence, the relationship may actually become stronger or weaker or even change sign. When including these suppressor variables into the respective analysis, the unwanted variance is removed from the relationship of interest. For instance, we would like to know whether differences in knowledge of history exist between younger and older people and administer a test to both groups that has to be completed within an indicated time limit. We may expect that older people know more about history than younger people but find that younger people have answered more of the test's questions. A possible explanation for this finding may be that older people are simply slower readers. Slow reading has prevented them from answering all questions but is otherwise not associated with knowledge of history. In this case, reading speed constitutes a suppressor variable we want to control for in our analysis to obtain the true relationship between age and knowledge of history.

Definition and Operationalization of the Measured Response-Variables

Similar to the operationalization of the treatments, also the dependent variable of interest has to be translated into an operational measure. Moreover, additional variables may have to be assessed, particularly to reveal processes underlying a cause-effect relationship or to rule out alternative explanations.

Operationalization of the Dependent Variable(s)

Using interviews and surveys, experimenters mostly refer to self-report measures to capture participants' response to experimental manipulations (Morales et al. 2017). This approach involves asking respondents (multiple) direct questions about, for instance, their feelings, attitudes, or behavioral intentions. These questions typically come in the form of rating scales or fixed-choice questions. Rating scales capture how strong participants agree or disagree with the statements and can reveal the degree of response. In contrast, fixed-choice questions force respondents to make a fixed-choice answer such as "yes" or "no," or "buy" or "not buy."

Sometimes, researchers are interested in aspects they cannot operationalize due to restrictions they face when conducting an experimental study. For instance, marketing researchers may be interested in factors influencing brand choice, but are not able to actually measure real choice behavior and therefore resort to scales assessing brand attitude as a proxy. Unfortunately, the link between attitudes and actual behavior is not really stable and highly contingent on a number of factors. For instance, attitudes are better able to predict behavior when they are formed based on actual experience with an object and often retrieved from memory (Glasman and Albarracín 2006). This implies that particularly in case of new products, researchers should refrain from using attitude measures if the theoretical interest lies in behavioral consequences.

An additional aspect that poses challenges to the measurement of outcome variables is the demand effect described in the section "[Assigning Participants to Experimental Treatments](#)." In a within-subject design, where scores of an outcome

variable are obtained repeatedly from the same participant, carryover-effects may occur or participants may guess the goal of the study and – consciously or not – act accordingly when providing scores for scale items that aim to assess the outcome variable. For instance, a huge stream of literature examines consequences of various treatments on participants' emotions (in a marketing context, this could be the impact of different advertisements on emotional reactions). While a first challenge in this context is the theoretical definition of emotions, how to measure them, particularly in within-subject designs, is even more difficult. Emotional reactions may manifest in various ways, such as facial action, vocal changes, cardiac responses, or subjective experiences. Particularly the latter are most often assessed based on self-report scales. In such cases, the standard deviations across the different measurement points may be compared to assess whether participants simply transfer a prior judgment to a subsequent assessment (stereotypic responding; Larsen and Fredrickson 1999).

Various outcome variables – including emotional reactions – can also be assessed without the need for participants' self-reports, thus circumventing this issue. In many marketing-related research questions, the degree to which participants attend to particular information depending on the structure of that information is of interest (for instance in the context of designing advertisements). In such cases, eye-tracking may be used to examine participants' attention to particular pieces of information. Also emotional reactions may be assessed continuously throughout the whole experiment based on physiological recording devices or via recording and coding participants' facial reactions during the experiment (Larsen and Fredrickson 1999). Since emotional reactions have been shown to systematically manifest in changes of facial muscles, several software solutions have been developed (e.g., IntraFace by Carnegie Mellon University's Human Sensing Laboratory) that code the recorded material into specific emotional reactions. More recent developments also show that mouse cursor movements may be used to index participants' emotional reactions (Hibbeln et al. 2017), providing an unobtrusive way to assess emotional reactions in online experiments.

Mediators

Assumptions on causal relationships in experiments should have a clear theoretical rationale, which is why researchers are increasingly interested in revealing the psychological processes underlying those relationships (Spencer et al. 2005). In other words, they search for variables that causally transmit the effects of the manipulated variable to the dependent variable – so-called mediators (Bullock et al. 2010). For instance, following the theory of cognitive dissonance, researchers postulate that the experience of incongruent information (e.g., positive impression of a brand and negative experience with one of its products) may cause an aversive arousal in individuals' minds, which in turn can initiate changes in their attitudes (Festinger 1957). In this case, the aversive arousal is the mediator of the incongruent information-attitude relationship. Such mediators can be particularly relevant for studies where no main effect of the independent variable on the dependent variable is found in initial studies. Often, researchers abandon their project at these stages in the

assumption that there is no effect at all. However, multiple psychological processes may simultaneously mediate a relationship, but with opposite signs, thereby causing the nonsignificant main effect (Zhao et al. 2010). For instance, advertising may (1) increase the consideration set of consumers, which in turn increases price sensitivity. On the other hand, it may (2) increase perceived differences in utility among competing products, which negatively affects price sensitivity (Mitra and Lynch 1995).

To operationalize the mediator variable, three approaches exist. First, most experiments use a so-called *measurement-of-mediation* design, where the proposed mediator/s is/are measured using survey item scales (Spencer et al. 2005). Then, regression-based causal mediation analyses are applied (e.g., Zhao et al. 2010). This approach, however, has been criticized for several limitations (see Spencer et al. 2005 for further discussion), of which the causal inference of the mediator-dependent variable relationship is the most decisive (Pirlott and MacKinnon 2016). Randomly assigning participants to levels of the manipulated independent variable and measuring both the mediator and outcome variables enables interpretation of the independent variable-mediator and the independent variable-outcome relationships. However, it is not possible to decipher whether the mediator causes the outcome, the outcome causes the mediator, or unmeasured confounding variables cause both the mediator and the outcome variable. To overcome the limitations of this approach and thus make strong inferences about the causal chain of events, researchers suggest manipulating not only the independent variable but also the mediator. In this respect, the experimental-causal-chain and moderation-of-process designs are the most used design approaches (Spencer et al. 2005). In *experimental-causal-chain* designs, two experiments are conducted. In the first experiment, participants are randomly assigned to the levels of the independent variable while the mediator and the outcome variables are measured. This allows for an unambiguous interpretation of the effect of the independent variable on the mediator and outcome variables, respectively. In a second experiment, the causal effect of the mediator variable on the dependent variable is tested. Here, participants are randomly assigned to levels of the manipulated mediator variable and the outcome variable is measured. The respective levels are defined based on the changes of the mediator variable caused by the independent variable in the first experiment. In contrast, in *moderation-of-process* designs, the independent and the mediator variables are simultaneously manipulated in a two-factor experimental design. These manipulations allow for inferences about the causal effects of both the independent variable and the mediator variable on the measured outcome variable. Moreover, a manipulation check that measures the mediator variable is applied, which is why also the effect of the independent variable on the mediator variable can be tested.

However, experimental-causal-chain and moderation-of-process designs have some drawbacks as well. Some psychological processes such as personal commitment are not easy to manipulate. Moreover, the manipulation of the mediator must be the same as the measured variable before, which constitutes a serious limitation of this approach. For instance, Postmes et al. (2001) failed to show a successful

manipulation of group norms when measuring this variable. Also, the additional manipulation of the mediator variable requires larger sample sizes.

Given the issues involved in the three designs, Spencer et al. (2005) make some suggestions regarding when to use which approach based on how easy it is to measure the proposed processes and how easy those processes are to manipulate. Specifically, the experimental-causal-chain design is the simplest and most straightforward approach if the mediator can be both measured and manipulated. If one has the resources, a series of studies conducting all three approaches would be the best option. In situations where the mediator can easily be manipulated but measuring it is difficult, the moderation-of-process design is recommended. The most prevalent approach in existing research – the measurement-of-mediation design – may be used if the manipulation of the mediator is difficult, but the mediator can be measured.

Decision About the Environmental Setting

Having identified and operationalized the sources of variation and response variables, experimenters need to decide about the environmental setting that best fits the defined requirements. Oftentimes, the environment of an experiment is chosen based on convenience aspects such as saving cost and time, or ease of application (Li et al. 2015). However, the experimental environment affects the controllability of the manipulation by the researcher as well as the participants' behavior. It is human nature to pay attention not only to the manipulated stimuli but also to the experimenter's environment (Harrison and List 2003). Individuals employ learned strategies and heuristics to cope with those influences, which is why insights from isolated snapshots in controlled settings can provide misleading insights of "true" behavior. Against this background, the next subsections provide a deeper understanding of three prevalent environmental settings of experiments (laboratory, field, and online).

Laboratory Experiments

Laboratory experiments describe experimental designs in controlled environmental settings. They have been almost neglected until the late 1940s, but have since become an integral part of today's marketing research. For instance, between 2000 and 2007 alone, more than 1200 laboratory experiments have been published in the four leading marketing journals (Baum and Spann 2011). This prevalence in academia may result from the special merits of laboratory experiments. Employing artificial environments, experimenters can eliminate many of the confounding influences (e.g., noise level, architectural design) that may otherwise affect the results of experiments (Falk and Heckman 2009; Harrison and List 2003). Therefore, a higher internal validity – referring to the extent to which an experimental manipulation is truly responsible for variations in the dependent variable (Shadish et al. 2002) – is assigned to laboratory experiments compared to other environmental settings. Moreover, experiments in controlled environments enable randomized allocation of participants to conditions, counterbalancing, and the use of standardized instructions, which facilitates later replication (Aaker et al. 2011).

However, there is an intense discussion on whether the findings from laboratory experiments are realistic and “right” for theory testing. Researchers criticizing controlled environmental conditions as being unrealistic argue that the context in which participants’ decisions are embedded (and the associated level of scrutiny) and the way they are selected to participate influence their behavior (List 2011). For instance, there is empirical evidence that participants in laboratory environments might make assumptions about the experimenter’s objectives and adjust their behavior to align with these expectations (Benz and Meier 2008). To ensure that none of the participants is aware of the true purpose of the study, researchers increasingly administer suspicion probes in their experimental material (e.g., Hattula et al. 2015). Moreover, in natural environments, individuals may adapt their behavior in ways that can hardly be captured in a laboratory environment (Levitt and List 2007). For instance, individuals can simply stop shopping and leave the particular store. Consequently, the results from the laboratory may not be generalizable to real markets and thus limit the external validity of the respective findings.

External validity denotes how far a causal relationship that has been uncovered in an experimental study can be generalized beyond the context of the experiment in terms of people, places, treatments, and outcomes (Shadish et al. 2002). An important prerequisite for external validity is that the experiment’s participants represent the true population of interest and that the experimental setting is perceived as realistic as possible. For instance, experimentation is often used to identify effective designs and layouts of advertisements. In such studies, participants are frequently exposed to variations of the focal advertisement only (in terms of pictures included, font sizes, etc.), whereas in real life, consumers normally are exposed to a sequence of advertisements for different brands and products (e.g., when watching TV-commercials or reading a magazine or newspaper). Thus, the attention participants pay to an isolated advertisement in an experimental study may be different from the attention they would pay for it in real life situations.

The extent of a laboratory experiment’s external validity therefore depends on its design and execution in the specific context. “The external validity of an experiment cannot be evaluated either a priori or a posteriori (e.g., on the basis of sampling practices or realism) in the absence of a fairly deep understanding of the structural determinants of the behavior under study” (Lynch 1982, p. 238). In this respect, Koschate-Fischer and Schandelmeier (2014) discuss three aspects that influence the generalizability of findings generated from laboratory experiments. First, the naturalness of a laboratory setting strongly depends on the operationalization of the independent and dependent variable(s). For instance, the levels of the manipulation of the independent variable should be different enough to represent meaningful categories (Zikmund and Babin 2006). Second, not the mundane realism (i.e., physically resembling the real world) of an experiment is important, but the experimental realism matters (Berkowitz and Donnerstein 1982). That is, experiments should be designed such that participants are caught up in the procedures and forget that they are part of an experiment – they should perceive the research setting as natural. Finally, laboratory experiments should be conducted in different contexts to provide valid results. For instance, the threats to external validity can be varied in

multiple experiments to examine their relevance (Falk and Heckman 2009) and thus to enhance the generalizability of the findings.

Field Experiments

Field experiments represent a conjunction of experimentation and fieldwork – they describe experimental designs conducted in mostly natural environmental settings (see also chapter ▶ “Field Experiments” by Valli, Stahl, and Feit in this volume). Literature in marketing has recently seen an increase in research using field experiments: more than 60% of all field experiments published in the leading marketing journals over the last 20 years were published in the most recent 5 years (Simester 2017). Following Harrison and List (2004), three types of field experiments can be differentiated (see Fig. 3). *Artefactual* field experiments are similar to laboratory experiments except for one feature – they involve a nonstandard pool of participants. Instead of recruiting students, participants of the experiment are drawn from the real market (List 2011). Thus, the respective research more closely features the real world actors of interest. Compared to artefactual field experiments, *framed* field experiments additionally consider a realistic task to avoid confounding effects that result from a laboratory setting. That is, the experiment is framed “in the field context of the commodity, task, stakes, or information set of the subjects” (List 2011, p. 5). Finally, *natural* field experiments are similar to framed field experiments, but here, participants naturally undertake the tasks and therefore are not aware of participating in an experiment. This combination of realism and randomization helps avoiding that participants adjust their behavior to align with assumed expectations of the experimenter (Benz and Meier 2008). Natural field experiments therefore maximize the generalizability and thus the external validity of experimental findings. They simulate as closely as possible the conditions under which a causal process occurs (List 2011).

However, the uncontrolled and frequently complex environmental settings in field experiments have been argued to limit the internal validity of those experiments (Aaker et al. 2011). Since confounding environmental characteristics are not held constant, establishing cause-effect relationships is difficult. Therefore, “an ideal field experiment not only increases external validity, but does so in a manner in which

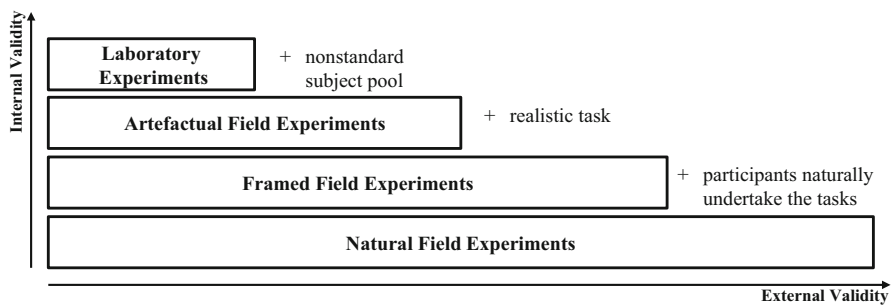


Fig. 3 Overview of laboratory and field experiments

little internal validity is foregone” (Harrison and List 2003, p. 37). To achieve this, several prerequisites have to be met. First, the treatment variable must be exogenous (Remler and Van Ryzin 2010). In other words, the dependent variable should not determine the manipulation of the independent variable. Second, the characteristics of the treatment and the control group must be truly comparable. For instance, finding comparable locations for both groups is crucial (Trafimow et al. 2016). Matching techniques can be applied, where locations are evaluated based on variables that might differentiate them. Finally, experimenters should capture as many control variables as possible to account for potential confounding effects (Arnold 2008).

At this point, one might expect a recommendation regarding whether to use laboratory or field experiments in market research. There is a lively debate on this question. Some researchers favor laboratory experiments due to their high internal validity, whereas others value the external validity of field experiments. “The obvious solution is to conduct experiments both ways: with and without naturally occurring field referents and context” (Harrison and List 2004, p. 1050). Combining the results from laboratory and field experiments can help understanding the mechanisms of behavior observed in the real market (Falk and Heckman 2009). For instance, differences in the findings between both settings indicate the need to study the environment in more detail in order to understand real world behavior of market actors.

Online Experiments

The third environmental setting is Internet based – the respective experimental design is mostly referred to as online experiments. Due to technological advancements and cost and distribution advantages (see Koschate-Fischer and Schandelmeier 2014 and Reips 2002 for further advantages), online experiments have become a standard instrument in today’s market research. Descriptions of this experimental setting highlight the possibility of an interaction of participants with the experimenter or other participants via the Internet (Morton and Williams 2010). This interaction is direct, but virtually mediated through the internet. Due to the fact that participants of online experiments remain in their natural environment, some researchers suggest that online experiments pertain to the category of field experiments (Koschate-Fischer and Schandelmeier 2014). This categorization may hold for settings where online marketing phenomena such as e-commerce strategies are studied and the use of a computer thus describes the natural environment of the participants. However, online experiments are also applied to study offline phenomena and to validate and/or extend results from field and laboratory studies (Reips 2002). In this latter case, scenario techniques ask participants to imagine offline situations, which is why they do not act in their situation-specific natural environment anymore. Depending on the actual context and execution, online experiments may therefore share more common properties with either field or laboratory settings (Morton and Williams 2010).

Given the use of online experiments to study offline behavior, the question arises as to whether data obtained online and offline provide equivalent results.

Researchers arguing against this equivalence highlight confounding factors that may affect individual behavior. Online respondents cannot scan, preview, review, skip, or change items, which is why they may experience a different level of self-generated validity (Feldman and Lynch Jr. 1988). Moreover, results may be different in a more public offline setting, where aspects such as social pressure are present that could influence individual shopping behavior (Bearden and Etzel 1982). Recent research, in contrast, indicates equivalence of online and offline data and demonstrates acceptable levels of accuracy, completeness, and response quality in some settings (Hauser and Schwarz 2016). Moreover, due to their broad coverage, online experiments provide good generalizability of findings to the population (Horswill and Coster 2001) and to a good number of settings and situations (Laugwitz 2001).

Nonetheless, careful evaluation of the applicability of online experiments for the individual research question and context is necessary. Online experiments are less suited for studies that require high levels of attention to the study material and instructions and that ask for a high motivation (Goodman et al. 2013). Also, participants in online settings seem to mentally process the material differently in that they pay more attention to minor aspects of instructions (Hauser and Schwarz 2016). Similarly, it is difficult to reliably assess factual answers such as competencies, since participants of online studies can simply search for answers in the internet (Goodman et al. 2013).

Determination of the Experimental Units and Assignment to Treatments

The next step in the process of designing an experimental study involves the determination of the experimental units and the rule by which the experimental units are assigned to the different treatments. Experimental units describe the entities upon which the relationships of interest are tested, i.e., which are subjected to the respective treatments. In marketing research, the experimental units are typically human participants (e.g., customers, managers). The high percentage of student samples in existing experimental research studying market phenomena has sparked a lively discussion on their qualification as experimental units (Koschate-Fischer and Schandelmeier 2014). Critics highlight that students are not representative for the decision-making of “real” customers because they are better educated and represent less mature personalities (Sears 1986). Moreover, they constitute a very homogeneous group with little variation in factors such as age or income, which makes the study of personal characteristics more difficult. Even more critical is the use of student samples to address business problems such as the selection of employees in organizational settings. Such issues are generally addressed by managers whose qualifications and job experience is key to make the respective decisions (Hakel et al. 1970). Advocates of student samples, however, argue that only little research exists that finds generalizable significant differences in the decision-making of students and “real” customers or managers, respectively. Such differences only exist if the above-mentioned characteristics such as age or job experience truly influence the

independent variable's effect on the dependent variable. This, however, is not very often the case for the phenomena studied in experimental research (Lynch 1982). The homogeneity in individual characteristics can even be beneficial to experiments because it reduces many confounding effects (see section “[Extraneous Variables](#)”). The conclusion from this discussion is that a careful evaluation of the suitability of students before conducting an experiment is required.

In the following, we will elaborate on the number of participants required and how to sample them, the question of whether and how to incentivize these participants, and finally how to assign participants to the different treatments.

Number of Participants and Sampling Procedure

One of the most frequently asked questions of students and practitioners in empirical projects is “How many participants do I need?” (Evans and Rooney 2013, p. 134). There is no universal answer to this question because different aspects have to be considered in determining this number. First, the required sample size depends on the specific experimental research design. A large number of treatments require a larger sample size especially in designs where each participant is exposed to only one of the multiple treatments (see section “[Assigning Participants to Experimental Treatments](#)”). In this respect, simple heuristics have been introduced suggesting to use approximately 30 participants for each experimental condition (Sawyer and Ball 1981). Usability professionals even argue that just 5 participants could be enough in qualitative experiments. They should reveal about 80% of all usability problems that exist in a product (Nielsen 2000). However, the number of participants should not be determined without considering the experimental setting itself. For instance, at least 39 users are required in eye-tracking studies to get stable heatmaps (Nielsen 2012).

More sophisticated approaches consider anticipated or previously observed effect sizes to calculate the statistical power and based on that estimate appropriate sample sizes (Fritz et al. 2012). The idea is that nonsignificant effects of independent variables in experiments are not necessarily the result of a missing relationship with the dependent variable but of low statistical power – a phenomenon called the dilemma of the nonrejected null hypothesis or simply Type II error. Significant effects may be found with quite small samples when the relationship between the manipulated variable and the dependent variable is strong and control is tight – that is, the effect size is large (Evans and Rooney 2013). However, small effect sizes require larger samples to achieve statistical significance. Experimenters make assumptions about the effect size by relying on previous research or on the smallest effect that would be meaningful. As an orientation, Cohen (1988) defines standardized effect sizes for common statistical tests. For instance, while for independent sample t-tests, he defines 0.20, 0.50, and 0.80 as small, medium, and large effect sizes, respectively, 0.10, 0.25, and 0.40 are the respective values for one-way ANOVAs. With respect to statistical power, a commonly suggested Type II error is 0.20, which indicates a power need of 0.80 (Ellis 2010). The sample size can then be estimated for a small, medium, or large anticipated effect size, considering the specified statistical power and significance level following Cohen's (1988) procedure. As an example, assuming a statistical power of 0.80 and a significance level of

0.05, the required total number of participants for an independent sample t-test (ANOVA with four group means) would be 786 (1096), 128 (180), and 52 (76) participants for a small, medium, and large effect size, respectively.

To obtain a pool of participants that constitutes a representative part of the total population of interest, two categories of sampling approaches can be applied: probability or nonprobability sampling (Evans and Rooney 2013). Probability sampling approaches subsume techniques where the likelihood of the selection of participants from a population is specified. The most prominent probability approach is *simple random sampling*. It assumes that all individuals of a population are included in the sample with the same probability. In an ideal case, all members of a population would be included in a list and then randomly picked until the desired sample size is reached. In contrast, *proportional random sampling* considers subgroups of the initial population, which are built first based on criteria such as age, gender, or income. Then, individuals are randomly selected from the respective subgroups. This technique is particularly useful when subgroups of the population such as women or older individuals represent the target segment. The third probability approach is *systematic random sampling*. Here, experimenters choose the first subject of a population and then every n th further subject is considered until the desired number of participants is reached. Finally, *multistage random sampling* combines at least two of the previously introduced approaches.

Nonprobability sampling approaches do not specify the likelihood of selecting an individual from the population, but select participants on the basis of accessibility or personal judgment (Evans and Rooney 2013). The most often used form of this sampling approach is *convenience sampling* (also known as haphazard or accidental sampling), where participants are selected based on simple recruitment. A common example in experimental research projects is the recruitment of students. Experimenters walk around the campus and ask them to participate in the respective experiment. This is convenient because it saves time and incentivizing students is less costly compared to other groups of a population.

Another nonprobability sampling approach is *volunteer sampling*. This technique describes the selection of participants from a group of volunteers and is used when it is difficult and unethical to randomly assign participants to the levels of a manipulation. This applies, for instance, to studies examining the effectiveness of pharmaceutical products. In an online context, opt-in panels are a popular volunteer sampling technique. People sign in or put their name on a mailing list to participate in experimental studies. Representatives of such opt-in panels are Amazon's Mechanical Turk (MTurk), CrowdFlower (CF), and Prolific Academic (ProA) (Peer et al. 2017). Particularly MTurk is increasingly drawing attention from practitioners and academics alike. It allows experimenters rapid data collection at low cost (ca. 10 cents per participant) and around the clock (Goodman et al. 2013). However, significant concerns exist regarding the use of MTurk in behavioral research. For instance, participants may pay less attention to instructions than traditional subject pools do because individuals are more likely to engage in distractions such as cell phone usage (Hauser and Schwarz 2016). Moreover, participants of MTurk may have different attitudes about money and time due to the low

compensation (Goodman et al. 2013). Those issues may cause a serious bias in the respective data. Given these concerns, several studies examined data generated by MTurk participants. The results are mixed in that some studies provide evidence for a higher risk-aversion and lower attention of MTurk participants in case of longer surveys (Goodman et al. 2013). Other research, however, demonstrated high accuracy and validity as well as a good replicability in behavioral outcomes (Buhrmester et al. 2011). This discussion indicates the need for a careful assessment of the suitability of MTurk participants for the underlying research question. Experimenters are recommended to “use screening procedures to measure participants’ attention levels and take into account that MTurk participants may vary from non-MTurk participants on social and financial traits” (Goodman et al. 2013, p. 222).

Finally, *quota sampling* describes a nonprobability sampling approach similar to convenience sampling, with one exception: participants with particular characteristics are selected until their proportion in the sample is large enough. For instance, this technique may be used if equal numbers of men and women are required.

Incentivization of Participants

The question what motivates individuals to participate in experiments is as old as experimental research itself. Besides altruistic (e.g., social obligations) and study-related (e.g., interesting topic) reasons, particularly egoistic motives have been shown to increase the willingness to contribute (Singer and Couper 2008). That is, individuals strive for a personal benefit when participating in experimental studies, which is why experimenters usually offer monetary and/or nonmonetary incentives (e.g., awarding of credit points or chance to win in a lottery drawing; Koschate-Fischer and Schandelmeier 2014). Such incentives have been shown to stimulate participation in experimental studies, especially when commitment to the task is low (Hansen 1980) and/or other motives are absent (Singer and Couper 2008).

Some aspects should be considered in this context. First, incentivization may not just impact the motivation of individuals to participate but also their performance in the experiment. An economic view holds that money generally lets individuals work harder, more persistently, and more effectively on a task, which improves the quality of the data received (Smith and Walker 1993). An increase in performance, however, has mostly been documented for mundane tasks such as memory or recall tasks, where financial incentives induce persistent diligence (Camerer and Hogarth 1999). In contrast, no effects on performance have been observed for more complex tasks. Even more problematic is that financial incentives may even decrease response quality – particularly in experimental settings where open-ended questions are asked. Here, the answers of incentivized participants were shorter and of lower quality compared to nonincentivized settings where participants contributed out of intrinsic motivation (Hansen 1980). Similarly, incentivization can produce inferior data quality in settings where incentives raise self-consciousness or may cause overreaction to feedback (Camerer 2011).

Second, one should consider that monetary and nonmonetary incentives are not equally attractive to potential participants. Generally, monetary incentives have been shown to be more effective to stimulate the willingness to participate in experiments

(Singer and Couper 2008). The amount of money offered, however, is not necessarily decisive for participation because participants see the monetary compensation more as a symbolic act (a thank you) than as a true income (Koschate-Fischer and Schandelmeyer 2014). This is why participants may continue participation in future research for even less than the current payment (Singer and Couper 2008). The amount should be appropriate for the task to be solved, which means that it should compensate for the physical and psychological risks involved. Nonmonetary incentives such as lotteries are attractive to individuals if there is a realistic chance of winning (Koschate-Fischer and Schandelmeyer 2014). Often, students are offered course credits as a compensation for their participation. This may be effective, but students may feel they have no other choice than participating to improve their grades. Meaningful extra credit options can reduce this coerciveness issue (Kalkoff et al. 2014). Similarly, the educational value of an experiment can be attractive to students. That is, they are more likely to participate in an experiment if they perceive that participation provides a valuable learning experience.

Finally, experimenters should be aware that response rates differ between incentives paid in advance (e.g., with the initial mailing) and those promised for the survey return (Singer et al. 1999). While both can be effective to increase response rates, prepaid incentives are more stimulating. This particularly holds for monetary incentives, where increases in response rates of more than 19% (compared to no incentivization) have been observed (Singer et al. 1999). In contrast, gifts as incentives accounted for increases of up to 8% only. With respect to promised incentives, charitable donations have been argued to be more effective to reduce costs and nonresponse bias than monetary incentives if the respective amount of money spent is low (Robertson and Bellenger 1978). This reasoning, however, could not be supported in an online context, where the monetary interest outweighed altruistic motives – particularly for longer studies (Deutskens et al. 2004).

Assigning Participants to Experimental Treatments

A fundamental issue pertaining to the design of experiments relates to the assignment of the experimental units (participants) to the different treatments. In a within-subject design, each participant of the experiment is (successively) exposed to multiple treatments, leading to multiple observations of the dependent variable from the same person. Thus, estimates of the causal effects of the treatments are obtained by measuring how the dependent variable changes with the successive exposure of the individual to the different treatments. In a between-subjects design, in contrast, each participant is (randomly) assigned to only one treatment combination, leading to one observation of the dependent variable per person. Hence, estimates of causal effects of the treatments are obtained by comparing the measure of the dependent variable between individuals of the different treatment groups.

Both approaches have their merits and weaknesses, and the application of one or another should be decided very carefully based on a number of factors. Consider, for instance, that the goal is to examine how far individuals infer quality from price. The hypothesis is that they indicate higher levels of perceived quality for the same type of product at a higher price as compared to a lower price. Examining this

question in a within-subject design would imply exposing the same individual repeatedly to the product stimulus at varying price levels. In a between-subjects design, it would imply the randomized assignment of individuals to either the high price condition or the low price condition and the respective measurement of perceived quality. Meta-analytical findings have repeatedly shown that in within-subject designs, the observed relationship between the product's price level and perceived quality is significantly stronger than in between-subjects designs (Völckner and Hofmann 2007). To provide guidance on the choice of design, we will subsequently elaborate on the statistical, theoretical, and psychological issues that may lead to differences between the two design options. It should also be noted that the designs can be combined into mixed designs – where one factor is manipulated between-subjects and another factor is manipulated within-subject – to profit from the advantages of both (Maxwell and Delaney 2004).

From a statistical viewpoint, within-subject designs result in data for two or more treatments per participant, whereas between-subjects designs yield data for only one treatment per participant. Moreover, each participant serves as his own control group when comparing the treatment effects in within-subject designs. Thus, internal validity does not depend on random assignment as in between-subjects designs since individual differences of the participants are removed from the error term. As a consequence, the number of participants needed to reach a certain level of statistical power is generally lower for within-subject designs as compared to between-subjects designs (Maxwell and Delaney 2004). Within-subject designs therefore are often employed in contexts where the costs per participant are relatively high and/or accessibility to the required infrastructure is limited. For instance, experimental studies that require functional magnetic resonance imaging (fMRI) mostly employ within-subject designs.

Also theoretical considerations may guide the choice of design. If the real-world phenomenon that is examined in an experimental setting can be described as *whether* at all to make a particular decision, a between-subjects design may be appropriate. A choice about *which* decision to make, however, is more akin to a within-subject design (Charness et al. 2012). Thus, the question is which design exhibits higher external validity in that it comes closer to the phenomenon as it unfolds in reality. For instance, a between-subjects design does not provide participants with a clear anchor or reference point. Going back to the price-perceived quality example, the phenomenon to be studied may be consumers' search behavior in a store, where the consumer is being exposed to different prices for products of the same category, and hence reference price effects are likely to occur. Alternatively, the researcher may be interested in the effect of a single advertisement where only one price cue is present (Völckner and Hofmann 2007). Whereas a within-subject design may provide higher external validity in the former scenario, the latter scenario may better map onto a between-subjects design. The results of the meta-analysis described above thus may suggest that price-quality inferences are less likely if no clear reference point exists.

The heaviest critique on within-subject designs is due to psychological issues inherent in this design (Greenwald 1976). A first phenomenon that can be observed

in some contexts is that participants simply become better in a task the more often they *practice* it. For instance, if the phenomenon to be studied involves the performance at a motor skill task (dependent variable) when being exposed to different types of distracting stimuli (treatment), participants in a within-subject design may become better in the task with every successive measurement simply due to practice of the task independent of the distracting stimuli. Such practice may confound the effect of the distraction treatment on task performance and thus threaten internal validity. A second phenomenon is the so-called *demand effect*: compared to a between-subjects design, participants in a within-subject design are more likely to guess the true purpose of the study and act in line with the inferred hypotheses. The more treatments a single participant is exposed to, the more likely is it that demand effects result (Greenwald 1976). For instance, successively exposing a participant to similar product stimuli with only price information being varied and asking this participant to indicate her or his quality perception of the product may indeed trigger thoughts on the purpose of the study. Finally, *carryover effects* may occur in a sense that the effect of a certain treatment persists over time and influences the subsequent measurement of the effect of another treatment. Such an effect may be observed if the effects of different drugs with unknown action times are examined in a within-subject design (Greenwald 1976; Maxwell and Delaney 2004). Moreover, individuals may use their evaluation of a prior treatment and transfer this prior judgment to a subsequent treatment if the prior judgment is accessible in memory and perceived as relevant and useful also in the new context (Lynch et al. 1988). This effect may occur when participants do not perceive that there is a substantial difference (e.g., new information) in subsequent treatments (Bornemann and Homburg 2011).

To counter some of these psychological issues, the order of the treatments in a within-subject design may be counterbalanced by randomly assigning participants to groups of equal size and presenting treatments to each group in a different order (e.g., with two treatments A and B, group 1 may first be given treatment A followed by B, while the reverse treatment order is administered to group 2). Maxwell and Delaney (2004) refer to such designs as *crossover designs*. To examine whether any effects caused by the order of the treatments exist, the factor “group” is included as a between-subjects factor in an analysis of variance. Nonsignificance of this “group” factor indicates that no order effects exist.

Preliminary Testing

As a final step in the design of an experimental study and before conducting the actual experiment, investigators have to conduct preliminary testing to ensure the adequacy of manipulations and measures (Perdue and Summers 1986). Weak experimental designs may make results unusable/uninterpretable and thus costly for researchers. This is especially the case for rather new research settings (Reynolds et al. 1993). Therefore, one should be able to modify potential shortcomings in advance and thus at a stage where corrections are less costly. Preliminary tests also make manipulation and confounding checks less necessary in the main experiment

(Perdue and Summers 1986). The only cases where skipping those preliminary analyses is acceptable relate to settings with very small target populations where at least some individuals of the preliminary test may equal those of the main experiment and to settings where the additional tests would adversely affect the main experiment (Reynolds et al. 1993).

Two types of preliminary testing may be considered. First, *pretesting* ensures the validity of aspects of the experiment such as instructions, tasks, and instruments. Participants of such pretests are instructed to evaluate these aspects in isolation, independent of the rest of the experimental design. Importantly, manipulation and confounding checks should be implemented in pretests to assess whether the manipulation adequately reflects the theoretical assumptions (Kuipers and Hysom 2014; Perdue and Summers 1986). Besides employing scale items, interviews with the participants or other qualitative techniques such as verbal protocols of scenarios or instructions can provide useful information on the credibility of the stimuli used. Moreover, pretesting can reveal issues related to missing response categories and the difficulty to answer particular questions (Reynolds et al. 1993).

In *pilot tests*, the second preliminary testing, the full experiment is provided to participants in situations comparable to those of the main experiment (Kuipers and Hysom 2014). Such pilot tests offer additional value to the experimenter since they provide information beyond individual parts of the experiment, including the measure of the dependent variable. Pilot testing can reveal whether there is enough variability in this measure. If this is not the case, the measure can be altered. Moreover, experimenters get information on readability and understandability of the instructions (e.g., logical flow), time required for completion, and the look of the design (Kuipers and Hysom 2014; Reynolds et al. 1993).

A general requirement for pre- and pilot testing is that participants should have the same characteristics as those targeted with the main experiment (Reynolds et al. 1993). This ensures that the adjustments made fit the requirements of this audience. Moreover, the same procedures and experimental instruments as in the main study are required to receive valid feedback for potential adjustments. Both the target population and the design of the instrument determine the sample size required for the preliminary analyses. The more subgroups of the total population to be considered and the more complex the experimental design (e.g., the more treatments), the more individuals are required. Usually, this sample size is rather small (5–10 to 50–100 participants; Reynolds et al. 1993). Finally, the feedback of participants should be captured directly after exposure to avoid any feedback bias that may result from later retrieval from memory (Perdue and Summers 1986).

Exemplary Experimental Study

We now illustrate the steps involved in the design of experiments with a real experiment that we conducted as part of a published research project (Albrecht et al. 2016). The project investigated the relevance of the interaction environment

for customer response to interactional service experiences and required the manipulation of two factors: the service experience and an environmental trigger.

Definition of Research Question. The starting point for this study was the observation that, in daily practice, buying behavior is affected by frontline employees' emotions as observed by the customer. However, little was known about how this relationship is influenced by the purchase environment – typically an important information source for customers to evaluate store experiences. Specifically, we intended to test our hypothesis that the presence of cues/triggers in the respective store environment that help explaining a given frontline employee's emotional display towards the customer may influence customer reactions. We expected that the impact of the presence of the environmental trigger would differently affect the customer response to positive versus negative emotions shown by the employee. Moreover, we were interested in the underlying psychological processes.

Determination and Operationalization of the Sources of Variation. Considering the *independent variables*, the experiment's objectives suggested two factors of interest: the emotional display of a frontline employee and the presence of a trigger in the store environment that may provide an explanation for the employee's emotion. The first factor, emotional display of the frontline employee, was varied at two levels: negative versus positive. The second factor, emotion trigger, also consisted of two levels: a control condition and a treatment condition, where an observable environmental trigger existed. In the control condition, no such emotion trigger was provided. We hence applied a full factorial design.

With respect to the *operationalization of the treatments*, a key challenge of our online experiment dealing with an offline phenomenon was to create a realistic purchase situation. To achieve that, we employed a scenario role-play-based approach and produced videotapes, one for each of the four treatments. Particularly, we hired a professional cinematographer and actor to create stimulus material in a local hardware store simulating a typical customer-employee interaction. The actor was instructed either to express the negative emotion of unfriendliness or to show the positive emotional display condition of smile. The customer was not shown explicitly but the camera represented the "eyes" of the customer such that each participant could put him/herself into the customer's shoes. In the emotion trigger condition, before the service interaction, the participant could hear the employee's phone ringing, see how the employee answered the call, and see and hear his reaction to the colleague on the phone. This reaction was either positive or negative in line with the emotional display manipulation. No such phone call trigger was provided in the control condition.

We applied *manipulation checks* to ensure that the service interaction scenario was perceived as realistic and that the manipulations worked as intended. In this respect, appropriate single/multi-item scales had already been validated in previous research. Therefore, we included those items as self-report measures in our study. These checks were administered after the assessment of the dependent variables to avoid any interference with the effect of the manipulation. Moreover, previous research suggested potential *extraneous factors* that we accounted for by assessing them as covariates: participants' susceptibility to catching emotions, preencounter mood, and age.

Definition and Operationalization of the Measured Response-Variables. With respect to the response variables, we assessed participants' purchase intention as a reaction to the employee's emotion. The examined process explanations comprised the perceived authenticity of the employee's emotional display and the perceived sympathy for this behavior. We decided to capture both the *dependent* and *mediator variables* with well-established self-report measures. We screened existing literature and identified multi-item scales that already worked well to capture purchase intention, perceived authenticity, and perceived sympathy. Participants rated the items on disagree-agree rating scales.

Decision about the Environmental Setting. We conducted the experiment online; hence, participants viewed the stimulus material on a computer monitor. We used the online setting because of cost advantages and the suitability of the stimulus material (the role-play video) for this type of environment. Alternatively, a laboratory setting would have been adequate as well.

Determination of the Experimental Units and Assignment to Treatments. The objective of the research project implies (adult) *customers* as the relevant experimental unit. We purposely did not refer to a student sample because of the above-mentioned homogeneity restrictions. We applied simple heuristics and set a minimum of 30 participants for each experimental condition. We recruited participants by posting the link to the study in relevant online communities and web forums with audiences from different social class, gender, and age categories. As such, we employed convenience sampling because we did not specify the likelihood of selecting an individual from the population, but selected participants based on accessibility in these online channels. In the end, we received 138 usable responses.

Participants were not paid an *incentive* but took part voluntarily. They were instructed to turn on the sound of their computer to be able to follow the videos. Moreover, we informed participants about the general purpose and procedure of the research and their right to decline participation or withdraw from the study. We told participants that they take part in a study on customer service. The experiment's objectives did not require deception – that is, we did not disguise any relevant information throughout the experimental study.

Given its multiple advantages and its easy application in online experiments, we applied randomization of the participants to the experimental conditions. Moreover, we chose a between-subjects design, where each participant was (randomly) assigned to only one of the four treatment combinations (negative/positive emotional display \times provision/nonprovision of the emotion trigger). We did so for theoretical considerations of realism. Typically, customers are confronted with a single employee showing either a positive or negative emotional display. Also, the trigger either exists or not, such that a mixed or within-subject design was not suited.

Preliminary Testing. We used different pretests to ensure the validity of our experimental manipulations. First, we provided the four videotapes to five doctoral students to ensure that participants had enough time to recognize all content of the videos. Second, we pretested the effectiveness of the manipulations and asked 246 persons to watch the videos and answer a few questions. Specifically, they were asked to rate the perceived unfriendliness and smiling of the frontline employee.

Finally, we provided the emotion trigger (cell phone call) sequence of the video to a university seminar class. Students watched the video sequence and wrote down their thoughts on what the talk was about and what the colleague on the phone had said to the frontline employee. We did so to ensure that the manipulation of the emotion trigger was perceived as “implicitly visible” to participants. All pretests confirmed the validity of our experimental manipulations.

Ethical Issues in Experimental Research

All research involving humans has to meet generally accepted ethical criteria to ensure the welfare of study participants and to protect them from physical and psychological harm. Building among others on regulatory requirements, the American Psychological Association (APA), for instance, has released the *Ethical Principles of Psychologists and Code of Conduct* (APA 2002) to provide guidance to researchers. The following aspects are part of the guidelines for research:

- *Institutional Approval*: Many organizations, particularly in the academic field, have created institutional review boards (IRBs) to protect the human dignity and welfare of participants of research projects. The IRB reviews research proposals submitted by researchers according to their conformance with ethical standards. Many academic journals now ask for such approval for submitted manuscripts.
- *Informed Consent*: Investigators are required to inform participants about the general purpose and procedure of the research and their right to decline participation or withdraw from the study.
- *Deception*: Investigators should generally refrain from deceiving participants unless the scientific value of the research is significant and the study cannot be realized without any deception. If unavoidable, such deception must not relate to aspects that may cause physical pain or severe emotional distress. After completion of the study, investigators are required to explain to participants the aspects involving deception and permit participants to withdraw their data.
- *Debriefing*: Investigators provide participants with the opportunity to obtain information about the results, conclusion, and purpose of the research and they correct potential misperceptions that participants may have.

While these aspects are relevant to all kinds of research involving humans, particularly issues related to deception are specific to experimental research since deception is sometimes used to ensure a high level of experimental control and to reduce the impact of extraneous factors.

Deception refers to the provision of false information or to withholding information to mislead participants into believing something that is not true (Hegtvedt 2014). Deception is distinct from the common practice to not fully inform participants about the hypotheses beforehand (e.g., through providing only partial information about the research question) to avoid demand effects (Hertwig and Ortmann 2008). An example of a rather serious form of deception is the provision of false feedback to

participants regarding their performance in a task they have completed, particularly if such feedback may affect their self-confidence in general (Hegtvedt 2014; Kuipers and Hysom 2014).

The controversy on the legitimacy of the use of deception has been quite intense, providing a long list of negative consequences of deception, such as embarrassment and a loss of self-esteem of participants at the individual level and resulting suspicion and negative attitudes towards research in general. This view is particularly prevalent among economists, who more or less generally reject deception. They also argue that if participants expect or are aware of deception, their behavior may no longer be shaped by the circumstances of the study (e.g., monetary rewards) but by psychological reactions to suspected manipulations (Hertwig and Ortmann 2008). Sieber (1992) argues that deception may be justifiable if (1) there is no other means to achieve stimulus control, if (2) responses to low-frequency events are studied, if (3) absolutely no risk of harm is associated with the deception, and if (4) the information would otherwise be unobtainable because of participants' anxiety or defensiveness. Pascual-Leone et al. (2010) offer a checklist that investigators may use to assess whether deception can be justified in a given context. As a general recommendation, investigators should employ such aids to determine whether there is really no way to avoid deception. If deception is used, it is important to conduct proper debriefing of participants to unravel the deceptive practice (Kuipers and Hysom 2014).

Conclusion

This chapter described the relevant steps involved when planning and executing experimental research in marketing. While experimentation is a central type of data collection in academic research in marketing, its use in corporate practice is still comparatively limited. Instead, companies nowadays embrace the blessings of big data analytics. However, the tremendous amount of historical data that companies create and collect poses challenges regarding the required data analysis skills, and not every company can afford to permanently employ the respective specialists. Experimentation, on the other hand, is technically relatively easy to implement and requires managers to directly focus on the causes and effects of interest instead of mining data that *might* provide useful insights. Specifically, the “test-and-learn” approach inherent in experimentation, where certain activities are directed towards one group of customers and other or no activities at all are directed to a control group, enables managers to develop a more direct feeling for relevant cause-effect relationships. The ease of implementation to a large extent depends on how easy the relevant outcomes can be assessed, which is why e-commerce and online business in general is at the forefront of corporate use of experimentation (remember the example of A/B testing from the introductory section). But also other businesses may easily implement and profit from experimentation (Anderson and Simester 2011). We hope that this chapter provides the necessary insights to accomplish such an endeavor.

References

- Aaker, D. A., Kumar, V., Day, G. S., & Leone, R. P. (2011). *Marketing research*. Hoboken: Wiley.
- Albrecht, C.-M., Hattula, S., Bornemann, T., & Hoyer, W. D. (2016). Customer response to interactional service experience: The role of interaction environment. *Journal of Service Management*, 27(5), 704–729.
- Albrecht, C.-M., Hattula, S., & Lehmann, D. R. (2017). The relationship between consumer shopping stress and purchase abandonment in task-oriented and recreation-oriented consumers. *Journal of the Academy of Marketing Science*, 45(5), 720–740.
- Anderson, E. T., & Simester, D. (2011). A step-by-step guide to smart business experiments. *Harvard Business Review*, 89(3), 98–105.
- APA. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, 57(12), 1060–1073.
- Arnold, V. (2008). *Advances in accounting behavioral research*. Bradford: Emerald Group Publishing.
- Baum, D., & Spann, M. (2011). Experimentelle Forschung im Marketing: Entwicklung und zukünftige Chancen. *Marketing – Zeitschrift für Forschung und Praxis*, 33(3), 179–191.
- Bearden, W. O., & Etzel, M. (1982). Reference group influence on product and brand decisions. *Journal of Consumer Research*, 9(April), 183–194.
- Benz, M., & Meier, S. (2008). Do people behave in experiments as in the field?—Evidence from donations. *Experimental Economics*, 11(3), 268–281.
- Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, 37(3), 245–257.
- Bornemann, T., & Homburg, C. (2011). Psychological distance and the dual role of price. *Journal of Consumer Research*, 38(3), 490–504.
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5.
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550–558.
- Camerer, C. F. (2011). *The promise and success of lab-field generalizability in experimental economics: A critical reply to levitt and list*. Available at SSRN 1977749.
- Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19(1), 7–42.
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8.
- Christian, B. (2012). The a/b test: Inside the technology that's changing the rules of business. http://www.wired.com/business/2012/04/ff_abtesting. Accessed 15 Mar 2018.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.
- Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods*, 14(3), 202–224.
- Cox, D. R. (1992). *Planning of experiments*. Hoboken: Wiley.
- Dean, A., Voss, D., & Draguljić, D. (2017). *Design and analysis of experiments*. Cham: Springer.
- Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, 15(1), 21–36.
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wikström, C., & Wold, S. (2008). *Design of experiments: Principles and applications*. Stockholm: Umetrics AB, Umeå Learnways AB.
- Evans, A. N., & Rooney, B. J. (2013). *Methods in psychological research*. Los Angeles: Sage.

- Falk, A., & Heckman, J. J. (2009). Lab experiments are a major source of knowledge in the social sciences. *Science*, 326(5952), 535–538.
- Feldman, J. M., & Lynch, J. G., Jr. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention and behavior. *Journal of Applied Psychology*, 73(3), 421–435.
- Festinger, L. A. (1957). *Theory of cognitive dissonance*. Stanford: Stanford University Press.
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2–18.
- Glasman, L. R., & Albarracín, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological Bulletin*, 132(5), 778–822.
- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3), 213–224.
- Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83(2), 314–320.
- Hakel, M. D., Ohnesorge, J. P., & Dunnette, M. D. (1970). Interviewer evaluations of job applicants' resumes as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. *Journal of Applied Psychology*, 54(1, Pt.1), 27–30.
- Hansen, R. A. (1980). A self-perception interpretation of the effect of monetary and nonmonetary incentives on mail survey respondent behavior. *Journal of Marketing Research*, 17(1), 77–83.
- Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., & Finkelstein, J. (2006). The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association*, 13(1), 16–23.
- Harrison, G. W., & List, J. A. (2003). *What constitutes a field experiment in economics? Working paper*. Columbia: Department of Economics, University of South Carolina <http://faculty.haas.berkeley.edu/hoteck/PAPERS/field.pdf>. Accessed 15 Mar 2018.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Hattula, J. D., Herzog, W., Dahl, D. W., & Reinecke, S. (2015). Managerial empathy facilitates egocentric predictions of consumer preferences. *Journal of Marketing Research*, 52(2), 235–252.
- Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1), 400–407.
- Hegtvedt, K. A. (2014). Ethics and experiments. In M. Webster Jr. & J. Sell (Eds.), *Laboratory experiments in the social sciences* (pp. 23–51). Amsterdam/Heidelberg: Elsevier.
- Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics and Behavior*, 18(1), 59–92.
- Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., & Weinmann, M. (2017). Inferring negative emotion from mouse cursor movements. *MIS Quarterly*, 41(1), 1–21.
- Horswill, M. S., & Coster, M. E. (2001). User-controlled photographic animations, photograph-based questions, and questionnaires: Three internet-based instruments for measuring drivers' risk-taking behavior. *Behavior Research Methods, Instruments, & Computers*, 33(1), 46–58.
- Kalkoff, W., Youngreen, R., Nath, L., & Lovaglia, M. J. (2014). Human participants in laboratory experiments in the social sciences. In M. Webster Jr. & J. Sell (Eds.), *Laboratory experiments in the social sciences* (pp. 127–144). Amsterdam/Heidelberg: Elsevier.
- Koschate-Fischer, N., & Schandelmeier, S. (2014). A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. *Journal of Business Economics*, 84(6), 793–826.
- Kuipers, K. J., & Hysom, S. J. (2014). Common problems and solutions in experiments. In M. Webster Jr. & J. Sell (Eds.), *Laboratory experiments in the social sciences* (pp. 127–144). Amsterdam/Heidelberg: Elsevier.
- Larsen, R. J., & Fredrickson, B. L. (1999). Measurement issues in emotion research. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: Foundations of hedonic psychology* (pp. 40–60). New York: Russell Sage.

- Laugwitz, B. (2001). *A web-experiment on colour harmony principles applied to computer user interface design*. Lengerich: Pabst Science.
- Levitt, S. D., & List, J. A. (2007). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics*, 40(2), 347–370.
- Li, J. Q., Rusmevichientong, P., Simester, D., Tsitsiklis, J. N., & Zoumpoulis, S. I. (2015). The value of field experiments. *Management Science*, 61(7), 1722–1740.
- List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *The Journal of Economic Perspectives*, 25(3), 3–15.
- Lynch, J. G. (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, 9(3), 225–239.
- Lynch, J. G., Marmorstein, H., & Weigold, M. F. (1988). Choices from sets including remembered brands: Use of recalled attributes and prior overall evaluations. *Journal of Consumer Research*, 15(2), 169–184.
- Madzharov, A. V., Block, L. G., & Morrin, M. (2015). The cool scent of power: Effects of ambient scent on consumer preferences and choice behavior. *Journal of Marketing*, 79(1), 83–96.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah: Lawrence Erlbaum Associates.
- Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research*, 44(5), 1157–1173.
- Mitra, A., & Lynch, J. G. (1995). Toward a reconciliation of market power and information theories of advertising effects on price elasticity. *Journal of Consumer Research*, 21(4), 644–659.
- Montgomery, D. C. (2009). *Design and analysis of experiments*. New York: Wiley.
- Morales, A. C., Amir, O., & Lee, L. (2017). Keeping it real in experimental research—Understanding when, where, and how to enhance realism and measure consumer behavior. *Journal of Consumer Research*, 44(2), 465–476.
- Morton, R. B., & Williams, K. C. (2010). *Experimental political science and the study of causality: From nature to the lab*. New York: Cambridge University Press.
- Myers, H., & Lumbers, M. (2008). Understanding older shoppers: A phenomenological investigation. *Journal of Consumer Marketing*, 25(5), 294–301.
- Nielsen, J. (2000). Why you only need to test with 5 users. <https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users>. Accessed 15 Mar 2018.
- Nielsen, J. (2012). How many test users in a usability study. <https://www.nngroup.com/articles/how-many-test-users>. Accessed 15 Mar 2018.
- Nisbett, R. E. (2015). *Mindware: Tools for smart thinking*. New York: Farrar, Straus and Giroux.
- Nordhielm, C. L. (2002). The influence of level of processing on advertising repetition effects. *Journal of Consumer Research*, 29(3), 371–382.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872.
- Pascual-Leone, A., Singh, T., & Scoboria, A. (2010). Using deception ethically: Practical research guidelines for researchers and reviewers. *Canadian Psychology*, 51(4), 241–248.
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163.
- Perdue, B. C., & Summers, J. O. (1986). Checking the success of manipulations in marketing experiments. *Journal of Marketing Research*, 23(4), 317–326.
- Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, 66(September), 29–38.
- Postmes, T., Spears, R., & Cihangir, S. (2001). Quality of decision making and group norms. *Journal of Personality and Social Psychology*, 80(6), 918–930.
- Rashotte, L. S., Webster, M., & Whitmeyer, J. M. (2005). Pretesting experimental instructions. *Sociological Methodology*, 35(1), 151–175.

- Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, 49(4), 243–256.
- Remler, D. K., & Van Ryzin, G. G. (2010). *Research methods in practice: Strategies for description and causation*. Thousand Oaks: Sage.
- Reynolds, N., Diamantopoulos, A., & Schlegelmilch, B. (1993). Pretesting in questionnaire design: A review of the literature and suggestions for further research. *Journal of the Market Research Society*, 35(2), 171–183.
- Robertson, D. H., & Bellenger, D. N. (1978). A new method of increasing mail survey responses: Contributions to charity. *Journal of Marketing Research*, 15(4), 632–633.
- Sawyer, A. G., & Ball, A. D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, 18(3), 275–290.
- Sawyer, A. G., Lynch, J. G., & Brinberg, D. L. (1995). A bayesian analysis of the information value of manipulation and confounding checks in theory tests. *Journal of Consumer Research*, 21(4), 581–595.
- Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51(3), 515–530.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Sieber, J. E. (1992). *Planning ethically responsible research: A guide for students and internal review boards*. Newbury Park: Sage.
- Simester, D. (2017). Field experiments in marketing. In E. Duflo & A. Banerjee (Eds.), *Handbook of economic field experiments* Amsterdam: North-Holland (pp. 465–497).
- Singer, E., & Couper, M. P. (2008). Do incentives exert undue influence on survey participation? Experimental evidence. *Journal of Empirical Research on Human Research Ethics*, 3(3), 49–56.
- Singer, E., Van Hoewyk, J., Gebler, N., & McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15(2), 217–230.
- Smith, V. L., & Walker, J. M. (1993). Rewards, experience and decision cost in first price auctions. *Economic Inquiry*, 31(2), 237–244.
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89(6), 845–851.
- Stuart, E. A., & Rubin, D. B. (2007). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155–176). New York: Thousand Oaks, CA: Sage.
- Thye, S. R. (2014). Logical and philosophical foundations of experimental research in the social sciences. In M. Webster Jr. & J. Sell (Eds.), *Laboratory experiments in the social sciences* (pp. 53–82). Amsterdam/Heidelberg: Elsevier.
- Trafimow, D., Leonhardt, J. M., Niculescu, M., & Payne, C. (2016). A method for evaluating and selecting field experiment locations. *Marketing Letters*, 7(3), 437–447.
- Trafimow, D., & Rice, S. (2009). What if social scientists had reviewed great scientific works of the past? *Perspectives on Psychological Science*, 4(1), 65–78.
- Verlegh, P. W. J., Schifferstein, H. N. J., & Wittink, D. R. (2002). Range and number-of-levels effects in derived and stated measures of attribute importance. *Marketing Letters*, 13(1), 41–52.
- Völkner, F., & Hofmann, J. (2007). The price-perceived quality relationship: A meta-analytic review and assessment of its determinants. *Marketing Letters*, 18(3), 181–196.
- Wetzel, C. G. (1977). Manipulation checks: A reply to kidd. *Representative Research in Social Psychology*, 8(2), 88–93.
- Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering baron and kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2), 197–206.
- Zikmund, W., & Babin, B. (2006). *Exploring marketing research*. Mason: Thomson South-Western.



Field Experiments

Veronica Valli, Florian Stahl, and Elea McDonnell Feit

Contents

Introduction	38
Motivation	38
Defining a Field Experiment	40
Experimentation: Causal Inference and Generalizability	44
Estimating the Causal Effect of a Treatment	44
Generalizability of Findings and External Validity	50
Sample Size	51
Experimental Design and Multivariate Experiments	53
Examples of Field Experiments	57
Case Studies	57
Conclusions	62
Cross-References	63
References	63

Abstract

Digitalization of value chains and company processes offers new opportunities to measure and control a firm's activities and to make a business more efficient by better understanding markets, competitors, and consumers' behaviors. Among other methodologies, field experiments conducted in online and offline environments are rapidly changing the way companies make business decisions. Simple A/B tests as well as more complex multivariate experiments are increasingly employed by managers to inform their marketing decisions.

V. Valli (✉) · F. Stahl
University of Mannheim, Mannheim, Germany
e-mail: veronica.valli@bwl.uni-mannheim.de; florian.stahl@bwl.uni-mannheim.de

E. M. Feit
LeBow College of Business, Drexel University, Philadelphia, PA, USA
e-mail: efeit@drexel.edu

This chapter explains why field experiments are a reliable way to reveal and to prove that a business action results in a desired outcome and provides guidelines on how to perform such experiments step by step covering issues such as randomization, sample selection, and data analysis. Various practical issues in the design of field experiments are covered with the main focus on causal inference and internal and external validity. We conclude the chapter with a practical case study as well as a brief literature review on recent published articles employing field experiments as a data collection method, providing the reader with a list of examples to consider and to refer to when conducting and designing a field experiment.

Keywords

Field experiment · A/B test · Randomized experiment · Online experiment · Digital experiment · Business optimization · Causal inference · Experimental design · Internal validity · External validity

Introduction

In God we trust, all others must bring data (Edward W. Deming¹).

Motivation

Digitalization of value chains and company processes offers new opportunities to measure and control a firm's activities and to make a business more efficient by better understanding markets, competitors, and consumers' behaviors. Among others, the advent of two main sets of methodologies is changing the way organizations do business in the current digital age:

1. *Big Data Analytics*: data mining, machine learning, and other statistical techniques allow practitioners to handle and analyze huge sets of data with a reasonable effort.
2. *Business Field Experiments*: studies conducted outside of the lab by means of easy-to-use software allow managers to reliably answer causality questions at reasonable costs. At the same time, field experiments have become a primary method for investigating scientific phenomena and that is why this chapter considers field experiments aimed at testing theories, of the same importance as those aimed at testing tactical business strategies.

¹Edward W. Deming was an eminent engineer, statistician, professor, and management consultant for more than half a century. His work on statistical process control and other strategies for data-driven decision making continues to be relevant today.

With the primary objective of informing marketing decisions, the fundamental value of market research is the collection, analysis, and interpretation of market-related information (Homburg et al. 2013). Depending on the objective, the research design can be exploratory, descriptive, or causal (Aaker et al. 2011). In particular, the *causal design* is the best approach to identify cause-effect relationships between variables based on preformulated hypotheses (Homburg 2015). Especially for practitioners, answers to the question “does A cause B?” are essential to derive managerial implications (Iacobucci and Churchill 2010), and, in such a context, an experiment is the most suitable and most popular method to establish causality (Crook et al. 2009; Homburg et al. 2013). For example, consider a marketer wishing to know the impact that a 20% discount will have on the proportion of customers making a purchase during a holiday sale. In such a case, comparing sales between a group of customers who were randomly chosen to be offered the discount and another group who was randomly assigned to not receive the offer will give a direct estimate of the incremental sales lift of the discount. For this reason, market researchers and other practitioners are increasingly making use of experiments in the field. Similarly academics have turned to field experiments, when once there was little experimentation outside of the lab. Field experiments are not only applied to inform almost every type of marketing decision (promotions, communications, visual designs, pricing, optimization of digital services, etc.) but also in disparate areas including business organization, product development, health care, human resource management, politics, and so on. As software tools and expertise grow, there are more and more A/B testing case studies showing that the practice of testing is becoming increasingly popular with small- and medium-sized businesses as well as with larger ones (see “A/B Testing Case Studies” on [Optimizely.com](https://www.optimizely.com) for many examples of online field experiments or Applied Predictive Technologies Case Studies on www.predictivetech.com for examples of offline field experiments).

The first field experiments in business practice date back to the first half of the 1900s when experiments revolutionized agriculture and created massive gains in farm productivity. Toward the end of that century, experiments became popular in manufacturing to improve production and quality. At their early stages, especially in firms that focused on product design and engineering, experiments were tremendously costly and often involved the destruction of expensive prototypes, such as in automotive crash testing. Nowadays, the digitalization of value chains has created a data-rich environment that offers both new challenges and new opportunities to managers, policy makers, and researchers, as also recognized in the recent (2014–2016) research priorities of the Marketing Science Institute (MSI). In such an environment, it is possible to measure market response at a much faster speed, allowing managers to track key economic parameters. These tracking skills allow companies to develop more effective business strategies to increase customer retention and loyalty or spending on products and/or services. This increased digitalization has also turned experiments into an economically feasible way to improve marketing decisions. Many marketers are embracing a *test and learn* philosophy with the aid of several platforms, such as Optimizely, Adobe Target, Applied Predictive Technologies (APT), Visual Website Optimizer

(VWO), Oracle Maxymiser, and Google Content Experiments, providing easy-to-use software to perform rigorous field experiments in the online and offline environments.

The primary scope of this chapter is to provide an answer to those readers who may be asking themselves: “why should I consider setting-up a field experiment to answer my research or business question?”

As a first answer, bear in mind the following hallmarks of well-designed field experiments:

- Field experiments are one of the most reliable ways to test a theory or to prove that a business action results in a desired outcome.
- Findings from field experiments have direct implications for business operations. In the language of experimentation, we say that they generalize well and have high external validity. On the other hand, lab experiments are acknowledged to have higher internal validity.
- Field experiments are easy to explain to business leaders and policy makers.

Throughout the following pages, we are going to explain each of the aforementioned points in depth advocating a major focus on business-related field experiments and online experiments (A/B tests).

Defining a Field Experiment

Field experimentation represents the conjunction of two methodological strategies: *experimentation* and *fieldwork*.

Defining an Experiment

Experimentation is a form of investigation in which units of observation are randomly assigned to treatment groups. Ex ante randomization ensures that the experimental groups have the same expected outcomes, which is fundamental to achieve an unbiased estimate of the causal effect of the treatment. Experimentation stands opposite to *observational investigations*, in which researchers attempt to draw inference from naturally occurring variations, as opposed to variations generated through random assignment (Gerber and Green 2008). However, some authors (e.g., Teele 2014) prefer to not exclude nonrandomized studies from the group of experiments, while others refer to studies without randomization as quasi-experiments (cf. Campbell and Stanley 1963).

An experiment involves the manipulation of the *independent* (or *explanatory*) variables in a systematic way which is then followed by the observation and measurement of the effect on the *dependent* (or *response*) variable, while any other variables that might affect the treatment are controlled or randomized over (Aaker et al. 2011; Iacobucci and Churchill 2010). For instance, in testing the impact of a 20% off promotion on sales, the researcher manipulates the independent variable

of promotion between the two levels of 20% and zero and measures customer purchases as the response variable.

From the perspective of Dunning (2012), true experiments (either in the lab or in the field) show three identifiable aspects:

1. The responses of experimental subjects assigned to receive one treatment are compared to the responses of subjects assigned to another treatment (often a control group which receives some type of baseline treatment that is essentially *no treatment* or the *state-of-the-art* condition). In the case of multivariate experiments, there are several treatment groups, which are all compared among each other.
2. The assignment of subjects to each group is done through a randomization device, such as a coin flip, a dice roll, or a digital algorithm.
3. The manipulation of the treatment is under the control of an experimental researcher.

Some *observational studies* share attribute number 1 of true experiments, in that treatment conditions' outcomes are compared. However, they do not share attributes number 2 and 3 as there is no randomization of treatment assignment and there is no treatment manipulation. On the other side, *natural experiments* share attribute 1 and partially attribute 2 since assignment is random or as-if random. However, in such cases, data comes from naturally occurring phenomena, and therefore the manipulation of treatment variables is not generally under the researcher's control. Natural experiments consider the treatment itself as an experiment and employ naturally occurring variations as a proxy for random assignment. In particular, the treatment is not assigned by a researcher but by some rule-based process that can be mathematically modeled (Teele 2014). Without it, other *confounder* variables could easily explain ex post differences between observed units (Dunning 2012).

Lab Versus Field Experiments

Depending on the setting employed, one can distinguish between laboratory and field experiments (Homburg 2015). In *laboratory experiments*, participants are tested in an environment which is created by the researcher and which thus differs from reality (Aaker et al. 2011). This unreal environment allows the experimenter to control other potential influences on the response but has the main drawback of making the respondent feel observed, which can lead to several kinds of response bias. In addition, the respondents who are willing to participate in a lab experiment may not represent the target population as a whole, and then findings might not be generalizable.

Outside of the lab environment, it is possible to run *field experiments*, in which the setting is an everyday life situation, often the exact same setting where the findings from the experiment will be deployed (Gerber and Greene 2012). In most field experiments, participants are not even conscious of taking part in an experiment (Aaker et al. 2011; Gneezy 2017) eliminating the risk of incurring a response bias.

Just as experiments are designed to test causal claims with minimal reliance on assumptions, experiments conducted in real-world settings are designed to make generalizations less dependent on assumptions (Gerber and Green 2012). Further, especially in digital environments such as websites, adequate sample sizes can be much more easily reached than in offline settings or labs, and randomization over large samples protects against the possibility that a variable other than the treatment is causing the response. Since the aim of this chapter is to provide a complete overview of the topic, a few issues discussed (e.g., issues related to causality, treatment effects, randomization, sources of bias, etc.) apply to experiments in general and therefore to both field and lab experiments. The reader will excuse the unavoidable overlap of some content with other chapters in this book.

Key Features of Field Experiments

Field experiments, either online or offline, can take many forms, but all have four key features that make them a field experiment: authenticity of treatments, representativeness of participants, real-world context, and relevant outcome measures. Indeed, the degree of fieldness of an experiment can vary dramatically; some field experiments may seem naturalistic on all dimensions, while others may be more dependent on assumptions. In a nutshell, what constitutes a field experiment depends on how the field itself is defined (Gerber and Green 2012). Harrison and List (2004) offer a classification system ranking field experiments depending on their degree of realism. The taxonomy they propose is based on six dimensions: (1) nature of the subject pool, (2) nature of the information that the subjects bring to the task, (3) nature of the commodity, (4) nature of the task, (5) nature of the stakes, and (6) nature of the environment that the subject operates in. Harrison and List (2004) propose the following terminology:

- The *conventional lab experiment* employs a convenient subject pool (typically students²), an abstract framing, and an imposed set of rules.
- The *artifactual field experiment* is akin to the lab experiment but involving a nonstandard (i.e., non-students) subject pool. With the term artifactual, the authors want to denote studies with an empirical approach that is artificial or synthetic in certain dimensions.
- The *framed field experiment* is akin to the artifactual field experiment but involving a realistic task and the natural environment of the tested subjects that are conscious of being tested. The term framed denotes the fact that the experiment is organized in the field context of the subjects (e.g., social experiments).
- The *natural field experiment* is akin to the framed field experiment involving the environment where subjects naturally undertake the tasks but with the subjects being unaware of participating in an experiment, that is, either online or offline depending on the nature of the setting under examination. Since participants in

²For an interesting discussion on the choice of participants for an experiment and the questionability of employing students, refer to Koschate-Fisher and Schandelmeier (2014).

this kind of experiments are a representative, randomly chosen, and non-self-selected subset of the treatment population of interest, the causal effect obtained from this type of experiment is the average causal effect for the full population, not for a nonrandom subset that chooses to participate (List 2011).

Online Experiments

Online experiments are a special form of field experiments and their simplest form is commonly referred to as *A/B test*. As shown in Fig. 1, this method involves random assignment of users to two different treatments, typically the current (or A) version and the new (or B) version (Kohavi et al. 2009). In particular, it involves the following steps:

- Randomly divide customers into groups.
- Expose each group to a different treatment.
- Measure one or more selected response variables (also called overall evaluation criteria or key performance indicators, such as conversion rates, click-through rate, revenues, etc.) for both groups.
- Compare groups by mean of data analysis to determine which treatment is better.

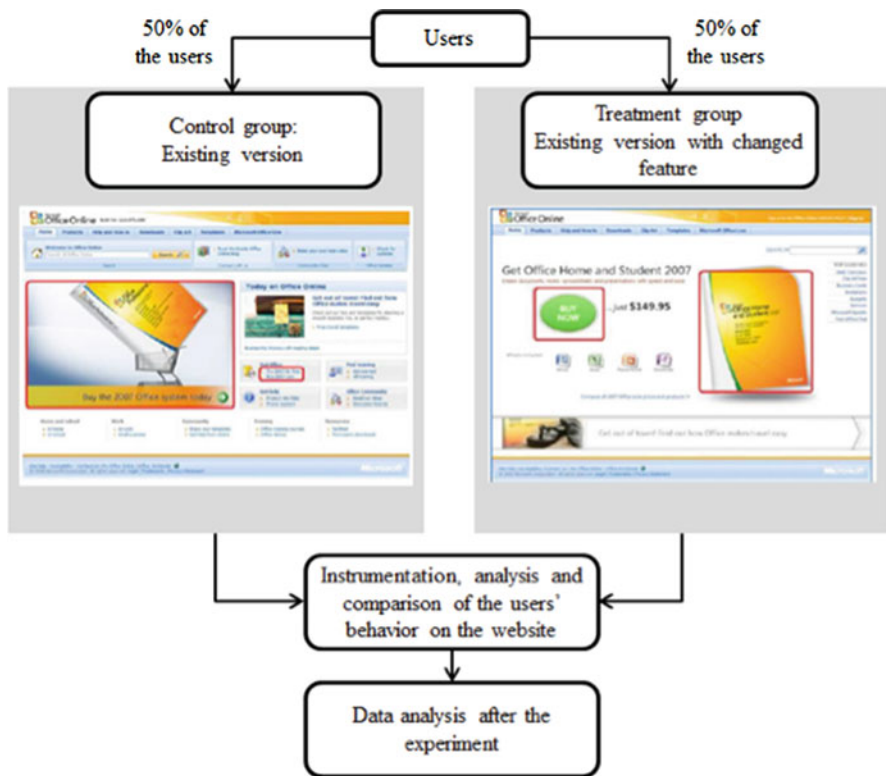


Fig. 1 Example of A/B test on Microsoft Office (Adapted from Crook et al. 2009 and Kohavi et al. 2009)

Field experiments did not start with digital marketing, and they are certainly not limited to digital marketing, but the digital environment has made testing easier and more popular as a way to inform managers' decisions. Managers are slowly accepting that carefully observing how customers respond to changes in marketing is more reliable than their experience and intuition.

Experimentation: Causal Inference and Generalizability

Whether it is a simple A/B test to choose the subject line for an email or a complex field experiment to test an economic theory, there are two main issues that the researcher must consider in designing an experiment. The first is whether the experiment has successfully measured the causal effect of the treatment within the context that the experiment is conducted (called *internal validity*). The second is whether the specific findings of the experiment can be generalized to other settings (called *external validity*). In this section, we discuss these two main issues in turn.

Estimating the Causal Effect of a Treatment

The Average Treatment Effect

Field experiments such as A/B tests allow managers to reveal the causal relationship between actions the company might take, such as price promotions (the *cause*), and consumers' purchase decisions (the *effect*). In other words, the goal of a field experiment is to determine whether a particular cause (such as a 20% price promotion) is responsible for an effect (such as a consumer's increased likelihood to purchase a particular product) and to exclude the reverse. Estimating the causal effect of an action has been a golden standard in the social sciences and in economic research for decades, and, as John List (2011) reminds us, economists have long worked on approaches that seek to separate cause and effect in naturally occurring data. For instance, instrumental variable regression aims at isolating cause-effect relationships. Field experiments use randomization as an instrumental variable, which, by construction, is uncorrelated with other variables that might affect the outcome (List 2011). However, there are a few key assumptions that must be met in order for experiments to provide reliable assessments of cause and effect (Gerber and Green 2012, Imbens and Rubin 2015). First, we provide a definition of *causal effect*: a causal effect is the difference between two potential outcomes, one in which a subject receives the treatment and the other in which the subject does not receive the treatment. In formulas:

$$\tau_i \equiv Y_i(1) - Y_i(0)$$

where (τ_i) is the causal effect of the treatment and $Y_i(1)$ is the potential outcome if the i th subject receives the treatment while $Y_i(0)$ is the potential outcome if the i th subject does not receive the treatment. For example, $Y_i(1)$ might be an indicator for whether the customer would make a purchase if she receives the promotion, and $Y_i(0)$ would be an indicator for whether the customer would make a purchase without the promotion.

Of course, it is typically not possible to directly observe both conditions for any given subject, but it is possible to estimate the average treatment effect (ATE) among all subjects, when certain assumptions are met. The ATE is defined as the sum of the subject-level treatment effects, $Y_i(1) - Y_i(0)$ divided by the total number of subjects. In formulas:

$$ATE \equiv \frac{1}{N} \sum_{i=1}^N \tau_i$$

The challenge in estimating the ATE is that at a given point in time, subject i is either treated or non-treated, and therefore either $Y_i(1)$ or $Y_i(0)$ is observed, but not both. Some statisticians conceptualize this as a missing data problem where either $Y_i(1)$ or $Y_i(0)$ is unobserved for each subject (Imbens and Rubin 2015).

Experiments, both in the lab and in the field, provide unbiased estimates of the ATE when the following assumptions are met (Gerber and Green 2012):

1. *Random assignment*: treatments are allocated such that all units have an equal probability between 0 and 1 of being assigned to the treatment group.
2. *Excludability*: the treatment must be defined clearly so that one can assess whether subjects are exposed to the intended treatment or to something else.
3. *Noninterference*: no matter which subjects the random assignment allocates to treatment or control, a given subject's potential outcomes remain the same.

Let us consider the three assumptions in more depth.

Random assignment is fundamental in experimentation, with roots that go back as far as Neyman (1923) and Fisher (1925). It implies that treatment assignments are statistically independent of the subjects' potential outcomes and addresses the missing data issue that challenges the estimate of the ATE, that is, the issue that at a given point in time, subject i is either treated or non-treated and therefore either $Y_i(1)$ or $Y_i(0)$ is observed, but not both. In fact, when treatments are allocated randomly, the treatment group is a random sample of the population in the experiment, and therefore the expected potential outcomes among subjects in the treatment group are identical to the average potential outcomes among the control group. Therefore, in expectation, the treatment group's potential outcomes are the same as the control group. When units are randomly assigned to treatment and control, a comparison of average outcomes in treatment and control groups, the so-called *difference-in-means* estimator, is an unbiased estimator of the ATE. In formulas, the estimator is:

$$\frac{1}{N} \sum_{i \in \text{treated}} Y_i - \frac{1}{M} \sum_{i \in \text{control}} Y_i$$

where N is the number of subjects in the treatment group and M is the number of subjects in the control group. We can see that the expected value of the estimator is equal to the ATE, meaning it is unbiased:

$$\begin{aligned} E \left[\frac{1}{N} \sum_{i \in \text{treated}} Y_i - \frac{1}{M} \sum_{i \in \text{control}} Y_i \right] &= E \left[\frac{1}{N} \sum_{i \in \text{treated}} Y_i \right] \\ - E \left[\frac{1}{M} \sum_{i \in \text{control}} Y_i \right] &= E[Y_i(1)] - E[Y_i(0)] = E[\tau_i] = ATE \end{aligned}$$

When random assignment is not used, there is always potential for a *selection bias*, where the treatment assignment is systematically related to potential outcomes. For example, if we want to measure the effect of a call from a sales agent and we do not randomize calls between customers, the sales agent may choose to call those customers that she/he feels are most likely to buy. This will produce an upward bias in our estimate of the ATE. The key idea is that randomized assignment allows us to use simple averages of the outcome for the treatment and control group to estimate the average treatment effect.

Excludability refers to the fact that each potential outcome depends *solely* on whether the subject *itself* receives the treatment and not on some other feature of the experiment. Therefore, when conducting an experiment, we must define the treatment and distinguish it from other factors with which it may be correlated. Specifically, we must distinguish between d_i , the treatment, and z_i , a variable that indicates which observations have been allocated to treatment or control. We seek to estimate the effect of d_i , and we assume that the treatment assignment z_i has no effect on the outcomes. In other words, the exclusion restriction refers to the assumption that z_i can be omitted from the potential outcomes for $Y_i(1)$ and $Y_i(0)$, and this restriction fails when random assignment sets in motion causes of Y_i other than the treatment. In real life, and therefore in field experiments in particular, it can become difficult to ensure excludability. Consider, for example, an A/B test investigating the impact of a discount on purchase decisions. If being assigned to receive a discount also means that the customer will get an email and customers in the treatment group do not get an email, then the excludability assumption is not met, and any observed difference between the treatment and control groups may be due to the email and not to the discount. A straightforward example of a research design that attempts to isolate a specific cause is a pharmaceutical trial in which the treatment group is given an experimental pill while the control group is given an esthetically identical sugar pill. The aim of administering a pill to both groups is to isolate the pharmacological effects of the ingredients, holding constant the effect of merely taking some sort of pill (*placebo effect*).

How can we make sure that the excludability assumption is met and that we are able to isolate the specific cause we intend to? Basically, by ensuring uniform handling of treatment and control groups, for instance, with double blindness, neither the subjects nor the researchers charged with measuring outcomes are aware of which treatments the subjects receive, so that they cannot consciously or unconsciously distort the results. Another procedure is parallelism when administering an experiment: the same procedures should be used for both treatment and control groups, and both groups' outcomes should be gathered at approximately the same time and under similar conditions (Gerber and Green 2012). In online experiments, meeting excludability assumptions might seem easier; however, consider, for instance, a test of several versions of the same webpage showing prices and promotions for a given brand. Randomization algorithms ensure that different customers shopping from different laptops and IP addresses see different versions. But if, unluckily, two people sitting one next to the other and surfing the same webpage from different terminals but in the same location see different versions (having been assigned to different treatment groups), we might incur in a violation of the exclusion restriction, as recognizing the different versions can confound the causal effect we set out to estimate. In such cases, precise geolocation and a randomization procedure that considers such geographical information could help solve the problem.

Noninterference refers to the fact that potential outcomes are defined over the set of treatments that the subject *itself* receives, not the treatments assigned to other subjects. This assumption is sometimes called the Stable Unit Treatment Value Assumption (SUTVA). Considering that each observational unit is either treated or not treated, the number of potential outcomes to take into account can quickly increase if we allow the outcome for subject i to depend on the treatment assignment of another subject j . The noninterference assumption cuts through this complexity by assuming that the outcome for i is not affected by the treatment of other subjects (Gerber and Green 2012; Imbens and Rubin 2015). Consider, for instance, when an A/B test is conducted on an e-commerce website offering promotions to a targeted subsample of existing customers and not to some others. Noninterference would assume that purchase decisions of subject i were only affected by his/her personal assignment to treatment or control group. But what if, for instance, two subjects belonging to the same household, say two sisters, are shopping from the same website and one falls into the treatment but the other one falls into the control? Then, we might have violation of noninterference as the treatment received by one sister can affect the other that therefore no longer constitutes an untreated control group. To prevent this from happening, researchers should try to design experiments in ways that minimize interference between units by spreading them out temporally or geographically or to design experiments in ways that allow researchers to detect spillover between units. Instead of treating interference as a nuisance, these more complex experimental designs aim to detect evidence of communication or strategic interaction among units.

Causality and Internal Validity

The previous section described the issues involved in estimating the causal effects as they are typically discussed in economics (Gerber and Green 2012) and statistics (Imbens and Rubin 2015). Psychologists also have a rich tradition of describing problems that can occur in experiments and have coined the term *internal validity* which refers to the extent to which we can say the observed effect in our study was caused by our treatment (Campbell 1957; Campbell and Stanley 1963; Shadish et al. 2002). Many of the ideas in this section are closely related to the previous discussion of the conditions necessary to estimate the causal average treatment effects, but using a different set of terms. Since both perspectives on experiments are common in marketing, we present both.

To achieve high internal validity, laboratory experiments are generally more suitable. This is because the controlled environment allows for better control of confounders. However, depending on the field considered, the natural environment can be highly controlled as well, especially in digital settings. In general, when considering studies that go beyond the randomized controlled experiment, there are many threats to internal validity, some of which we have discussed previously and most of which apply to both field and lab experiments:

- *Selection bias*: when assignment to treatment is not random and certain types of people are more likely to receive one of the treatments, in other words the experimental groups systematically differ from each other either because of self-selection (e.g., by voluntarily choosing whether to receive the treatment) or by incorrect assignment (Campbell 1957; Iacobucci and Churchill 2010; Shadish et al. 2002). For example, when running an offline field experiment to test the effect of marketing actions on purchase intentions, a selection bias could emerge due to self-selection of respondents into treatments. When treatments are not randomly assigned, the subjects or the experimenters may assign certain types of subjects to treatment and other types to the control. For example, if we are studying the effect of receiving emails on customer's purchase rate using observational data collected by the company, we have to consider that customers get to self-select whether to sign up for the mailing list, and so those who sign up may be systematically more likely to purchase than those who do not sign up. This is less likely to happen in online field experiments, as assignment to treatment groups is handled by the computer systems and mostly unnoticed by users who are often completely unaware of being tested.
- *Differential attrition*: when certain types of subjects drop out of one of the treatments. It implies that certain types of participants leave during the run of the experiment or do not take part in the final measurement (Aaker et al. 2011; Shadish et al. 2002), and this attrition is different for the treatment and the control groups. For instance, if you were testing an increase in the frequency of direct marketing, customers who have less affinity for the brand may be more likely to ask to be put on a "do not call" list when they are in the high-frequency condition. These participants would not complete the treatment and so typically would not

be counted in the analysis of the response. The direct consequence of differential attrition is that the average of the experimental group might differ if the exited participants were still involved (Iacobucci and Churchill 2010; Shadish et al. 2002).

- *Time effects*: when treatments are administered at two different times, outside events, learning, or other changes are confounded with the treatment (Shadish et al. 2002).
- *Confounding variables*: when other variables are correlated with the treatment and have an effect on the outcome, a cause-effect relationship between the confounder and the dependent variable can be mistakenly assumed to be a causal effect of the treatment.
- *Noncompliance*: subjects assigned to the experiment do not get the specified treatment. This can happen because of individuals' voluntary decision to use a different treatment than the one they were assigned, because they do not like it or they think another treatment would be better.
- *Diffusion of the treatment across groups*: subjects assigned to one treatment find out about the other treatment.
- *Demand effects*: participants guess the hypothesis of the experiment and try to cooperate by exhibiting behavior that confirms the hypothesis.
- *Experimenter bias*: experimenter makes subjective measurements and inadvertently favors the hypothesis in those measurements. An experimenter bias may exist when the mere presence or interaction with the interviewer has an effect on the respondent's responses. Being interviewed about personal purchase intentions might arouse a sense of self-exposure that could lead to biased responses not reflecting the private true intentions. This is more often the case in face-to-face interviews and is quite unlikely to happen in lab experiments or in online field experiments.
- *Hawthorne effect*: it is also possible that individuals being part of an experiment and being monitored change their behavior due to the attention they are receiving from researchers rather than because of the manipulation of the independent variables. The Hawthorne effect was first described in the 1950s by researcher Henry A. Landsberger during his analysis of experiments conducted during the 1920s and 1930s at the Hawthorne works electric company in Illinois. His findings suggested that the novelty of being research subjects and the increased attention deriving from this could lead to temporary increases in workers' productivity. This is sometimes also referred to as the *John Henry effect* and is closely related to the *placebo effect* in medicine. This issue is easily overcome in many field experiments where subjects are unaware of being a subject in a test but is more likely to happen in lab experiments (Landsberger 1958).
- *Ambiguous temporal precedence*: In some experiments, it can be unclear whether the treatment was administered before or after the effect was measured. For instance, if purchases and promotional emails are tracked at a daily level, it can be difficult to discern if a customer who received an email on a particular day and also made a purchase that same day received the email before she made the

purchase. If the treatment does not occur before the outcome is measured, then the causality may be reversed.

Generalizability of Findings and External Validity

Often, we are interested in whether the conclusions of our experiment can be applied to a specific business decision. For instance, if we test a new product display in 30 stores within a chain and find that the new product display increases sales, then we want to know whether this finding will generalize to other stores in the chain or to other retailers. *External validity* refers to the extent to which the specific findings of the experiment can be generalized to other target populations or other similar situations (Campbell 1957; Shadish et al. 2002). If the study shows high external validity, we can say that the results can be *generalized*. Field experiments are largely acknowledged to better generalize to real situations than lab experiments because of the real setting in which they are deployed, although some have cautioned that field experiments conducted in one setting cannot always be generalized to other settings (Gneezy 2017).

The major threat to external validity is that some idiosyncrasy of the test situation (*context effect*) produced the effect, but the effect goes away in the target business environment. For instance, while an ad may perform well in a copy test where customers are brought into a lab setting and exposed to the ad and then surveyed on their purchase intent, those results may not generalize to ad exposures in the real world, perhaps because people do not pay as much attention to ads in the real world as they do in the lab. Or a finding from a field experiment showing that price promotions increase sales of packaged goods may not extend to a different product category. For those familiar with regression, another way to conceptualize context effects is that there is an interaction between the treatment and some context variable that was held fixed in the experiment, such that the effect of the treatment is different depending on the value of that context variable (Campbell and Stanley 1963).

Another key element in designing an experiment with good external validity is determining which subjects to include in the experiment. Note that the assignment of subjects to treatments is closely related to the internal validity of the test, while the selection of subjects to include in the experiment is closely related to the external validity. The best way to enhance external validity is to test the research hypotheses on the entire population that the researcher hopes to learn about, e.g., all the customers in a CRM system or all the stores in a chain. This approach also maximizes the power of the test to detect differences between treatments, which we will discuss in the next section. Obviously, this is rarely possible outside of some digital marketing contexts either because of the high costs of applying treatments and measuring outcomes and/or the riskiness of the treatment.

To reduce risks and costs, researchers frequently rely on samples of subjects from the target population. Some sampling strategies that are available to use are (from ideal to worst):

- *Simple random sample*: take a random draw from the target population using, for instance, a coin flip or a dice roll. This gives to each subject an identical probability of entering the sample, ensuring that the sample will be representative of the target population.
- *Cluster sample*: when it is easy to measure groups or clusters of subjects, randomly sample from among the clusters.
- *Stratified sample*: use a procedure to make sure that the sample contains different types of subjects.
- *Convenience sample*: sample in some way that is easy for the researcher, e.g., an academic might conduct the experiment with students or a company might conduct the experiment using store locations that are nearby.

For instance, if a publishing company wants to evaluate whether a given promotion strategy works better than another and decides to run a field experiment, they have to consider the target population from which to sample. If their goal is to learn how their current customers respond, they might focus on customers from their current mailing list. However, if they hope to learn about how *potential* customers respond to the promotions, they might choose to sample customers from a larger list of avid readers. In either case, once the target population is identified, the ideal strategy for selecting a group of customers to include in the experiment is to either use all the customers in the target population, assigning some to treatment and some to control, or to select smaller treatment and control groups randomly from the mailing list. The simple random sample ensures that the subjects in the study represent the target population. A convenience sample, by contrast, may not properly represent the target population; for example, students may not behave in the same way as other types of customers. If the company plans to study separate subgroups within the target population, they may find a stratified sample useful for ensuring that there is sufficient sample size within each subgroup. Another potential threat to generalizability is the representativeness of the subjects in the test. A common criticism of experiments conducted with students, for instance, through surveys or lab experiments, is that the results may not reliably extend to the entire population of reference. Similarly, in online experiments the researcher should keep in mind that mostly heavy users of the website or app are more likely to be included in field experiments than light users. Most online tests include in the sample all the visitors in a fixed period, and this group will naturally include more frequent users than infrequent users. To overcome such issues, companies should consider test designs that assign treatments to users (rather than to sessions), track users across visits, and cap the number of times each user is exposed to the treatment.

Sample Size

A key question in designing any experiment is determining how many subjects to include in the test. Sample sizes for an A/B test are typically determined by considering

the hypothesis test comparing the two groups. The typical A/B test in marketing estimates the average treatment effects by comparing the proportions of people who respond to two different stimuli. Following the traditional one-tailed test for comparing proportions, we begin with a null hypothesis that the proportion of people who respond will be the same in both groups versus an alternative that the A group responds in greater proportion than the B group:

$$\begin{aligned} H_0 &= \pi_A = \pi_B = \pi \\ H_1 &= \pi_A - \pi_B = \delta > 0 \end{aligned}$$

Our goal is to plan the number of subjects to include in the treatment and control groups so that we will be able to correctly retain the null hypothesis if there is no difference between treatments and reject the null if there is a difference of at least δ . In the extreme, if we have no subjects, then we clearly will always retain the null hypothesis no matter what. There are four aspects of the experiment that influence the expected required sample size for an A/B test:

- The expected proportion π
- The expected (minimum) difference between the two groups δ
- The desired confidence $1-\alpha$ (where α is the significance)
- The desired power $1-\beta$

The *confidence* is the likelihood that you will retain the null hypothesis and decide that there is no difference when there really is no difference. *Power* is the likelihood that you will reject the null and detect a difference when indeed there is a difference of at least δ . Both should be considered carefully in the design of an experiment. Consider, for example, an A/B test designed to determine the effect of an ad on the proportion of people who buy. In this case, we want high confidence to prevent the possibility of concluding that the ad has a positive effect when it, in fact, does not. We also want high power, to prevent concluding that that the ad does not work when, in fact, it does. For a given sample size, power and confidence can be traded off. Lewis and Rao (2015) find that for display advertisements, even A/B tests with very large sample size conducted at a traditional confidence level of 0.95 do not have sufficient power to detect whether an ad has positive ROI. Thus, it is critical to consider power when planning an A/B test.

The sample size for each group in a comparative A/B test can be accurately estimated by (Ledolter and Swersey 2007):

$$N \approx \frac{2\pi(1 - \pi) [z_{1-\alpha} + z_{1-\beta}]^2}{\delta^2}$$

where z_x is the cumulative normal distribution evaluated at x . This can be computed, for example, using the Excel formulas: $z_{1-\alpha} = NORM.S.INV(1 - \alpha)$ and $z_{1-\beta} = NORM.S.INV(1 - \beta)$.

One can see from this formula that if the researcher wants to detect a small difference, δ , in the response rate between the A and B groups, then a larger sample

size is required. Similarly, if the researcher wishes to reduce the chance of an erroneous conclusion (i.e., that there is a difference when there is not or that there is not a difference when there is), then $z_{1-\alpha}$ and $z_{1-\beta}$ will be larger and the required sample sizes will be higher.

Note that this formula depends on the size of the difference that the marketer wishes to detect. In practice, it is very important to consider δ carefully. When a very large amount of data is available (for instance, from e-commerce websites), generating large datasets and big samples is much easier than few years ago. In such cases, it might happen that very negligible effects become significant (e.g., WTP is \$10 in treatment group and \$9.99 in control). While this effect is statistically significant, it does not really tell much about our business/research question and may not be useful for making decisions. So, in situations where N is not limited by the budget, it may be sensible to choose a smaller N so that the difference to detect, δ , is a difference that would be meaningful to the business. This is sometimes referred to as aligning practical and statistical significance.

Experimental Design and Multivariate Experiments

Managers frequently want to measure the effect of several different marketing actions (i.e., they are interested in more than one treatment). For instance, a publisher might be interested in assessing how different discount levels perform in combination with different ways of communicating the discount. They might be interested in measuring the effect of two levels of discount (say 5€ and 10€) while at the same time understanding the effect of communicating the price reduction in terms of price discount (e.g., “subscribe for one month and save x €!”) or in terms of bonus time (e.g., “subscribe for 1 month and get x weeks free!”). A multivariate experiment can be used to simultaneously measure the effect of the discount level and the message type while also determining if there is any additional effect of combining two treatments together. When the combined effect of two treatments is better than the sum of the individual effects, there is an *interaction* effect. Detecting interactions is the main reason why companies conduct multivariate tests. In addition, multivariate tests can reduce required sample sizes and increase the amount that can be learned in the time frame of a single test.

Before approaching the technicalities of multivariate testing, we define some useful terminology. The *factors* are those variables (continuous or categorical) whose effect we want to study, e.g., ad copy, font, photo, and color in an advertisement or seed type, fertilizer, and amount of water for an agricultural experiment. In the experiment, each factor is tested at multiple *levels*, the different versions we want to test. The simplest A/B test comparing two treatments has 1 factor with two levels.

Multivariate tests are experiments where two or more factors are tested. Multivariate tests should be carried out when the researcher wants to know the relative effects of the different factors or when there might be combinations of levels that perform especially well together. If the effect of the two factors together is more (or less) than the sum of their separate effects, we say the two factors interact with each other. For instance, the text color and the background color of a call-to-action

button typically interact: when the colors are the same, customers cannot read the button and do not respond.

For a better understanding of multivariate experiments, consider the following experiment (adapted from Ledolter and Swersey 2007) that was conducted by a credit card company who wanted to increase the response rate, that is, the number of people who respond to a credit card offer. The marketing team decided to study the effects of interest rates and fees, using the four factors shown in the following table.

	Factor	Level 1 (-)	Level 2 (+)
A	Annual fee	Current	Lower
B	Account-opening fee	No	Yes
C	Initial interest rate	Current	Lower
D	Long-term interest rate	Low	High

We could choose to study these factors with a series of A/B tests. Suppose we all agree that factor A (annual fee) is likely to be most important. Then we can run an A/B test on annual fee, holding the other factors at the control levels. The combination of factors and levels is clearly summarized in the following *design matrix*:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
1	-	-	-	-	20,000
2	+	-	-	-	20,000

Suppose our first test found that the lower annual fee increased the response rate. So, we can fix the factor A to “+” and in our next A/B test, we can look at factor B:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
3	+	-	-	-	20,000
4	+	+	-	-	20,000

Putting a sequence of these A/B tests together, we might end up with the following runs:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
1	-	-	-	-	20,000
2	+	-	-	-	20,000
3	+	-	-	-	20,000
4	+	+	-	-	20,000
5	+	-	-	-	20,000
6	+	-	+	-	20,000
7	+	-	-	-	20,000
8	+	-	-	+	20,000

Looking back at the resulting set of runs, we might notice several serious problems:

- Before we run the first A/B tests, we do not really know which factor is most influential, so it is difficult to know where to start.
- We could be wasting time with the sequential process.
- We are sometimes running the same condition more than once, which is inefficient (runs 2, 3, 5, and 7 are all the same).
- Because we have not tested all combinations of factors, we have little information about the interactions between factors.
- If there are interactions, testing the factors in a different sequence could lead to different conclusions about which combination is best.

To overcome these issues, it is recommended to make use of a proper experimental design (commonly referred to as design of experiment, or DOE). In this example, a better approach creates a single test that includes every possible combination of levels (*full factorial design*) which allows us to see if there are certain combinations of factors which are particularly good and to reduce the sample sizes for each run. The full factorial design matrix, in this case, looks like this:

Run	A Annual fee	B Account-opening fee	C Initial interest rate	D Long-term interest rate	Sample
1	–	–	–	–	7500
2	+	–	–	–	7500
3	–	+	–	–	7500
4	+	+	–	–	7500
5	–	–	+	–	7500
6	+	–	+	–	7500
7	–	+	+	–	7500
8	+	+	+	–	7500
9	–	–	–	+	7500
10	+	–	–	+	7500
11	–	+	–	+	7500
12	+	+	–	+	7500
13	–	–	+	+	7500
14	+	–	+	+	7500
15	–	+	+	+	7500
16	+	+	+	+	7500

Note that the number of possible combinations for a design can be computed by multiplying together the number of levels (2) for each of the four factors ($2 \times 2 \times 2 \times 2 = 2^4 = 16$ combinations). It has become common to describe an experiment with multiple factors using this shorthand. For example, a $2^3 \times 5^1$ full factorial experiment has three factors that have two levels and one factor that has five levels, which is a total of 40 different combinations of the factors.

A full factorial design allows us to estimate the *main effects* of the factors and all *interactions* between factors. The *main effect* of a factor is defined as the change in the response variable when the level of the factor is changed from low to high and corresponds to the average treatment effect for an A/B test that we discussed in “[Motivation](#)”. For a full factorial design, we can compute the main effect, by averaging the response rate across the runs when the level is at the high level and comparing that to the average across the runs at the low level. A two-way *interaction* occurs when the effect of one factor depends on the level of another factor (e.g., does the impact of having an annual fee depend on whether or not there is an account-opening fee?). *Three-* and *four-way interactions* are similar to two-way interactions, but are difficult to think about (e.g., is the effect of C different when both A and B are at their high levels?). Luckily, those higher-order interactions are usually negligible in most business settings. To estimate main effects and interactions for multivariate experiments, most researchers use *regression analysis*, fitting a model that relates the outcome measure to the various factors. If the subjects in a multivariate test are assigned to conditions randomly, the estimates of main effects and interactions that we get from this regression represent the *causal* effect of those treatments, just as in single-factor experiments.

In the example above, we show a full factorial test, where all the possible combinations of factors are tested. However, as the number of factors increases, the number of combinations increases rapidly. Therefore, researchers who use multivariate tests frequently spend a lot of time thinking about which combinations of factors they should include in their experiment and which they can leave out. One approach is *fractional factorial* design, which reduces the number of combinations to a half or a quarter of the possible combinations, by eliminating the possibility of estimating high-order (three-way and higher) interactions. A newer approach for determining which combinations of factors to include in a multivariate test is *optimal design*, which characterizes how much we learn from an experiment by considering how precisely we will be able to estimate the parameters of our regression model. Optimal designs choose the design matrix so as to get the best possible standard errors and covariance matrix for the parameter estimates. (See Goos and Jones [2011](#) for more detail.) Optimal design typically requires specialized software (e.g., JMP from SAS or the AlgDesign package in R) where the user inputs the factors and levels and the software finds the best combination of factors to test.

An important feature of good multivariate experimental designs is *orthogonality*. When two variables are orthogonal in an experiment, it means that the various combinations of the two factors occur exactly the same number of times. A nice property of orthogonal design is that the estimate of the effect of one factor will not depend on whether or not the other factor is controlled for in the regression. When the two factors are always set at the same level (e.g., the account opening fee is always paired with the annual fee), it is impossible to estimate separate effects for each factor, and this is called a *confound* in the multivariate design, which is the opposite of orthogonality. Full and fractional factorial designs maintain orthogonality, while optimal designs are not necessarily orthogonal, but are usually nearly orthogonal.

One common application of multivariate testing in marketing is in testing various features of direct mail offers: from the color of the envelope to the celebrity

endorser's appeal. In this type of experiment, the direct marketer typically sends out a number of different direct marketing offers with varying levels of the features and then measures the number of customers who respond. In this context, additional cost is incurred for each different version of the mailing, and so fractional factorial and optimal design approaches, which reduce the number of required combinations, are valuable. Applying an optimal design or an orthogonal, fractional factorial design instead of a one-factor-at-a-time method increases the efficiency at evaluating the effects and possible interactions of several factors (independent variables).

Another important application of multivariate experimental design is *conjoint analysis*. In conjoint analysis, customers are asked to evaluate or to choose from a set of hypothetical products, where the products vary along a set of features. These product features become the factors in a multivariate experimental design. A common approach to creating the questions to include in a conjoint survey is to use optimal design (Sándor and Wedel 2001).

Examples of Field Experiments

Case Studies

Field Experiments in Business

Field experiments are rapidly becoming an important part of business practice, and many marketing-oriented firms now employ a testing manager, who is responsible for designing, executing, and reporting on field experiments to answer important questions. These testing managers often specialize in a particular part of the business or communication channel. For instance, one might find different specialists in website testing, email testing, and direct marketing experiments, all within the same company. Regardless of the specific platform, the goal of these testing managers is to find treatments to test, to determine how to measure the response to the treatments, to ensure that the test is designed so that it can be interpreted causally, and to analyze and report on the results. In the next subsection, we describe the testing program employed by the donation platform for the 2012 US presidential campaign for Barack Obama.

A major focus for the 2012 US presidential campaigns was fundraising. Several changes in regulation had made donations to political campaigns more important than ever, and so there was a major focus on the web platform where potential donors were encouraged to make small- and medium-sized donations. In their ongoing efforts to improve the platform, the team conducted more than 240 A/B tests over 6 months to determine which marketing messages worked best (Rush 2012a).

An important consideration for any testing team is deciding which features of the website platform to test. The ultimate determination of which features are worth testing should depend on the potential returns the firm can gain by acting on the findings of the test. The potential returns depend both on how much better the new treatments perform (which is of course unknown before the test) and how many customers will be affected by the treatment. Consequently, most testing teams

choose to test features of their marketing that are seen by many customers and that they believe have a large potential to increase sales or other desired outcomes.

The team managing the donation platform for the Obama campaign tested several areas of the website including imagery, copy, and the donation process. Figure 2 shows an example of an image test that was used on the splash page, where potential donors arrived after clicking on a link describing a special campaign where donors could win a “Dinner with Barack” (adapted from Rush 2012b). The objective of the test was to learn whether the focused shot showing the candidate smiling (which they labeled as *control*) would perform better than the wide shot showing several attendees at a previous event chatting with the candidate and his wife (which they labeled as *variation*). Previous tests had shown that large images of the smiling candidate increased the donation rate, so the team hypothesized that the control image would perform better. The images were assigned randomly in real time to all visitors who clicked on a link to the splash page. The team used the Optimizely web-testing platform, which, like other web-testing platforms, handles the random assignment of treatments automatically and integrates with the web analytics platform to measure the response. The team assessed the performance of the two images, by comparing the percentage of people who made donations in the control group relative to the variation group. The team found that the wider shot showing previous guests at the table with the candidate resulted in a 19% increase in donations. Based on this finding, they quickly decided to change the splash page to the variation image for the remaining duration of the campaign.

Figure 3 shows another example of a test described by Rush (2012b) that involved website copy. The website had a feature that invited donors to store their payment information so that they could make donations in the future with one click. This was a very successful tool – by the end of the campaign more than 1.5 million



Fig. 2 Image test for Obama campaign (Adapted from Rush 2012b)

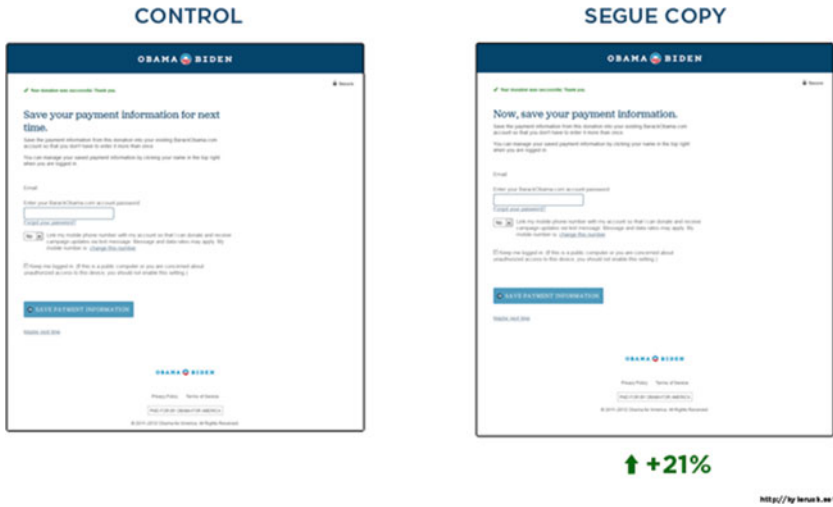


Fig. 3 Copy test for Obama campaign (Adapted from Rush 2012b)

Quick Donate users donated \$115 million – and so the team was anxious to find ways to get more donors to sign up for Quick Donate.

Figure 3 shows two versions of the page that donors saw just after making their donation. The control page asked customers: “save your payment information for next time,” while the treatment page made it seem as if saving the payment information was part of the current process by saying: “now, save your payment information.” When users were randomly assigned to the two treatments, the percentage of customers who saved their payment information was 21% greater among those who saw the *segue copy*.

This example raises a key issue that testing managers face in practice: how to measure the effect of the treatment. In this case, the team chose to compare treatments based on how many customers signed up for the Quick Donate program, and this is a logical choice as that is the immediate goal of these marketing treatments. However, Quick Donate sign-ups do not result in an immediate monetary gain for the campaign. One might also legitimately prefer to compare these two treatments based on how many actual donations are received in the subsequent month for those in each group, although this would require more time and tracking capability to measure effectively.

In field experiments in digital marketing, it is common to measure a variety of outcomes within the same experiment, both those that are directly related to the short- and long-term effects of the treatment and potential side effects such as increased costs or increased complaints. (Medical experiments face a similar challenge in defining response measures: in testing a new cancer treatment, researchers must decide whether to compare treatments based on a near-term outcome such as the recurrence of cancer in the subsequent 5 years or a longer-term outcome such as mortality in the next 20 years.)

The straightforward randomization and measurement available on the web platform allow for easy causal interpretation of the results, which in turn makes it easy for decision makers to act immediately on the findings without much risk of paralysis by analysis. As Rush describes, “In looking at the overall results I think you could say our efforts paid off. We increased donation conversions by 49%, sign up conversions by 161% and we were able to apply our findings to other areas and products.” And this sort of result is not unique: spurred on by a number of popular business books with titles like *Always Be Testing* (Eisenberg and Quarto-von Tivadar 2009), *Experiment!* (McFarland 2012), and *A/B Testing* (Siroker and Koomen 2013) where other examples of the Obama campaign’s optimization are reported, many firms are finding ways to making field experiments a regular part of how they make decisions.

Field Experiments in the Academic Literature

Field experiments are becoming popular as a tool for exploring marketing theory (Gneezy 2017), and there are many online and offline field experiments reported in the academic literature.

Offline field experiments can, for instance, be run in retail stores like Chen et al. (2012) did to test how different types of promotions can impact the volume of purchases. They tested whether the bonus pack or an equivalent price decrease of a product has an impact on the sales figures changing the promotion type on a weekly basis for 16 weeks. The employment of only one store allowed keeping all external factors constant (e.g., store layout, employees, background of customers, neighboring environment), increasing internal validity at the expense of external validity.

Furthermore, field experiments are often conducted over a long period of time in order to identify long-term effects. For example, Bawa and Schoemaker (2004) conducted two field experiments each one over a 2-year time frame aimed at estimating the long-run effect of free sampling on sales. In both cases, they recorded the sales data of the customers over 1 year (panel data). After delivering the sample at the end of the first year, the volumes were registered for another year. Of course, the longer the time frame, the higher the probability that external factors can influence the participants. In general, problematic marketing-related extraneous factors depend on the respective context and on the research topic.

Online-controlled experiments have gained popularity because of the increased digitalization of companies that are more and more engaging in a test and learn mentality. As we have discussed, A/B tests can easily be implemented to examine how users react to different webpage layouts and designs. An example is Yang and Ghose (2010), who measured the impact of different search advertising strategies on the click through rate, conversion rate, and revenues. All of these measures give an indication of how the customers use the website.

A study revealing how the use of field experiments can shed new light on existing and well-established theories is the recent paper by Anderson and Simester (2013). Standard models of competition predict that firms will sell less when competitors target their customers with advertising. This is particularly true in mature markets with many competitors that sell relatively undifferentiated products. However, the

authors present findings from a large-scale randomized field experiment that contrast sharply with this prediction. The field experiment examines the effect of competitors' advertising on sales at a private label apparel retailer. To examine this effect, the researchers sent competitive advertisement mailings to the treatment group. As customers normally have no comparison of whether other people receive the same or different mailings, they do not realize that they are part of an experiment. Results show that, surprisingly, for a substantial segment of customers, the competitors' advertisements increased sales at this retailer.

Recommended readings for those interested in online advertising are the field tests employed by Goldfarb and Tucker (2011a, c). In the same area, Blake et al. (2015) and Kalyanam et al. (2015) published large-scale field experiments aimed at studying the causal effectiveness of paid search ads. They find somewhat contradictory results: Blake et al. (2015) showed that returns from paid search ads for eBay are minimal, while Kalyanam et al. (2015) find that search ads are effective for other retailers. In a recent working paper, Simonov et al. (2015) have also confirmed that search advertising does have some benefit for less-well-established brands. They use a large-scale, fully randomized experiment on Bing data studying 2500 brands. These experiments rely on treatment and control groups made up of various geographic regions where advertising can be turned on or off; using such *geo-experiments* to measure ad effectiveness has also been suggested by researchers at Google (Vaver and Koehler 2011).

Randomized holdouts take this idea of non-exposure to customer-level experiments and are rapidly becoming popular in many industries. In a randomized holdout experiment, the marketer selects a group of customers at random to not receive planned marketing communication, such as an email, a catalog, or a promotional offer. Comparing the treated and the holdout group allows the marketer to make a causal measurement of the treatment effect, i.e., the incremental sales lift of the marketing. Hoban and Bucklin (2015) report on randomized holdout experiments in display advertising, Zantedeschi et al. (2016) report on randomized holdouts for catalog and email campaigns, and Sahni et al. (2015) report on randomized holdouts for discount offers. All of these studies find positive incremental effects of marketing. However, Lewis and Rao (2015) report similar experiments on display advertising and find effect sizes that are so small that it would be difficult to accurately measure the returns on advertising.

Lambrecht and Tucker (2013) run a field experiment with an online travel firm to examine whether *dynamic retargeting*, a new form of personalized advertising that shows consumers ads that contain images of products they have looked at before on the firm's own website, is more effective than simply showing generic brand ads. Even if this new strategy integrates the usage of both internal and external browsing data, results revealed that dynamic retargeted ads are on average less effective than traditional retargeting.

Ascarza et al. (2016) analyze retention campaigns based on pricing plan recommendations, and the results emerging from their field experiment surprisingly show that being proactive and encouraging customers to switch to cost-minimizing plans can increase rather than decrease customer churn.

As the MSI-Tier 1 priorities suggest, the customer journey is developing into a multimedia, multiscreen, and multichannel era (mobile = physical + digital worlds). Considering multichannel customer management literature, Montaguti et al. (2016) test the causal relationship between multichannel purchasing and customer profitability. Within a field experiment, they show that multichannel customers are indeed more profitable than they would be if they were single-channel customers providing insights on how multichannel shopping leads to higher profit.

Andrews et al. (2015) had the opportunity to collaborate with one of the world's largest telecom providers managing to gauge physical crowdedness in real time in terms of the number of active mobile users in subway trains. Their research examines the effects of hyper-contextual targeting with physical crowdedness on consumer responses to mobile ads, and results based on a massive field experiment counting a sample of 14,972 mobile phone users suggest that, counterintuitively, commuters in crowded subway trains are about twice as likely to respond to a mobile offer by making a purchase vis-à-vis those in non-crowded trains.

Dubé et al. (2015) implemented another massive field experiment to test an information theory of prosocial behavior. A long literature in behavioral economics has generated a collection of empirical examples where economic incentives counterintuitively reduce the supply of prosocial behavior. The data comes from two field experiments involving a consumer good bundled with a charitable donation. Considering a population of 15 million subscribers living 2 km from a theater and who purchased a ticket via phone in the previous 6 months, the sample consisted of 4200 randomly chosen individuals. Results suggest that price discounts crowd out consumer self-inference of altruism.

Nevertheless, the aforementioned papers are only some of those interesting works published involving the use of field experiments. We leave to the reader's curiosity the task to look for other field experiments!

Conclusions

As can be seen from the previous section, there are numerous examples of both companies and academics using field experiments to answer tactical questions and test marketing theory. The increasing use of field experiments in marketing is also enhancing the collaboration between firms and academia. The big challenge and opportunity here are the reconciliation of academics doing "big stats on small data" with practitioners doing "small stats on big data."

This chapter has laid out the key ideas one should think about when designing field experiments. For the reader interested in more detail, a major author of reference is John A. List, who focuses on field experiments in economics. In List (2004) the author presents a series of field experiments he conducted about theories of discrimination, and in a slightly more recent paper (2006), he reviews a broad set of field experiments to explore the implications of behavioral and neoclassical theories as well as of topics ranging from the economics of charity to the measurement of preferences. Furthermore, in 2011 he proposed 14 tips to follow for

improving academic's chances of executing successful field experiments in collaboration with companies. We suggest practitioners to refer to this checklist, before implementing their experiment ideas.

Of course, it is unavoidable to meet some challenges in the implementation and use of field experiments. First of all, as pointed out by Levitt and List (2009), field experiments do not provide the same extent of control as laboratory experiments. Therefore, internal validity is often lower, and, because of a lower level of control, potential confounding variables should be identified before starting and recorded during the experiment in order to control for them using statistical methods (Homburg 2015; Gneezy 2017). Pre-testing and continuous monitoring during the experiment are helpful to identify excluded effects and record general trends like a change of the general market conditions which can impact the sales volume independently of the experiment (Gerber and Green 2012; Gneezy 2017). This issue further reveals that researchers should put much effort and time into the planning stage and in the experimental design. On top of that, a relatively high level of knowledge of the whole experimental design and of the underlying constructs is required upfront (Levitt and List 2009). Other challenges concern privacy and security regulations that unavoidably tend to limit collection/retention of data (Goldfarb and Tucker 2011b). Future researchers should focus on the development of analytics that can overcome such limitations and on the proactive development of methods for protection of customer privacy.

In summary, this chapter outlines and argues that field experiments are, next to big data analytics, one of the major advances of the digital age which allow firms to reveal the causality between two processes, actions or observations. Managers and researchers have now to accept the challenge by ensuring that the causal inferences of their field experiments are both correct and useful in terms of advancing management and marketing practice. We hope this chapter encourages and helps managers in considering field experiments as a state-of-the-art market research approach for collection, analysis, and interpretation of market-related information.

Cross-References

- ▶ [Analysis of Variance](#)
- ▶ [Experiments in Market Research](#)

References

- Aaker, D. A., Kumar, V., Day, G. S., & Leone, R. P. (2011). *Marketing research*. Hoboken: Wiley.
- Anderson, E. T., & Simester, D. (2013). Advertising in a competitive market: The role of product standards, customer learning, and switching costs. *Journal of Marketing Research*, 50(4), 489–504.
- Andrews, M., Luo, X., Fang, Z., & Ghose, A. (2015). Mobile Ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science*, 35(2), 1–17.

- Ascarza, E., Iyengar, R., & Schleicher, M. (2016). The perils of proactive churn prevention using plan recommendations: Evidence from a field experiment. *Journal of Marketing Research*, 53(1), 46–60.
- Bawa, K., & Shoemaker, R. (2004). The effects of free sample promotions on incremental brand sales. *Marketing Science*, 23(3), 345–363.
- Blake, T., Nesko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica*, 83(1), 155–174.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Chen, H., Marmorstein, H., Tsiros, M., & Rao, A. R. (2012). When more is less: The impact of base value neglect on consumer preferences for bonus packs over price discounts. *Journal of Marketing*, 76(4), 64–77.
- Crook, T., Brian, F., Ron, K., & Roger, L. (2009). *Seven pitfalls to avoid when running controlled experiments on the web*. In Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining.
- Dubé, J.-P., Luo, X., & Fang, Z. (2015). *Self-signaling and pro-social behavior: A cause marketing experiment*. Fox school of business research paper no. 15-079. Available at SSRN: <http://ssrn.com/abstract=2635808> or <https://doi.org/10.2139/ssrn.2635808>
- Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge: Cambridge University Press.
- Eisenberg, B., & Quarto-von Tivadar, J. (2009). *Always be testing: The complete guide to Google website optimizer*. New York: Wiley.
- Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.
- Gerber, A. S., & Green, D. P. (2008). Field experiments and natural experiments. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology, Oxford handbooks online* (pp. 357–381). Oxford: Oxford University Press.
- Gerber, A. S., & Green, D. P. (2012). *Field experiments. Design, analysis, and interpretation*. New York: Norton.
- Gneezy, A. (2017). Field experimentation in marketing research. *Journal of Marketing Research*, 46, 140–143.
- Goos, P., & Jones, B. (2011). *Optimal design of experiments: A case study approach*. New York: Wiley.
- Goldfarb, A., & Tucker, C. E. (2011a). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3), 389–404.
- Goldfarb, A., & Tucker, C. E. (2011b). Privacy regulation and online advertising. *Management Science*, 57(1), 57–71.
- Goldfarb, A., & Tucker, C. E. (2011c). Advertising bans and the substitutability of online and offline advertising. *Journal of Marketing Research*, 48(2), 207–227.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Hoban, P. R., & Bucklin, R. E. (2015). Effects of internet display advertising in the purchase funnel: Model-based insights from a randomized field experiment. *Journal of Marketing Research*, 52(3), 375–393.
- Homburg, C. (2015). *Marketingmanagement. Strategie – Instrumente – Umsetzung – Unternehmensführung. Lehrbuch*. Wiesbaden: Springer Gabler.
- Homburg, C., Kuester, S., & Krohmer, H. (2013). *Marketing management. A contemporary perspective*. London: McGraw-Hill Higher Education.
- Iacobucci, D., & Churchill, G. A. (2010). *Marketing research. Methodological foundations*. Mason: South-Western Cengage Learning.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social and biomedical sciences: An introduction*. New York: Cambridge University Press.
- Kalyanam, K., McAteer, J., Marek, J., Hodges, J., & Lin, L. (2015). *Cross channel effects of search engine advertising on brick and mortar retail sales: Meta analysis of large scale field experiments on Google.com*. Working paper.

- Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1), 140–181.
- Koschate-Fischer, N., & Schandelmeier, S. (2014). A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. *Journal of Business Economics*, 84, 793–826.
- Landsberger, H. A. (1958). *Hawthorne revisited*. Ithaca: Cornell University.
- Lambrecht, A., & Tucker, C. E. (2013). When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, 50(5), 561–576.
- Ledolter, J., & Swersey, A. J. (2007). *Testing 1-2-3. Experimental design with applications in marketing and service operations*. Stanford: Stanford University Press.
- Levitt, S. D., & List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1), 1–18.
- Lewis, R. A., & Rao, J. M. (2015). The unfavourable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 1941–1973.
- List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics*, 119(1), 48–89.
- List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25(3), 3–16.
- McFarland, C. (2012). *Experiment! Website conversion rate optimization with A/B and multivariate*. Berkeley: New Riders.
- Montaguti, E., Neslin, S. A., & Valentini, S. (2016). Can marketing campaigns induce multichannel buying and more profitable customers? A field experiment. *Marketing Science*, 35(2), 201–217.
- Neyman, Jerzy. (1923[1990]). On the application of probability theory to agricultural experiments: Essay on principles. Section 9. *Statistical Science*, 5 (4), 465–472. Translated by Dabrowska, Dorota M. and Terence P. Speed.
- Rush, K. (2012a). *Meet the Obama campaign's \$250 million fundraising platform*. Blog post 27 Nov 2012.
- Rush, K. (2012b). *Optimization at the Obama campaign: a/b testing*. Blog post 12 Dec 2012.
- Sahni, N., Dan, Z., & Pradeep, C. (2015). *Do targeted discount offers serve as advertising? Evidence from 70 field experiments*. Stanford University Graduate School of Business research paper no. 15-4. Available at SSRN: <http://ssrn.com/abstract=2530290> or <https://doi.org/10.2139/ssrn.2530290>
- Sándor, Z., & Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38(4), 430–444.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont: Wadsworth Cengage Learning.
- Simonov, A., Nosko, C., & Rao J. M. (2015). *Competition and crowd-out for brand keywords in sponsored search*. Available at SSRN: <http://ssrn.com/abstract=2668265> or <https://doi.org/10.2139/ssrn.2668265>
- Siroker, D., & Pete K. (2013). *A/B testing*. Wiley.
- Teele, D. L. (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. New Haven & London: Yale University Press.
- Vaver, J., & Koehler, J. (2011). *Measuring ad effectiveness using geo experiments*. Google Research working paper.
- Yang, S., & Ghose, A. (2010). Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science*, 29(4), 602–623.
- Zantedeschi, D., McDonnell Feit, E., & Bradlow, E. T. (2016). Modeling multi-channel advertising response with consumer-level data. *Management Science*, Articles in Advance. <http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2451>



Crafting Survey Research: A Systematic Process for Conducting Survey Research

Arnd Vomberg and Martin Klarmann

Contents

Introduction: Relevance of Survey Research	68
Understanding Survey Bias	70
Fundamentals of Survey Research	70
Psychology of Survey Response	71
Measurement Theory	72
Sources of Systematic Errors in Survey Research: Measurement Errors	74
Sources of Systematic Errors in Survey Research: Representation Errors	85
Survey Research Process	89
Selection of Research Variables	89
Selection of Survey Method	91
Questionnaire Design	93
Data Collection	104
Measurement Evaluation	105
Data Analysis	106
Endogeneity in Survey Research	106
Conclusion	110
Cross-References	110
References	111

Abstract

Surveys represent flexible and powerful ways for practitioners to gain insights into customers and markets and for researchers to develop, test, and generalize theories. However, conducting effective survey research is challenging. Survey researchers must induce participation by “over-surveyed” respondents, choose

A. Vomberg (✉)

Marketing Department, University of Groningen, Groningen, The Netherlands

e-mail: A.E.Vomberg@rug.nl

M. Klarmann

Department of Marketing & Sales Research Group, Karlsruhe Institute of Technology (KIT),
Karlsruhe, Germany

e-mail: martin.klarmann@kit.edu

appropriately from among numerous design alternatives, and need to account for the respondents' complex psychological processes when answering the survey. The aim of this chapter is to guide investigators in effective design of their surveys. We discuss state-of-the-art research findings on measurement biases (i.e., common method bias, key informant bias, social desirability bias, and response patterns) and representation biases (i.e., non-sampling bias and non-response bias) and outline when those biases are likely to occur and how investigators can best avoid them. In addition, we offer a systematic approach for crafting surveys. We discuss key steps and decisions in the survey design process, with a particular focus on standardized questionnaires, and we emphasize how those choices can help alleviate potential biases. Finally, we discuss how investigators can address potential endogeneity concerns in surveys.

Keywords

Survey research · Biases · Survey design · Survey research process · Measurement theory · Common method bias · Key informant bias · Social desirability · Response styles · Non-sampling bias · Non-response bias · Item reversal · Order bias

Introduction: Relevance of Survey Research

Surveys are ubiquitous, used to inform decision makers in every walk of life. Surveys provide practitioners with deeper insights into the attitudes of their customers (e.g., Hohenberg and Taylor (chapter ► [“Measuring Customer Satisfaction and Customer Loyalty”](#)) in this handbook) and employees (e.g., employee satisfaction surveys). Surveys are also helpful in exploring theoretical mechanisms for theory testing and development, as survey research can contribute to generalizing experimental findings to different persons and settings (Krosnick 1999; MacKenzie and Podsakoff 2012). Many relevant and important research questions would be difficult to study without relying on survey data (Hulland et al. 2018). Often, adequate secondary data are not available and experimental manipulations are not feasible. Thus, unsurprisingly, marketing research has a “rich tradition...in survey research” (Rindfleisch and Heide 1997, p. 30).

Surveys represent a versatile and powerful research instrument that is applicable in various contexts. For instance, investigators rely on surveys to study:

- Customer attitudes (e.g., customer satisfaction, customer loyalty, voice of the customer surveys)
- Employee attitudes (e.g., employee satisfaction, employee commitment)
- Service quality (e.g., surveys about hotel service)
- Product quality (e.g., surveys with package inserts)
- Performance evaluations (e.g., training evaluation surveys)
- Product feedback (e.g., new product/concept testing surveys)

Recent findings demonstrate the superiority of survey research over other methods. For instance, a recent meta-analysis reveals that direct survey-based techniques more validly indicate consumers’ willingness-to-pay than indirect methods (Schmidt and Bijmolt 2019). Similarly, survey research might deliver the most valid results in studies of sensitive topics (John et al. 2018).

However, despite its important benefits, survey research is in decline (Hulland et al. 2018). Possibly, awareness of potential biases that can occur in survey research may have nurtured skepticism toward surveys, rendering findings less trustworthy or credible. Thus, a critical challenge for survey research lies in separating noise and bias from a survey. As an understanding of how biases emerge will help investigators enhance the validity of their surveys, we discuss the most commonly identified biases in survey research.

Researchers need to make various decisions when developing their surveys. We introduce a systematic process to survey research design that will help investigators organize and structure survey development by answering guiding questions for each stage of the survey research process. In addition, we outline how those decisions can help to alleviate potential biases – an important consideration, as biases from survey research can to a large extent be attributed to “haphazard decisions” (Schwarz 2003, p. 588), investigators make when constructing surveys. While we focus primarily on procedural remedies to avoiding biases (ex ante bias prevention), we also briefly address statistical techniques (ex post bias corrections) and direct readers to further literature. Such statistical techniques represent important supplements to effective survey design.

After reading this chapter, researchers will have an in-depth understanding of the various biases that may affect the results of survey research. In addition, researchers will comprehend the general survey process and know which decisions in survey development will help to reduce potential biases. Figure 1 shows how we have

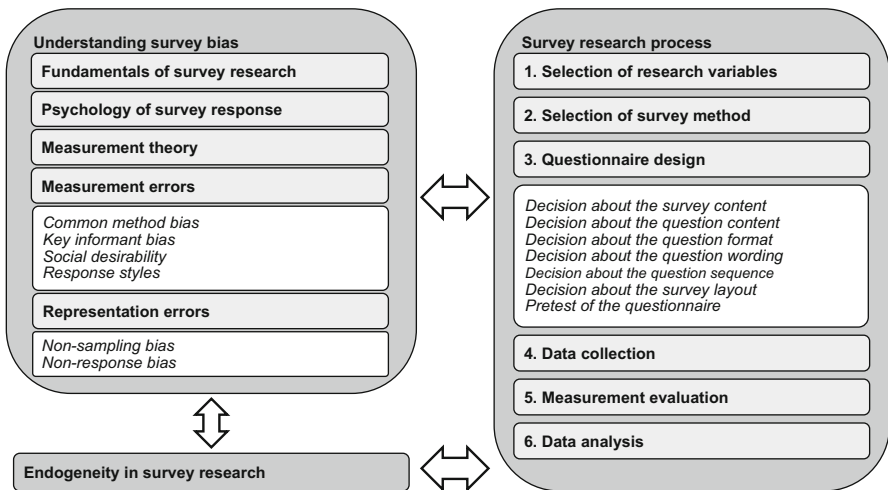


Fig. 1 Overview of chapter structure

structured this chapter. We first foster an understanding of survey bias (section “[Understanding Survey Bias](#)”) by discussing the psychology of survey response (section “[Psychology of Survey Response](#)”) and measurement theory (section “[Measurement Theory](#)”). We then discuss in detail important sources of systematic errors in survey research, which we classify into measurement errors (section “[Sources of Systematic Errors in Survey Research: Measurement Errors](#)”) and representation errors (section “[Sources of Systematic Errors in Survey Research: Representation Errors](#)”). We subsequently outline the survey research process (section “[Survey Research Process](#)”), with a particular focus on how to design the questionnaire (section “[Questionnaire Design](#)”). We briefly address the issue of endogeneity (section “[Endogeneity in Survey Research](#)”) and end by summarizing and aligning the sections (section “[Conclusion](#)”).

Understanding Survey Bias

Fundamentals of Survey Research

A survey comprises a “cross-sectional design in relation to which data are collected predominantly by questionnaire or by structured interview on more than one case (usually quite a lot more than one) and at a single point in time” (Bryman and Bell 2015, p. 63). Surveys can be categorized by several aspects, such as

1. The method in which they are administered to the participant: written, online, telephone, or personal surveys
2. Time horizon: cross-sectional versus longitudinal surveys
3. The role of the respondent: self-reports versus key informants

Survey administration can be classified into personal, telephone, written, and online surveys. We discuss these different forms when outlining the selection of the survey method (section “[Selection of Survey Method](#)”), focusing primarily on written and online surveys because these are the dominant forms of survey research (Hulland et al. 2018).

The time horizon can be purely cross-sectional or longitudinal. While cross-sectional surveys are administered at a single point in time, longitudinal surveys comprise repeated observations for different time periods (e.g., Heide et al. 2007; Jansen et al. 2006; Wathne et al. 2018). Cross-sectional surveys are the dominant form of applied research, encompassing 92.1% (Hulland et al. 2018, p. 94) and 94% (Rindfleisch et al. 2008, p. 262). Examples of longitudinal surveys include

- American Customer Satisfaction Index (theacsi.org; Fornell et al. 1996), which tracks the evolution of customers’ satisfaction with several companies over time
- Harris Poll EquiTrend study (e.g., Vomberg et al. 2015) or Young & Rubicam Brand Asset Valuator (e.g., Mizik and Jacobson 2008), which gauge consumers’ brand perceptions

- The CMO survey, which regularly surveys the opinions of chief marketing officers (e.g., cmosurvey.org)

Longitudinal surveys can be designed in various ways. The American Customer Satisfaction Index and the Harris Poll EquiTrend Study are repeated cross-sectional surveys in which different respondents are sampled each time. Alternatively, researchers can conduct a panel survey in which mostly the same respondents are surveyed each time (chapter ► [“Panel Data Analysis: A Non-technical Introduction for Marketing Researchers”](#) by Vomberg and Wies).

Regarding the role of the respondent, survey participants can either provide self-reports or act as key informants. In self-reports, participants assess questions for themselves. For instance, they may indicate their level of satisfaction or their attitude toward a focal brand. In contrast, key informants provide answers for a higher-order social entity. For instance, employees may indicate the strategic orientation of their company. Key informants are commonly relied on in organizational contexts. We elaborate later on a potential bias stemming from the use of key informants (section [“Key Informant Bias”](#)).

Psychology of Survey Response

Survey researchers need to be aware of the psychological processes that typically occur when participants answer a questionnaire. In business research, participants usually cannot offer predefined responses but form their evaluations when answering the questionnaire. For instance, when customers are asked about their satisfaction with a company, they are not likely to retrieve such an evaluation directly from memory, but instead tend to reflect on their answer when completing the questionnaire. Thus, survey questions trigger a cognitive process of response generation.

It is beyond the scope of this chapter to outline the variety of models that have been proposed to capture these cognitive processes. Therefore, we only briefly summarize the model of Tourangeau et al. (2000), which investigators frequently refer to when studying respondent behaviors (e.g., MacKenzie and Podsakoff 2012; Podsakoff et al. 2003; Weijters et al. 2009).

Tourangeau et al. (2000) argue that respondents pass through five stages when replying to survey questions: (1) comprehension, (2) retrieval, (3) judgment, (4) response selection, and (5) response reporting. In the comprehension stage, participants attend to the survey question and deduce its intent. Respondents then generate a retrieval strategy and search their memories for relevant information. Retrieval thus entails the process of bringing information held in long-term memory to an active state, in which it enters the short-term memory to be used (belief-sampling model). Respondents integrate this information into a judgment (e.g., their satisfaction with a certain product). Finally, when selecting their response, participants map the judgment onto the offered response categories and report their answer (e.g., Krosnick 1999; Tourangeau et al. 2000; Weijters and Baumgartner 2012).

Each of these stages is quite complex and involves a significant amount of cognitive effort by the participant. Thus, during this process, participants may not

be motivated to process survey items in sufficient detail to provide a valid statement. However, even motivated respondents may retrieve biased information. The accessibility-diagnostics theory argues that respondents retrieve information that is accessible to them and has a high diagnosticity (Feldman and Lynch 1988). For instance, very salient but exceptional events (e.g., a specific negative incident with a company) are likely to be more accessible than regular events and thus lead respondents to provide a distorted picture of their true attitudes and opinions. In addition, information provided in earlier questions may represent a source of information that respondents use to form their answers. For instance, the sequence of questions may influence what information respondents retrieve when answering subsequent questions (section “[Decisions About the Question Sequence](#)”). Even if consumers retrieve accurate information, they must make substantial efforts to condense this complex information into rather simple answer categories, such as scales from 1 to 7 (Homburg et al. 2012c).

These examples highlight the complexity of survey response. In the following, we outline the consequences of these psychological processes for the interpretation of survey data.

Measurement Theory

Reliability and Validity: Fundamentals

Measurement theory claims that any observed value x for a question (e.g., an observed value for the liking of a brand) is the sum of a true value t (also referred to as trait) and measurement error, which can have a random e and a systematic component s . Hence, any observed value can be understood in the following way (Eq. 1):

$$x = t + s + e \quad (1)$$

Importantly, the random error component poses threats to the reliability of a survey question, and the systematic error component can affect the question’s validity. A survey question can be considered reliable when it produces the same results under the same measurement conditions, whereas the question has validity when it actually measures what it purports to measure. An intuitive example to understand the concept of random and systematic error is the following. Imagine that 100 researchers measure the time it takes for a participant to run a certain distance. Usually, when all researchers compare their results, their observed measurements will differ slightly. Thus, individual measurements likely suffer from random measurement error. However, since this error is assumed to be randomly distributed among participants, its expected value is zero: $E(e) = 0$. Hence, when taking the average value, researchers likely obtain an unbiased measure (Iacobucci 2013).

With respect to survey research, a characteristic such as imprecise wording can raise fundamental threats to reliability: participants could interpret words such as “usually” or “almost” differently, adding noise to the data

(section “[Decisions About the Question Wording](#)”). However, since the expected value of the random error is zero, survey researchers can ask multiple questions when measuring abstract concepts such as commitment or satisfaction. Just as in the stop watch example, averaging multiple measurements may help to safeguard against reliability concerns (section “[Decisions About the Question Content](#)”).

In contrast, systematic errors affect the validity of survey questions. Validity refers to the degree to which a measure really measures what it is supposed to measure. Since the expected value of the systematic error is not zero ($E(s) \neq 0$), repeated measurements cannot alleviate potential validity concerns. Intuitively, this can be explained by continuing with the stop watch example. If all researchers had received stop watches that systematically add 10 s, even the average of the individual measurements will be biased. We discuss sources of systematic errors in survey research in sections “[Sources of Systematic Errors in Survey Research: Measurement Errors](#)” and “[Sources of Systematic Errors in Survey Research: Representation Errors](#).”

Reliability and Validity: Implications for Survey Research

A natural follow-up question is the extent to which random and systematic errors influence the results of survey research. Many times survey researchers are interested in establishing relationships between variables. In the simplest case, investigators can focus on bivariate correlation coefficients, and in the following we discuss the bivariate correlation coefficient between a variable x (e.g., customer satisfaction (CS)) and a variable y (e.g., word-of-mouth behavior (WOM)). In line with Eq. 1, we assume that both variables are measured with error. We apply two common assumptions in deriving the correlation coefficient: we assume (1) uncorrelated random measurement errors (i.e., $\text{Cov}(e,t) = 0$; $\text{Cov}(e,s) = 0$; $\text{Cov}(e_x, e_y) = 0$) and (2) no correlation between the true value and the systematic measurement error ($\text{Cov}(t,s) = 0$). These assumptions lead to the following correlation coefficient (Eq. 2) (e.g., Baumgartner and Steenkamp 2001; Homburg and Klarmann 2009):

$$\begin{aligned}
 r(x,y) &= \frac{\text{Cov}(CS; WOM)}{\sqrt{\text{Var}(CS) \cdot \text{Var}(WOM)}} = \frac{\text{Cov}(x;y)}{\sqrt{\text{Var}(x) \cdot \text{Var}(y)}} = \\
 &= \frac{\text{Cov}(t_x + e_x + s_x; t_y + e_y + s_y)}{\sqrt{\text{Var}(t_x + e_x + s_x) \cdot \text{Var}(t_y + e_y + s_y)}} = \quad (2) \\
 &= \frac{\text{Cov}(t_x, t_y) + \text{Cov}(s_x, s_y)}{\sqrt{[\text{Var}(t_x) + \text{Var}(e_x) + \text{Var}(s_x)] \cdot [\text{Var}(t_y) + \text{Var}(e_y) + \text{Var}(s_y)]}}.
 \end{aligned}$$

Equation 2 offers three important insights. First, a common concern regarding survey research is that participants may provide inflated answers that bias the results (e.g., De Jong et al. 2015). For instance, managers may be tempted to exaggerate their performance. However, while these over- or understatements affect mean

values, mean values do not directly affect Eq. 2. Thus, (systematic) over- or understatements do not bias the relationships between variables in survey research.

Second, in many research applications, investigators are interested in the direction of a relationship rather than in the size of the coefficient. Since the denominator in Eq. 2 contains only variances that cannot become negative, only the numerator is responsible for the direction of the correlation coefficient. Increased variances in the denominator can only reduce the size of the correlation coefficient. Consequently, Eq. 2 reveals that random measurement errors cannot change the sign of the correlation coefficient since the random error is only part of the denominator. From the perspective of a survey researcher, Eq. 2 implies that random measurement errors can lead to only conservative results by decreasing statistical power. Thus, random measurement errors may obscure an effect that is present, leading to Type II errors. However, random measurement errors cannot artificially create relationships.

Third, the impact of systematic measurement errors (also referred to as method error) can be twofold since the systematic error is part of both the numerator and the denominator in Eq. 2. If sources of systematic errors affect both variables independently (i.e., $\text{cov}(s_x, s_y) = 0$), then systematic errors have the same impact as random errors: they can lower statistical power but cannot artificially create effects. However, more likely sources of systematic errors (e.g., key informant bias) affect both variables simultaneously (i.e., $\text{cov}(s_x, s_y) \neq 0$). Thus, systematic errors can affect not only the strength but also the direction of an effect. Systematic errors could be responsible for either detecting artificial relationships in cases in which there is no true relationship (i.e., Type I error) or masking existing relationships (i.e., Type II error) (Baumgartner and Steenkamp 2001; Homburg et al. 2012c). Huang and Sudhir (2021) provide compelling empirical evidence for the latter. They show that systematic biases in survey research can lead to conservative estimates, that is, an underestimation of true effect sizes.

Sources of Systematic Errors in Survey Research: Measurement Errors

In this section, we review key sources of systematic biases that may affect the validity of survey research. A bias is “any force, tendency, or procedural error in the collection, analysis, or interpretation of data which provides distortion” (Crisp 1957, cited in Tortolani 1965, p. 51). Knowledge regarding these threats is important because, to a certain extent, researchers can safeguard against them when designing the survey instrument.

In the following, we describe the most commonly discussed biases in survey research. Following prior investigators, we categorize those biases into measurement errors and representation errors (e.g., Baumgartner and Steenkamp 2006). Measurement errors reflect tendencies of respondents to answer to survey questions on some grounds other than the item content. Specifically, we discuss common method bias (section “[Common Method Bias](#)”), key informant bias (section “[Key Informant Bias](#)”), social desirability bias (section “[Social Desirability](#)”), and response patterns

(section “[Response Styles](#)”). Representation errors reflect biases due to the selection of the sample of respondents. These biases could follow from unrepresentative sampling frames – non-sampling bias (section “[Non-sampling Bias](#)”) – or participants’ unwillingness or inability to respond – unit non-response bias (section “[Non-response Bias](#)”).

Common Method Bias

Conceptualization. Common method bias (CMB) – one of the most frequently discussed threats to survey research – can largely undermine the validity of a study’s findings (Hulland 2019; Palmatier 2016). However, no agreement on a definition of CMB presently exists. CMB definitions differ widely in scope. Podsakoff et al. (2003, p. 880) favor a broad definition and refer to CMB as “variance that is attributable to the measurement rather than the construct of interest.” Under this definition, CMB can be regarded as an umbrella term for various biases that can be classified into four categories (Podsakoff et al. 2003, p. 882):

- *Common source rater:* for example, consistency motif, social desirability (Section “[Social Desirability](#)”), response patterns (section “[Response Styles](#)”), or a tendency toward satisficing
- *Item characteristics:* for example, item ambiguity (section “[Decisions About the Question Wording](#)”), common scale formats or anchors (section “[Decision About the Question Format](#)”)
- *Effects due to item context:* for example, item embeddedness (section “[Decisions About the Question Sequence](#)”)
- *Measurement context:* for example, independent and dependent variables measured at the same point in time

A narrow definition of CMB focuses on the last category and attributes the distortion of the sample’s covariance structure to the use of the same data source to measure both independent and dependent variables (e.g., Klarmann 2008; Rindfleisch et al. 2008). We adopt this definition and discuss other biases that Podsakoff et al. (2003) mention separately.

CMB can emerge when a single informant provides information on the independent and dependent variables. As we elaborate below, the severity of CMB will then depend on various factors, such as the type of collected measure (lower CMB threat for objectively verifiable measures), the complexity of the tested relationships (lower CMB threat for quadratic and interactive relationships), whether a time lag occurs in data collection (lower CMB threat when independent and dependent variables are collected at different points in time), or the response format (lower CMB threat if the independent and dependent variables are measured in different scale formats).

CMB can undermine the validity of survey research in two ways. First, it can lead to biases in estimates of construct reliability and validity, as it can inflate reliability and validity measures by 18 to 32% (Podsakoff et al. 2012, p. 543). Thus, CMB can lead researchers to mistakenly believe that they have validly measured the constructs of interest when in reality they have captured method artifacts.

Second, CMB can bias the parameter estimates between two constructs. Notably, however, investigators document vastly different results when quantifying the impact of CMB: some evidence indicates that, on average, CMB inflates correlation coefficients between constructs by approximately 45%, with a range of 27% to 304% (MacKenzie and Podsakoff 2012, p. 543; Podsakoff et al. 2012, p. 545). However, a meta-analysis of 42,934 correlation coefficients showed correlations based on single-source self-reports to be on average only 0.02 higher than correlation coefficients from different sources (Crampton and Wagner 1994).

Reasons for occurrence. To effectively address CMB, understanding the reasons for its occurrence is important. Given the pivotal role CMB assumes in marketing research projects, the literature not surprisingly offers plenty of explanations for why CMB may occur (Podsakoff et al. 2003 provide a detailed discussion). In the following, we highlight selected explanations.

First, when reading the questions, respondents may start developing implicit theories of the relationships between the constructs. In effect, their answers then represent reflections of these theories. For instance, managers might be asked to evaluate the innovativeness of their business unit and their competitive position. If the participating managers develop the implicit theory that innovativeness represents a focal driver of performance, then their answers referring to the business unit's innovativeness might be influenced by prior answers on the unit's competitive position (e.g., Podsakoff et al. 2003).

Second, respondents strive for consistency in their cognitions. Therefore, they may try to provide answers that they think are consistent. They may search for similarities in the questions and adapt their answers accordingly. As a consequence, stated answers may not truly reflect their behaviors or opinions.

While these two explanations in general outline why CMB can emerge, an important question is, in which situation is CMB more likely to occur? Empirical evidence suggests that two factors drive potential CMB concerns: (1) the nature of the constructs under investigation and (2) the complexity of the investigated relationships.

The literature rather consistently states that the likelihood of CMB to occur depends on the nature of the questions. For example, the percentage of common method variance is lower in marketing (16%) than in psychology or sociology (35%) (Cote and Buckley 1987). Constructs investigated in psychology or sociology are likely to be abstract and complex and thus be harder to answer and may trigger cognitive processes that increase covariation between systematic error components (MacKenzie and Podsakoff 2012; Podsakoff et al. 2012; section "Psychology of Survey Response"). Empirical evidence confirms these expectations. In general, concrete, externally oriented, and verifiable constructs are less prone to CMB than are abstract, internally oriented, and non-verifiable constructs (Chang et al. 2010; Rindfleisch et al. 2008).

Second, with increasing complexity of the investigated relationships, CMB is less likely to occur. Analytical evidence demonstrates that CMB can deflate quadratic and moderating relationships but cannot create them (Siemsen et al. 2010). For such complex relationships, participants are unlikely to develop corresponding implicit

theories that affect their responses. Therefore, if the main interest of research lies in identifying quadratic relationships or interaction effects, CMB is not likely to undermine researchers' findings (Vomberg et al. 2020).

Procedural remedies. A general recommendation is that investigators should favor procedural remedies over statistical remedies when addressing potential biases in survey research. While statistical controls are rather symptom-based – that is, they target the consequences of CMB only in the analysis stages – procedural remedies try to eliminate sources of CMB in the moment of collecting the data. Naturally, in many situations researchers will not be able to completely alleviate CMB concerns with procedural remedies, and in these cases, statistical remedies can help to elevate the credibility of the findings. In the following, we discuss four procedural remedies: (1) use of different data sources, (2) temporal separation, (3) proximal separation, and (4) psychological separation of scale of scale formats.

(1) *Different data sources.* An investigator nullifies the risk of CMB when relying on different data sources for the independent and dependent variables (Rindfleisch et al. 2008, p. 274; Ostroff et al. 2002). This approach makes it impossible for participants to develop implicit theories between independent and dependent variables.

First, researchers can survey different respondents to evaluate the independent and dependent variables (Gruner et al. 2019) – that is, use dyadic data: one portion of the sample is used to estimate the independent variables and the remaining portion evaluates the dependent variables. Empirical evidence supports the effectiveness of dyadic data to attenuate CMB. Correlations of independent and dependent variables when rated by one respondent ($r = 0.359$) dropped by 49% when different respondents evaluated them ($r = 0.184$) (Podsakoff et al. 2012).

Although this procedure appears promising, it requires large sample sizes, and therefore is not appropriate for all kinds of surveys. Particularly in organizational research, researchers observe generally declining response rates, making it especially challenging to recruit additional respondents (section “Non-response Bias”). It might also be problematic in small companies where the owner is in charge of most of the decisions (Rindfleisch et al. 2008).

Second, researchers can rely on a combination of secondary data and survey data. Most commonly, researchers collect independent variables via a survey study and evaluate performance outcomes from profit and loss statements (e.g., Vomberg et al. 2020). Research suggests that this approach is also effective. Obtaining independent and dependent variables from different data sources can reduce their correlations by 49% (Podsakoff et al. 2012, p. 548).

However, relying on different sources might not be feasible in several settings. For instance, use of different data sources is not viable when the two variables cannot be validly inferred from different sources (e.g., self-referential attitudes and perception constructs; Johnson et al. 2011), when archival data cannot adequately represent the construct, or when such data are available only for prohibitively high costs in terms of money or time.

(2) *Temporal separation of measurement.* An alternative might be to separate the collection of independent and dependent variables temporally by including a

time lag (e.g., Homburg et al. 2020; Jansen et al. 2005; Vomberg et al. 2020). Empirical research indicates that temporal separation of measurement can be effective for reducing potential CMB. However, effectiveness depends on the length of the time lag, with longer time lags triggering the possibility that intervening events might introduce new sources of biases (Podsakoff and Organ 1986; Chang et al. 2010). While some research indicates that temporal separation is effective for same point in time versus two weeks later (Johnson et al. 2011) and same point in time versus one month later (Ostroff et al. 2002), other work finds no significant improvement with relatively long time lags (30 vs. 36 months) (Rindfleisch et al. 2008).

(3) *Proximal separation*. Researchers can increase the proximal distance between independent and dependent variables in the questionnaire. As we elaborate in section “[Decisions About the Question Sequence](#),” separating measures significantly reduces shared variance (Weijters et al. 2009). For instance, concern regarding CMB can be alleviated by first measuring the dependent and then the independent variables.

(4) *Psychological separation of scale formats*. Researchers frequently rely on common scale formats for a variety of questions within a questionnaire. While common scale formats make answering the survey questions easier, they enhance the probability that cognitions formed when answering one question may be retrieved when answering another question. However, relying on different scale formats when assessing the independent and dependent variables reduces potential CMB (e.g., Vomberg et al. 2020), as different formats disrupt potential consistency bias (Feldman and Lynch 1988; Podsakoff et al. 2012; Rindfleisch et al. 2008).

Initial empirical evidence supports the effectiveness of using different scale formats. For instance, labeling anchor points differently for independent and dependent variables shrinks CMB by 18% (Johnson et al. 2011). Even larger decreases of 38% or 60% occur when the scale format is changed completely (Arora 1982; Kothandapani 1971).

Statistical remedies. The literature contains discussions of several statistical remedies for reducing CMB. Since these remedies are beyond the scope of the article, we briefly mention only a few. Podsakoff, MacKenzie, and Podsakoff (2012) provide a detailed overview of several statistical techniques, and Hulland et al. (2018) outline how to test for CMB. Hulland et al. (2018) discuss the correction-based marker variable technique, which is one of the most frequently applied approaches to address CMB in marketing research. This approach suggests that a marker variable that is theoretically unrelated to the constructs under investigation resembles the amount of CMB (Lindell and Whitney 2001). Grayson (2007, Appendix B) intuitively summarizes this approach.

Other approaches include (1) the measured latent factor technique, or measured response style technique, in which the researcher directly measure sources of CMB (e.g., Wathne et al. 2018), (2) the unmeasured latent factor technique, in which researchers model a method factor in structural equation models (e.g., Homburg et al. 2011), and (3) endogeneity-correction approaches, in which researchers employ instrument-free techniques (e.g., Vomberg et al. 2020).

Key Informant Bias

Conceptualization. Survey participants can act either as respondents providing self-reports or as key informants. While respondents describe their own beliefs, attitudes, opinions, or feelings, key informants generalize about patterns for a higher social unit (e.g., a company's culture) on the basis of their experiences (Seidler 1974). For example, in an employee satisfaction survey, employees act as respondents when they evaluate their own job satisfaction level. However, they act as key informants when they assess an organizational culture (higher social entity). Key informant bias can occur in the latter case; it comprises the distortion of the sample's covariance structure that arises because the data collection has taken place through key informants. As we elaborate below, the severity of key informant bias will depend on various factors, most pressingly on the expertise of the key informant (less bias if the key informant is knowledgeable, has a higher hierarchical position, and has longer job tenure) and the type of collected data (less bias for objectively verifiable measures).

To the best of our knowledge, only two studies systematically evaluate the occurrence of key informant bias (Homburg et al. 2012b; Kumar et al. 1993), and the two studies deliver conflicting conclusions. Kumar, Stern and Anderson (1993, p. 1646) report high levels of key informant *disagreement*, whereas Homburg et al. (2012c, p. 605) find in their meta-analytical study high levels of key informant agreement (0.872).

In addition, Homburg et al. (2012c) investigate how large the systematic error provoked by key informant bias needs to be to create artificial effects or to mask true effects. They demonstrate that the requisite size depends largely on the study's sample size and the strength of the effect, and conclude that in the studies of their meta-analysis there is a risk that small and medium-sized effects were not correctly identified.

Reasons for occurrence. Key informant bias can emerge if investigators have unrealistic expectations about the key informant's knowledge. For instance, it might not be feasible for key informants to provide complex judgment about organizational characteristics that may result in random measurement error. In addition, key informant bias may emerge owing to positional bias or knowledge deficiencies (Phillips 1981). For instance, key informants' backgrounds can substantially influence the answers they provide (Homburg et al. 2005). When marketing and sales managers were asked to evaluate the marketing function (same higher social entity), mean values differed considerably between key informants from the two departments (Fig. 2). Sales managers perceive marketing managers (on a scale from 0 to 100) to have a rather low level of customer know-how (mean value = 37), whereas marketing managers on average consider themselves to have rather high levels customer know-how (mean value = 59).

In addition, in line with empirical findings on CMB, the risk of key informant bias largely hinges on the nature of the studied constructs. Key informant reports are reliable for objective and salient issues from the present (e.g., performance outcomes). However, for more abstract measures (e.g., organizational culture) key informants tend to be less accurate. Furthermore, a higher hierarchical position and tenure also increase reliability (Homburg et al. 2012c; Kumar et al. 1993).

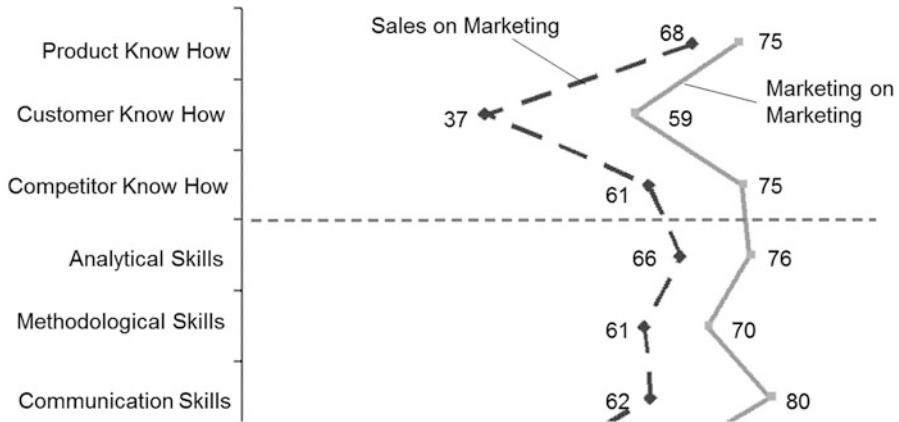


Fig. 2 Marketing and sales managers' evaluations of the marketing function – comparison of mean values (Homburg et al. 2005, p. 14, reproduced with the kind permission of the publisher)

Procedural remedies. The most important and effective remedy to alleviate concerns regarding key informant bias is the careful selection of key informants. The investigator needs to carefully align the contacted key informants with the objective of the study. Typically, key informants are not chosen randomly and are not deemed representative for the population. Instead, researchers select respondents who have special qualifications, such as their position in the company or their knowledge (Phillips 1981; Kumar et al. 1993).

The question of whom to contact as key informants depends primarily on the research context. For instance, if the research objective focuses on strategic issues, high-ranked key informants are probably appropriate. However, if the study investigates operational aspects, such as specific sales approaches used, then key informants with regular customer contact are more promising.

In addition, questions that require less demanding social judgments and instead are impersonal and focus on objective and observable phenomena seem preferable (Phillips 1981; Homburg et al. 2012c; Klarmann 2008). However, whether this condition is realistic depends largely on the context of the study. Nevertheless, even if more complex or abstract constructs are the focus of the study, the researcher might lower the risk of key informant bias with careful wording of the questionnaire items (section “[Decisions About the Question Wording](#)”).

Finally, if key informant response accuracy is expected to be low or empirical evidence for reliability is scarce in the particular research domain, the researcher may use triangulation – that is, combining methods in the study of the phenomenon. For instance, investigators can try to survey multiple key informants per company. However, since triangulation is costly, sometimes difficult to implement (Rindfleisch et al. 2008), and devoid of substantial additional value if the responses of the first respondent have been accurate (Homburg et al. 2012c), this option should be carefully chosen.

Statistical remedies. Statistical procedures address either the reliability or validity of the key informant study. They aim either to establish or enhance reliability or validity (Homburg et al. 2012c) provide a systematic overview of various approaches). For instance, authors can establish reliability by demonstrating key informants' job tenure or experience with the company (Kumar et al. 1993) or by measuring their perceived knowledge level (e.g., Ghosh and John 2005; Kumar et al. 2011). In the case of triangulated data, researchers can also employ correlational approaches, such as intraclass correlation ICC(1) (Bliese 2000; chapter ► “Multi-level Modeling” Haumann et al. in this handbook) or the absolute deviation agreement index (Burke and Dunlap 2002; LeBreton and Senter 2008). To enhance validity, researchers can integrate factors representing the data sources into a structural equation model, allowing estimation of trait relationships while controlling for systematic error (e.g., Cote and Buckley 1987).

Social Desirability

Conceptualization. As one researcher noted over 20 years ago, “One well-known phenomenon in survey research is overreporting of admirable attitudes and behaviors and underreporting of those that are not socially respected” (Krosnick 1999, p. 545). This phenomenon is called social desirability and represents the distortion of the sample's covariance structure that arises owing to the tendency of respondents to reply in a manner that others will view favorably (e.g., De Jong et al. 2010). For instance, in a study on consumer innovativeness, 41% of the respondents indicated they owned, had repurchased, or had seen products that did not actually exist in the market (Tellis and Chandrasekaran 2010).

While social desirability is often considered a source of bias in survey research (e.g., Podsakoff et al. 2003), surprisingly few investigators explicitly address it in their studies (Steenkamp et al. (2010) and Tellis and Chandrasekaran (2010) represent some notable exceptions). As a consequence, knowledge about the impact of social desirability on survey results is scarce.

Reasons for occurrence. Theory proposes two factors to explain the occurrence of socially desirable responding: the level of awareness (self-deception vs. impression management) and the domain of content (agency-based vs. communion-based contexts). Self-deception occurs when participants *unconsciously* dissemble, and impression management occurs when participants *consciously* dissemble (e.g., Krosnick 1999; Paulhus 1984).

Regarding the domain of content, some respondents are more likely to engage in socially desirable responding because they have a need to be perceived as more powerful than others (egoistic response tendencies, arising in agency-based settings). Other respondents offer socially desirable responses because they have a need to be perceived as exceptionally good members of society (moralistic response tendencies, emerging in communion-based contexts) (Paulhus and John 1998; Paulhus 2002; Steenkamp et al. 2010).

Procedural remedies. Self-deception and impression management are particularly important in controlling or reducing potential social desirability bias. Socially desirable responding owing to self-deception might be part of a respondent's

personality, and therefore should not be controlled for to prevent elimination of a central component of individual differences in personality. However, the impression management component should be controlled for because this component embodies a conscious bias (Paulhus 1984).

Several methods are available to prevent or at least reduce socially desirable responses to questionnaires. Survey researchers should assure respondents that their answers are anonymous, reinforce that the items have no right or wrong answers, emphasize that people hold different opinions about the issues addressed in the questionnaire, and encourage respondents to answer as honestly as possible. Paradoxically, socially desirable responding remains an issue under anonymous conditions when there is no one to impress (Mick 1996).

Another remedy is to allow respondents to report on the external world rather than answering questions about themselves. The solution is to ask indirect questions in the neutral third-person form that are not affected by the social desirability bias. This approach is based on the assumption that respondents project their opinion onto others and consequently give more honest answers, an assumption that Fisher (1993) demonstrates empirically. In addition, the social distance between interviewer and respondent should be reduced (Nederhof 1985), and socially sensitive questions should be eliminated or placed at the end of the questionnaire to avoid carry-over effects (Baumgartner and Steenkamp 2006).

Finally, a third way to deal with a social desirability bias is through randomized response techniques (e.g., De Jong et al. 2010, 2015; Himmelfarb and Lickteig 1982; Warner 1965). In this approach, respondents are asked a sensitive question (e.g., “Are you willing to pay higher prices for sustainable products?”) with response options of “yes” and “no.” Prior to answering the question, they flip a coin and adapt their answers following the outcome of the coin flip. If the coin flip returns “heads,” respondents should answer the question with “no” regardless of whether they truly engaged in the questioned behavior. However, if the flip returns “tails,” respondents should answer the question truthfully with “yes” or “no.” Since investigators cannot see the outcome of the coin flip, they cannot tell whether a particular “no” response denotes a negative answer that reveals the respondent’s true attitude or a coin flip that has come up heads (or both). Note that in this simplistic illustration, only the socially less desirable answer (i.e., “no”) is concealed (De Jong et al. (2010) describe more advanced designs). In theory, this technique should counteract a social desirability bias. However, initial empirical evidence indicates that randomized response techniques can also deliver worse results than directly asking participants (Holbrook and Krosnick 2010; John et al. 2018).

Statistical remedies. The literature offers two main approaches to check or control for social desirability: (1) including a measured latent factor capturing social desirability in the empirical model (e.g., Podsakoff et al. 2003) or (2) correlating separate survey measures with a measured social desirability score (Steenkamp et al. (2010) develop a systematic process to check for social desirability concerns). The latter approach demonstrates whether a social desirability bias likely affects the constructs in the study and the first approach is intended to control for social desirability in empirical models. However, both approaches require the investigator to measure the

social desirability construct. To do so, literature suggests relying on Paulhus's (2002) *Balanced Inventory of Desirable Responding* (e.g., Steenkamp et al. 2010).

Response Styles

Response styles lead to a distortion of the sample's covariance structure that arises because, regardless of the question content, respondents favor certain response categories (e.g., Van Rosmalen et al. 2010). The most frequently discussed response styles are respondents' tendency to select specific subsets of response options such as disproportionately favoring the positive side of a response scale (acquiescence response style, or yay-saying).

Response styles can compromise the comparability of the data: the same responses can have different meanings for different respondents. Participants may display different response styles between countries (e.g., Tellis and Chandrasekaran 2010) or between different modes of data collection. Response style biases are higher in telephone interviews than in written and online surveys (Weijters et al. 2008). Response styles thus undermine the ability to validly compare mean values. Furthermore, response styles can also influence construct variances and correlations between constructs. For instance, researchers found that response styles account for 8% of construct variance in their sample and for 27% of the variance in the observed correlations (Baumgartner and Steenkamp 2001).

However, despite their potentially biasing effects, response styles have not received much attention in the marketing literature. In the following, we discuss two common groupings of response styles: (1) acquiescence, disacquiescence, and net acquiescence and (2) extreme responding, midpoint responding, and response range.

Acquiescence, disacquiescence, and net acquiescence. Acquiescence, also called yay-saying, is the respondent's tendency to agree with items regardless of their content, whereas disacquiescence, or nay-saying, is the respondent's tendency to disagree with items regardless of their question content (Baumgartner and Steenkamp 2001; Tellis and Chandrasekaran 2010). Net acquiescence is acquiescence minus disacquiescence and reflects the tendency to show greater yay- than nay-saying (Baumgartner and Steenkamp 2001; Tellis and Chandrasekaran 2010). However, many researchers do not distinguish between acquiescence and net acquiescence (e.g., Greenleaf (1992) uses the label acquiescence but actually measures net acquiescence). We focus on acquiescence and disacquiescence.

Acquiescence poses problems for segmentation research because it can lead to the emergence of clusters that reflect response styles rather than attitudinal information (Greenleaf 1992). In addition, it may falsely heighten correlations among items that are worded in the same direction (Winkler et al. 1982). Remarkably, although acquiescence and disacquiescence appear to be opposites, empirical evidence demonstrates only small- to medium-sized negative correlations between them (Baumgartner and Steenkamp 2001: $r = -0.16$; Tellis and Chandrasekaran 2010: $r = -0.31$).

Dispositional (e.g., personality traits) and situational factors (e.g., item ambiguity) may explain why the three response styles occur (Knowles and Condon 1999).

Regarding dispositional factors, research has demonstrated inconsistent results for demographic variables (particularly gender). For instance, research shows that respondents with an acquiescence response style tend to be extroverted, impulsive, emotional, and undercontrolled. Respondents prone to a disacquiescence response style are likely to be introverted, cautious, rational, and overcontrolled. Acquiescence may be evoked by a desire to please, to be agreeable in social situations, or to display deference to the researcher (e.g., Krosnick 1999).

However, findings on the participants' cultural backgrounds are more consistent. Respondents from countries scoring high on collectivism (uncertainty avoidance) tend to display more (less) acquiescence in their response style (Tellis and Chandrasekaran 2010).

Regarding situational factors, participants are more likely to display an acquiescence response style if response categories are fully labeled (Weijters et al. 2010a), if response categories contain a neutral point (Weijters et al. 2010a), or if items are ambiguous (Podsakoff et al. 2003). Similarly, the personal situation of the participant can provoke an acquiescence response style. An acquiescence response style is likely to occur if participants read items uncritically (Messick 2012), if they experience time pressure (Baumgartner and Steenkamp 2006), or if their cognitive capabilities are exceeded (e.g., items at the end of a long questionnaire) (MacKenzie and Podsakoff 2012).

Extreme responding, midpoint responding, and response range. A respondent with an extreme response style tends to favor the most extreme response categories regardless of the item content. Midpoint responding refers to the tendency to use the middle-scale category regardless of the question content. Response range refers to the tendency to use a wide or narrow range of response categories around the mean response (Baumgartner and Steenkamp 2001).

Thus far, few conceptual and empirical investigations have focused on midpoint responding and response range. Current knowledge suggests that response ranges (and also an extreme response style) may relate to the characteristics of the respondent; they concern rigidity, intolerance of ambiguity, and dogmatism and are associated with higher levels of anxiety and possibly deviant behavior (Hamilton 1968).

In the last few years, investigators have become increasingly interested in extreme responding, with research linking it to personality (Cabooter et al. 2012; Naemi et al. 2009), scale format (Weijters et al. 2020), language (De Langhe et al. 2011; Weijters et al. 2013), and culture (De Jong et al. 2008) – although these papers do not necessarily use the term extreme response style (e.g., De Langhe et al. (2011) talk about “anchor contraction”). In addition, a recent literature stream on item response trees (IRTtree) models investigated extreme (and midpoint) responding (e.g., Böckenholt 2012, 2017; Zettler et al. 2015).

Procedural remedies. Thus far, few procedural remedies have been identified for addressing biases from response styles. To the best of our knowledge, the literature provides only initial remedies for midpoint responding and for an acquiescence and a disacquiescence response style.

First, to address midpoint responding, some authors suggest eliminating the middle response category or including a “don't know” category (Baumgartner and

Steenkamp 2006; Schuman and Presser 1996). Second, since acquiescence is more likely for positively worded items and disacquiescence for negatively worded items (Baumgartner and Steenkamp 2001), investigators suggest introducing doubly balanced scales (Tellis and Chandrasekaran 2010). However, researchers need to carefully evaluate this option since negatively worded items may lead to misresponse (Weijters et al. 2010b; section “[Decisions About the Question Wording](#)”).

Finally, response patterns are often problematic in international market research. Our recommendation is that researchers should make sure that all respondents receive a similar response format with equally familiar labels, preferably in their native language.

Statistical remedies. To investigate whether response styles likely bias the survey results, investigators should correlate the focal constructs of interest with measures of the different response styles – that is, use a measured response style technique. If this analysis does not demonstrate potential biases, then the investigators should proceed as planned. However, if potential threats to validity emerge, the authors should include measures of different response styles in their models (e.g., Weijters et al. 2008).

The literature proposes various ways in which response styles can be measured. For example, response styles can be calculated on the basis of items that are included in the survey – it is not necessary to add items to specifically measure response styles (Weijters et al. 2008, 2010b; Tellis and Chandrasekaran 2010). Alternatively, a latent-class model can identify different response styles (Rosmalen et al. 2010).

Sources of Systematic Errors in Survey Research: Representation Errors

Non-sampling Bias

Conceptualization. Investigators in many research projects are interested in generalizing findings to an overall population. Therefore, representativeness of the sample is important. That is, the sample needs to display approximately the same characteristics as the overarching population. However, as Fig. 3 illustrates, two biases may threaten representativeness: non-sampling bias and nonresponse bias. Non-sampling bias refers to the distortion of the sample’s covariance structure that arises because the population is not adequately represented in the original sample (i.e., the sample that received the questionnaire does not resemble the overall population). Non-response bias is a distortion of the sample’s covariance structure that arises because the structure of the final sample does not correspond with the structure of the original sample. We discuss non-sampling bias and nonresponse bias in section “[Non-response Bias](#).”

Whether non-sampling bias represents an important threat to a study depends on the research objective. If the researcher is interested in making generalizations to a particular target group, then non-sampling bias represents a severe threat. Thus, in the language of experimental research, if the research objective focuses on external validity, representative heterogeneous samples are required. However, if the

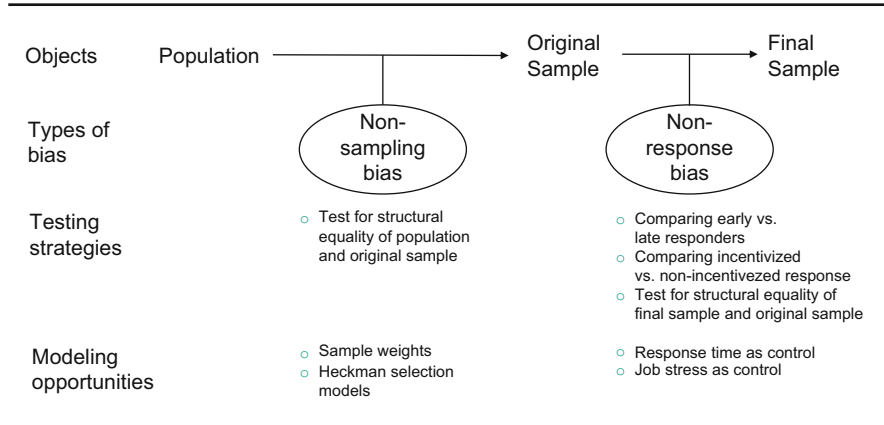


Fig. 3 Illustration of non-sampling and nonresponse bias. (Adapted from Homburg and Krohmer 2008)

researcher is interested in testing the veracity of proposed theoretical effects, non-sampling bias is a lesser issue. Stated differently, when the emphasis lies on internal validity, homogeneous samples can be adequate.

Short, Ketchen, and Palmer (2002) study the effects of non-sampling bias systematically. They link CEO duality, in which a firm's CEO concurrently chairs the board of directors, to a company's return on equity. The authors draw on different sampling frames, for which they observe substantially different effects: no effects in two cases, a positive effect in one case, and a negative effect in one case.

While sampling biases can have important consequences for research findings, the issue of sampling is often not systematically addressed. A review of studies published in leading management journals revealed that less than 40% of the studies included a discussion of representativeness (Short et al. 2002). Similarly, a review of marketing studies found that only 55% used an explicit sampling frame (Hulland et al. 2018). Finally, an important finding is that in general, samples chosen in social science are biased toward Western, educated, industrialized, rich, and democratized countries (Henrich et al. 2010).

Reasons for occurrence. Like resource constraints, the lack of suitable sampling frames can be a principal reason for non-sampling bias. For instance, while past consumer research could rely on telephone books for sampling frames, telephone books no longer validly resemble the population of interest (Iacobucci and Churchill 2010).

Similarly, for organizational studies, identification of adequate sampling frames can be challenging. For example, companies can be classified along numerous criteria (McKelvey 1975). In addition, the question arises regarding the organizational level at which the study should be conducted. A focus on strategic business units or product groups may present additional challenges for identifying adequate sampling frames (Sudman and Blair 1999).

Procedural remedies. The only procedural remedy is to base the sample on a sampling frame that is representative of the overall population. However, as mentioned, various reasons may obstruct knowledge of the overall population. Investigators may rely on sampling frames provided by data bases (e.g., COMPUSTAT), commercial mailing lists (e.g., Heide et al. 2014), or releases from Federal Statistical Offices regarding industry compositions (e.g., Vomberg et al. 2020). Nexis Uni (previously LexisNexis) may also serve to define the population for a study (Homburg et al. 2020). Nexis Uni has the advantage of including private companies, which are common for the business-to-business sector.

Statistical remedies. Statistical remedies can comprise methods to establish and to enhance representativeness. First, researchers can establish representativeness of their sample by comparing the structural equivalence of the original sample and the population (e.g., Vomberg et al. 2020). For instance, a χ^2 test may demonstrate the representativeness of the sample (Homburg et al. 2020).

Even if the distribution in the population is not known, researchers may compare the composition of two samples. For instance, Homburg et al. (2015b) compared the structural equivalence of samples obtained in two waves of data collection (years 1996 and 2013) in terms of industry sectors, sales volume, and number of employees.

Second, researchers may rely on Heckman selection models to control for non-sampling bias (Vomberg et al. 2020). Certo et al. (2016) provide a detailed discussion of Heckman selection models in the context of strategic management and also provide Stata code. Third, researchers may rely on sampling weights to enable the generalizability of their findings (Andreß et al. 2013; McElheran 2015; Raval 2020).

Non-response Bias

Conceptualization. In contrast to non-sampling bias, which refers to structural differences between the population and the original sample, non-response bias refers to structural differences between the original and the final sample (Fig. 3). Non-response bias may result from participants' failure to answer the survey (i.e., unit nonresponse) or from return of incomplete questionnaires (i.e., item nonresponse) (Klarmann 2008).

Analysis of reported response rates from leading management and organizational journals for the years 1975, 1985, and 1995 showed that the average response rate reported in these journals was 55.6% (with a large standard deviation of 19.7%) (Baruch 1999). However, responses have declined substantially over time – a trend that has been observed in various settings such as household studies (De Heer and De Leeuw 2002) or online surveys (Cook et al. 2000). Since the average US consumer receives more than 550 unsolicited surveys per year (compared with 50 to 100 for Germany, the UK, or France) (Iacobucci and Churchill 2010, p. 192), “oversampling” likely (partly) explains the dramatic reduction in response rates since Baruch's study. Today, response rates over 20% in organizational studies constitute the exception (Heide et al. 2007, 2014) and response rates around 10% represent the rule (Klarmann 2008).

In addition, response rates differ substantially between target groups. Top management studies in particular typically yield low response rates (Anseel et al. 2010).

However, participant drop-out does not necessarily threaten the results of a study. Non-response bias is a threat only when the reason for drop-out is related to the survey content (e.g., customers take part in a customer satisfaction survey depending on their satisfaction levels, or people with less discretionary time are less satisfied and less likely to respond to a satisfaction survey) and when sources that positively and negatively relate to survey content are not balanced (e.g., only unsatisfied customers do not reply in a customer satisfaction survey) (Thompson and Surface 2007).

Reasons for occurrence. Unit non-response is typically the result of respondents' refusal to participate in a survey. A review of factors that may drive unit non-response revealed personal factors, organizational factors, and survey-related factors (Klarmann 2008):

- *Personal factors:* Personal attitudes toward the survey (e.g., Helgeson et al. 2002; Rogelberg et al. 2001), involvement with the research topic (e.g., Groves et al. 2004), authorization (e.g., Tomaskovic-Devey et al. 1994; Gupta et al. 2000), and demographic criteria (e.g., Gannon et al. 1971; Gupta et al. 2000) have a strong impact on response rates.
- *Organizational factors:* Organizational factors such as industry profitability, dependence (i.e., subsidiary), and company size also influence the response rate (Tomaskovic-Devey et al. 1994).
- *Survey-related factors:* Number of contacts with participants (i.e., pre-notifications and reminder) (e.g., Yu and Cooper 1983), personalization of the survey (Yu and Cooper 1983), incentives (Church 1993; Yu and Cooper 1983; Yammarino et al. 1991), and length of the questionnaire (Yammarino et al. 1991) affect respondents' decisions to participate.

Procedural remedies. In the literature, four measures are discussed to diminish unit non-response bias (e.g., Klarmann 2008; Rogelberg and Stanton 2007; Anseel et al. 2010 for an in-depth review and evaluation of which techniques are effective for a particular target group):

- *Activities that increase the opportunity to participate in the survey:* These measures include (1) a deadline that can be determined by the participant (e.g., have the survey run for sufficient time so that vacation time does not impede response), (2) reminder notes, and (3) different modes of participation (e.g., written, online, per telephone)
- *Activities that emphasize the importance of the survey:* Emphasis can be achieved by (1) pre-notifying participants personally that they will receive a survey in the future, (2) reflecting participants' interests (e.g., an engaging title), (3) informing the participants about the survey goals.
- *Activities that decrease the perceived costs of the participation:* Researchers should (1) manage survey length and (2) carefully consider the survey design

- *Activities that raise the perceived utility of participation:* Researchers may provide incentives to participants. Common incentives in business-to-business contexts include social incentives (e.g., researchers donate of 10–15€ per participant for a social cause; Vomberg et al. 2020) and reports which outline the study results or benchmark a firm to sample averages; the latter two are more suitable incentives if participants are interested in the survey content. In business-to-consumer contexts payments (e.g., 5€ per participant; Homburg et al. 2019b), advance incentives (e.g., small gift included with the survey such as a pen), or a raffle (e.g., ten randomly chosen participants win a gift coupon from Amazon.com) represent commonly selected incentives. Thereby, the former two are likely effective even if participants are not involved with the survey content.

Statistical remedies. Researchers can use several different approaches to detect and to address a potential nonresponse bias (Rogelberg and Stanton 2007; Klarmann 2008). To establish that the final sample resembles the original sample, investigators can conduct a χ^2 goodness-of-fit test (Homburg et al. 2012b). If incentives are used for only a subset of respondents, then investigators could compare the answers of incentivized to non-incentivized participants. Researchers can also conduct follow-up interviews with non-respondents using a short version of the questionnaire (i.e., focal constructs) to determine the reasons for non-response (e.g., Groves 2006; Homburg et al. 2007). Finally, if researchers suspect that factors like job stress influence a participant's likelihood to respond, then they should also measure such variables and add them as control variables in their statistical models (Homburg et al. 2010).

Survey Research Process

In this section, we outline the general phases of survey research. Figure 4 shows that the typical survey research process starts with selection of research variables and ends with data analysis. Since we focus on the design of survey research in this section, we discuss questionnaire design in greater detail and comment only briefly on issues such as data analysis.

Selection of Research Variables

Which Relationships Are of Particular Interest?

In the first stage, the most important question is which relationships should be investigated. The answer is obviously determined by the research question the investigator wants to address. For instance, researchers might be interested in the role of formal and informal organizational elements for reacquiring customers (Vomberg et al. 2020). In this case, their conceptual model could contain formal reacquisition policies and an informal failure-tolerant organizational culture (Fig. 5).

<i>Phase</i>	<i>Illustrative questions</i>
1. Selection of research variables	Which relationships are of particular interest?
2. Selection of survey method	Should data be collected personally, by telephone, in written form or as an online survey? How can data be collected in an online context?
3. Questionnaire design	What questions should be asked? How should the questions be worded? In what order should the questions be asked?
4. Data collection	What is the ideal structure of the sample? How large should the sample be? How can we achieve a high response rate?
5. Measurement evaluation	To what extent do the survey questions measure what they are supposed to measure?
6. Data analysis	How are the examined phenomena related? How can the results be illustrated?

Fig. 4 Phases of the typical survey process

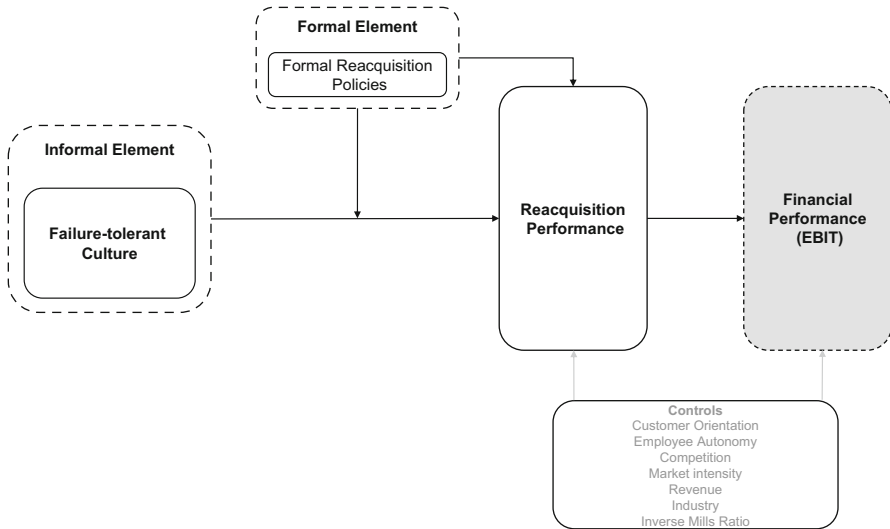


Fig. 5 Illustrative conceptual model (Vomberg et al. 2020, p. 121, reproduced with the kind permission of the publisher)

Importantly, the identification of variables requires familiarity with the research topic and is informed by prior research, conceptual considerations, or, for instance, in-depth interviews with practitioners. The conceptual model should

be as coherent and complete as possible, encompassing potential mediating and/or moderating effects if they are of interest for the subsequent analysis (e.g., Koschate-Fischer and Schulle (chapter ► [“Mediation Analysis in Experimental Research”](#)) in this handbook).

Selection of Survey Method

Should Data be Collected Personally, by Telephone, in Written Form, or as an Online Survey?

We discuss data collection modes only briefly because our focus is on written and online surveys. The personal interview takes place face-to-face between the interviewer and the interviewee. The interviewer reads out the items on the questionnaire to the interviewees and collects their answers. The telephone interview can be a short, cold-call interview or a planned interview. The written survey usually involves sending questionnaires to the respondent by mail, with the respondent then returning the completed questionnaire. The online survey can be conducted in various ways – as a questionnaire attached to an e-mail or as a link provided on a website, in an e-mail, or in a QR-code.

Table 1 which is based on Homburg (2020) systematically compares the appropriateness of the four survey methods with regard to the (1) suitability for the object of study, (2) extent of data, (3) quality of data, and (4) length and costs of the project. These general evaluations may differ according to the survey context. In addition, the four forms of survey method likely differ in the sources of biases that may relate with them. Usually, personal interviews and telephone surveys are more prone to social desirability bias or acquiescence compared to self-administered questionnaires (i.e., written and online surveys) (Krosnick 1999; MacKenzie and Podsakoff 2012). In telephone interviews, midpoint response styles are less common than in mail or online surveys. In addition, telephone surveys tend toward acquiescence. Compared to mail surveys, online surveys display fewer disacquiescence and extreme response styles (Weijters et al. 2008; section [“Response Styles”](#)).

How Can Data Be Collected in an Online Context?

Current trends in consumer research suggest that researchers increasingly rely on crowdsourcing platforms to obtain convenient and inexpensive samples. Mechanical Turk (MTurk), which was launched by Amazon in 2005, currently is the prevalent option. On MTurk, researchers act as employers and hire and compensate workers for participation in surveys, giving MTurk several likely advantages. Since incentives are comparably low on MTurk, researchers can conduct studies at lower costs without necessarily compromising data quality. In addition, participants likely have strong incentives to fill in the survey with due diligence, since they receive compensation only after passing screening and attention questions. Finally, MTurk allows researchers to implement their studies quickly (e.g., Chandler and Paolacci 2017; Levay et al. 2016).

However, skepticism regarding the use of MTurk is increasing. First, while MTurk samples are likely more heterogeneous than student samples, they are not

Table 1 Comparison of different survey methods

Criteria	Standardized verbal survey	Standardized telephone survey	Standardized written survey	Standardized online survey
Suitability for the object of study				
Possibility to explain a complex issue	Very good	Good	Rather low	Rather low
Possibility to use complex scales and filter questions in the questionnaire	Only when computer-aided	Only when computer-aided	Low	Very good
Possibility to demonstrate trial product samples	Very good	Rather low	Rather low	Rather low
Possibility to include multimedia (e.g., video and sound)	Rather good	Medium	Low	Very good
Extent of data				
Subjective length of the survey evaluated by the interviewee	Short	Medium	Long	Very long
Possibility to question a large sample	Low	Medium	Very high	High
Response rate	High	Medium	Low	Low
Risk of aborting the interview	Low	Medium	High	Very high
Quality of data				
Possibility of inquiries when comprehension problems occur	Very good	Good	Very low	Very low
Possibility of distortion due to the social interaction with the interviewer	Very high	High	None	None
Possibility to modify the research instruments during the field phase	Very good	Good	Low	Good
Length and costs of the project				
Length of the field phase	Medium	Short	Long	Long
Costs	Very high	High	Low	Low

necessarily representative of the overall consumer population. Second, participants on MTurk can be experienced survey takers who answer differently than less experienced survey respondents. Third, participants on MTurk can misrepresent themselves to participate in attractive studies with high payouts. However, misrepresentation is a general threat to online research (Goodman et al. 2013; Goodman and Paolacci 2017; Hulland and Miller 2018; Wessling et al. 2017).

TurkPrime and Prolific Academic (ProA) have recently emerged as alternatives to MTurk. TurkPrime is a website that utilizes MTurk but claims to overcome potential disadvantages of MTurk (e.g., participant misrepresentation) (Litman et al. 2017).

Graduate students from Oxford and Sheffield Universities launched ProA (<http://www.prolific.co>) in 2014. ProA provides several demographic details about its participant pool. Peer et al. (2017) who compare ProA with MTurk note that ProA participants are less experienced with research paradigms and provide more honest answers than MTurk participants. Marketing researchers increasingly rely on ProA for their studies (e.g., Castelo et al. 2019; Hagen 2020).

Qualtrics, Survey Sampling, Inc. (SSI), or Critical Mix, whose samples are likely more representative than MTurk samples, are other potential ways to collect data. However, these platforms are also more expensive. Finally, for short surveys (up to 10 questions) Google Surveys may offer a valid option. On Google Surveys, respondents take part because they want to read an article, lowering the likelihood of misrepresentation threats (Wessling et al. 2017; Hulland and Miller 2018).

Questionnaire Design

The design of the questionnaire usually represents the crucial stage in the collection of survey data. Mistakes that occur in this stage are normally hard to revise in later stages. Figure 6 outlines the phases of questionnaire design and presents examples of questions during the different phases.

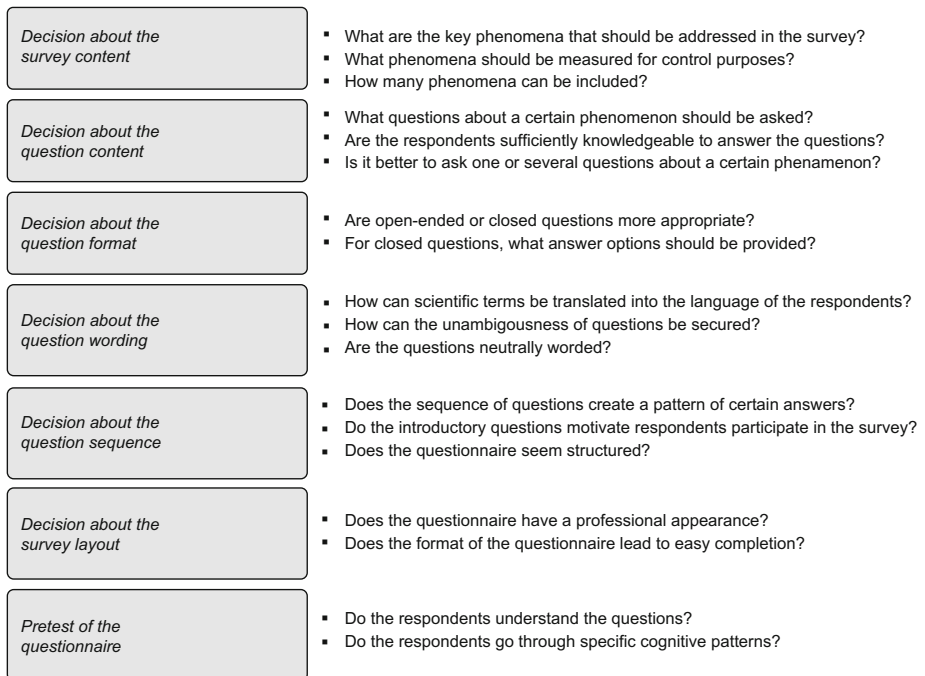


Fig. 6 Phases of the typical questionnaire design process

Decision About the Survey Content

Three fundamental questions determine survey content: What key phenomena should the survey address? What phenomena should be measured for control purposes? How many phenomena can be included in the survey?

What Are the Key Phenomena that Should Be Addressed in the Survey?

The key phenomena are provided by the research questions and the formulated hypotheses. Drawing on the conceptual framework (section “[Selection of Research Variables](#)”), investigators decide on the phenomena/constructs that should be included in the survey. In our example (Fig. 5), investigators would include variables that capture formal reacquisition policies, failure tolerance, and reacquisition performance. If financial performance cannot be obtained from archival data bases, then measures of financial performance should also be collected.

What Phenomena Should Be Measured for Control Purposes?

Besides including variables for substantive reasons, researchers should include theoretically motivated control variables to reduce “noise” in the independent and dependent variables. More importantly, however, survey researchers rely on control variables to account for non-randomness in their data. In contrast to experimental data, survey data are a form of observational data, and thus effects observed from survey data might be correlational and not causal in nature. To increase survey researchers’ ability to infer causality, control variables serve to rule out rival explanations (section “[Endogeneity in Survey Research](#)”). For instance, in our example (Fig. 5), customer orientation may represent an important control variable. Theory suggests that customer-oriented companies might be more successful in winning customers back (i.e., customer orientation may affect the dependent variable) and also are more likely to have formal reacquisition policies in place (i.e., customer orientation may influence the independent variable). However, additional variables have been included to account for further potential differences.

To systematically collect control variables, researchers should review prior literature, follow theoretical considerations, and rely on their own plausibility considerations. They should then store the identified variables in a long list, such as an Excel sheet containing potentially relevant variables. However, in many cases, researchers end up with a list that captures too many variables to be assessed in one questionnaire and have to make a trade-off between questionnaire length and the completeness of the variables.

How Many Phenomena Can Be Included?

Consideration of questionnaire length is essential to reducing potential biases. Excessive length of the questionnaire can lead to respondent fatigue, resulting in careless and biased responses (MacKenzie and Podsakoff 2012) and contributing to an acquiescence response style (section “[Response Styles](#)”). In addition, long questionnaires can lead to survey non-response (section “[Non-response Bias](#)”), which may threaten the representativeness of the survey or lead to sample sizes that are too small for statistical analysis. As rules of thumb, we recommended the following:

- Telephone interview: 20–30 mins (planned interview; less for a cold call)
- Pop-up online survey: 25 questions
- Written survey: 100 questions

Researchers should also carefully evaluate options discussed in section “[Sources of Systematic Errors in Survey Research: Measurement Errors](#)” on how to control for biasing effects, such as using dedicated variables to account for potential CMB through the measured latent factor approach. However, this approach requires that the questionnaire becomes longer and that other substantially important variables be dropped from the questionnaire. Therefore, when designing their surveys, researchers need to critically reflect which biases are likely to present threats to the validity of their findings.

However, we point out that these rules of thumb are general, as situational factors also influence the length of questionnaires. For instance, questionnaires for high-level corporate decision makers need to be shorter, whereas surveys of lower-level employees might be longer. In addition, whether participants are willing to answer longer questionnaires depends heavily on their involvement with the research topic. If participants consider the survey interesting, they may be willing to answer longer questionnaires (section “[Nonresponse Bias](#)” discusses measures to stimulate involvement).

Decisions About the Question Content

A next critical step is to decide on the question content. Typical questions in this step are: What questions about a certain phenomenon should be asked? Are the respondents sufficiently knowledgeable to answer the questions? Is it better to ask one question or more about a certain phenomenon? Which questions should be selected?

What Questions about a Certain Phenomenon Should Be Asked?

To determine which questions are appropriate for certain phenomena, researchers first need to conceptually define their constructs of interest. These definitions logically precede the operationalization of the construct – that is, the process of coming up with items for their measurement.

The next important decision is whether to rely on new or established scales to measure the constructs. For many constructs, various established scales already exist in the literature. Researchers can find scales in the following ways:

- Manuals such as the *Handbook of Marketing Scales* (Bearden and Netemeyer 1999) provide an overview of established scales.
- The website inn.theoriezeit.org allows online search for established measurement scales
- Articles published in leading journals typically provide the items for the scales they used (e.g., Homburg et al. 2011).

Researchers can systematically collect these established questions with appropriate references in an Excel sheet.

An important advantage of relying on existing scales is that they enjoy wide acceptance in the academic community because prior research has rigorously

demonstrated their psychometric properties (e.g., Homburg et al. 2015a). In addition, development of original questions can sometimes lead to problems in evaluating their reliability and validity in later stages (section “[Measurement Evaluation](#)”), as even after rigorous pretesting, newly developed scales may not display required psychometric properties in this handbook).

However, a middle way is possible: researchers can combine newly developed items with established items to better fit their research context. In addition, researchers can adapt established items to the context of their study. For instance, Kumar et al. (1995) relied on established items from Price and Mueller (1986) but adapted them to context of their study (e.g., established item: “To what extent are you fairly rewarded for the amount of effort you put forth” vs. adapted item: “How fair are your firm’s outcomes and earnings compared to the contributions we make to this supplier’s marketing effort”).

Are the Respondents Sufficiently Knowledgeable to Answer the Questions?

When selecting questions for the constructs of interest, researchers need to have their target respondent characteristics in mind (e.g., educational background). Researchers should especially evaluate whether respondents are sufficiently knowledgeable to respond to the questions – a pivotal concern in key informant studies (section “[Key Informant Bias](#)”). If respondents lack the ability to answer the particular questions, they may take on an acquiescence response style (section “[Response Styles](#)”).

Is it Better to Ask One Question or Several about a Certain Phenomenon?

When deciding about the questions’ content, researchers need to determine the number of questions they intend to ask per variable because adding redundant or unnecessary questions increases the questionnaire’s length, perhaps squeezing out questions to collect information on other constructs or evoking the respondent’s refusal to participate in the survey (Bergkvist and Rossiter 2007). However, relying on too few questions may result in problems regarding the reliability or validity of the measurement (section “[Reliability and Validity: Fundamentals](#)”).

In the marketing literature, debate is ongoing as to whether and when single items can be used (Bergkvist and Rossiter 2007; Diamantopoulos et al. 2012). Usually, single items are sufficient when a construct can be precisely captured (e.g., demographic information, revenues). However, in the case of less concrete constructs (e.g., customer satisfaction, company culture, and leadership styles) multiple-item scales offer several benefits. First, multiple-item scales help to control random measurement error. Addition of further items to the measurement of one construct typically increases the overall reliability of the scale (section “[Measurement Theory](#)”). Second, multiple-item measures are more likely to capture all facets of the construct of interest (Baumgartner and Homburg 1996). For instance, the construct transformational leadership, which captures leader behaviors that make followers more aware of the importance of their job and that inspire employees to strive for the benefit of the company, is composed of facets such as “identifying and articulating a vision,” “providing an appropriate model,” or intellectually stimulating employees, thus requiring multiple items for its measurement (Podsakoff et al. 1990). Third,

multiple-item measures enable evaluation of the psychometrics properties for the respective scales (section “[Measurement Evaluation](#),” in this handbook).

In the following, we provide analytical evidence that increasing the number of items can enhance the overall reliability of a construct (Moosbrugger 2008). Equation 3 formally defines the reliability of a construct (x) measured with a single item (e.g., the researcher measures customer satisfaction with a single question).

$$\text{Rel}(x) = \frac{\text{Var}(t)}{\text{Var}(x)} = \frac{\text{Var}(t)}{\text{Var}(t + e)}. \tag{3}$$

The observed variance of the construct x (customer satisfaction) is split into a true value t and a random error component e . Recall that reliability refers to random but not to systematic errors (section “[Reliability and Validity: Fundamentals](#)”). Therefore, we do not include a systematic error in Eq. 3. The resulting reliability measure from Eq. 3 ranges from 0 to 1, with larger values indicating higher levels of reliability.

When measuring a construct with multiple items (x_1, x_2, \dots, x_k), the overall reliability of the scale (e.g., reliability of the overall customer satisfaction measure) can be calculated according to Eq. 4. For simplicity, Eq. 4 assumes that each individual item has the same individual reliability $\text{Rel}(x)$.

$$\text{Rel}_{k \times x} = \frac{k \times \text{Rel}(x)}{1 + (k - 1) \times \text{Rel}(x)}. \tag{4}$$

Table 2 is based on Eq. 4 and demonstrates that the overall reliability of a construct increases with the number of items. To be more precise, Table 2 demonstrates the reliability of a scale depending on the number of items and individual reliability of an item (i.e., $\text{Rel}(x) = 0.40$ and $\text{Rel}(x) = 0.60$). For instance, if an individual item has a reliability of $\text{Rel} = 0.40$ and the researcher uses this single item

Table 2 Illustration of how number of items increases construct reliability

# Items	Scale reliability	
	Individual item reliability	
	0.40	0.60
1	0.40	0.60
2	0.57	0.75
3	0.67	0.82
4	0.73	0.86
5	0.77	0.88
6	0.80	0.90
7	0.82	0.91
8	0.84	0.92
9	0.86	0.93
10	0.87	0.94

to measure a construct, then the overall reliability of the scale is also $Rel_{1 \times x} = 0.40$. However, if the investigator asks five items to measure one construct (i.e., the questionnaire contains five different questions that all measure customer satisfaction) and each item has an individual reliability of $Rel(x) = 0.40$, the overall reliability of the scale increases to $Rel_{5 \times x} = 0.77$ – a result that clearly demonstrates the advantage of using multiple items. In addition, Table 2 demonstrates that scale reliability does not increase linearly with the number of items (i.e., moving from one to two items increases scale reliability more than moving from nine to ten items).

Diamantopoulos et al. (2012) replicate the analytical evidence in a simulation study and demonstrate that in most instances multiple-item scales outperform single-item scales. Situations in which single items outperform multiple items are not realistic in applied research, as researchers would not have a priori knowledge about which item is most adequate in the specific context.

Against these observations, we recommend relying on single items for concrete constructs (e.g., Carson and Ghosh 2019) but using multiple items for complex constructs (Diamantopoulos et al. 2012; Hulland et al. 2018). In the case of multiple items, as a general rule, researchers should select between four to six items (Table 2). In addition, to avoid reactance from respondents when evaluating similar questions, in the beginning of the survey researchers should point out the necessity of repeating similar questions.

Which Questions Should Be Selected?

A final and important decision is which questions should be selected. The domain-sampling model addresses this question theoretically. The domain-sampling model postulates that a given construct has a broad universe of possible items, behaviors, and responses that can serve as its observable markers or indicators (Nunnally 1967). Relying on concepts from sample selection theory (e.g., the concept of representativeness), the domain sampling model proposes that a valid representation of a construct is achieved when representative items are selected. Representative items can be identified through random sampling – that is, researchers randomly pick items from their list of items that could be used to measure the construct. However, when a given domain is adequately specified (e.g., through prior research or conceptual considerations), the researcher can deliberately select (rather than sample) a set of items (Little et al. 1999).

Decision About the Question Format

The important issue in this step is whether open-ended or closed questions are more appropriate, and for closed questions, what answer options should be provided?

Are Open-Ended or Closed Questions More Appropriate?

Open-ended questions allow participants to freely articulate their thoughts and opinions, while closed questions require participants to choose answers from a given list. Closed questions can have important advantages particularly for written and online surveys. First, the meaning of provided answers to closed questions might be clearer to the researcher than answers to open-ended questions, which likely require subjective interpretation. Relatedly, many statistical techniques

(e.g., regression analysis) require quantitative and structured data (however, the automated analysis of qualitative data is evolving; e.g., Humphreys (chapter ► [“Automated Text Analysis”](#)) in this handbook). Second, open-ended questions require great effort on the part of the respondent and may lead to participant refusal. Third and as a consequence of the aforementioned, researchers typically are able to ask more closed questions than open-ended questions – an aspect particularly relevant in organizational studies, for which participants are hard to recruit.

However, open-ended questions offer two potential benefits. First, participants can provide unusual and spontaneous answers to open-ended questions whereas their ability to respond is constrained by closed questions – if the investigator’s list of questions and answer options omits important aspects, participants cannot deliver insights in this regard. Second, open-ended questions require less in-depth knowledge than closed questions, and participants can provide insights into topics the investigator has not considered.

However, we question whether these potential benefits arise for written or online surveys. First, we doubt that in all domains participants can better evaluate than the investigator what could be relevant for the research question (Schuman and Presser 1981). Second, investigators can rely on pretesting to discover and fix omissions in their closed questions. Schuman and Presser (1979) demonstrate that answers to closed and open-ended questions are comparable when closed questions are adapted to insights from pretests (section [“Pre-test of the Questionnaire”](#)).

Our experience mirrors this discussion. We have noted that the quality of answers to open-ended questions in written surveys tends to be low. Consequently, we recommend focusing on closed questions in written and online surveys. However, we recommend that if possible, researchers change scales between dependent and independent variables to alleviate potential CMB (section [“Common Method Bias”](#)). We also recommend including an open-ended question at the end of the questionnaire allowing participants to provide further insights (e.g., “Are there additional issues you want to address?”).

Relatedly, including an open-ended question has an additional advantage: it lowers the probability that respondents engage in response substitution, which occurs when respondents want to express attitudes and beliefs that the researcher has not asked about. Informing respondents at the beginning of the questionnaire that they will have an opportunity to express any other thoughts in an open-ended format can reduce the biasing effects of response substitution (Gal and Rucker 2011).

For Closed Questions, What Answer Options Should Be Provided?

When selecting answer categories for closed questions, researchers have various options. For instance, closed questions can be multichotomous (e.g., automotive industry, financial industry, consumer durables), dichotomous (e.g., purchase vs. no purchase), or measured on a scale (e.g., Likert scale, frequency scale). When a Likert scale is chosen, a frequently asked question is how many response categories should be provided. The empirical evidence on this matter demonstrates that reliability tends to increase with an increasing number of scale options. However,

the marginal utility also tends to decrease. For instance, the quality of the provided answers is better for six than for four response categories (Preston and Colman 2000), and indicators with more response categories tend to display higher reliability (Alwin and Krosnick 1991). However, for more than seven response categories very few additional gains are observed. Thus, we recommend relying on five to seven answer categories.

In addition, participants can adapt their responses to seemingly arbitrary choices of scale labeling. For instance, Schwarz et al. (1991a) relied on two different labels for the same scale. In the first case, the scale ranged from 0 (“not at all successful”) to 10 (“extremely successful”). In the second, the anchor values ranged from -5 (“not at all successful”) to $+5$ (“extremely successful”). The result was that 34% of the respondents selected values between -5 and 0 on the scale from -5 to $+5$, but only 13% chose the equivalent values between 0 and 5 on a scale from 0 to 10. Since such behaviors are hard to anticipate, this example emphasizes the need for conducting pretests (section “[Pretest of the Questionnaire](#)”).

Decisions about the Question Wording

Ambiguous question formulations can obviously bias the responses to questionnaires. If respondents do not comprehend the question, they cannot provide adequate information (Baumgartner et al. 2018). Thus, to ensure comprehension, question wording should be (1) simple, (2) unambiguous, and (3) neutral.

Simplicity is important, as complex items likely encourage respondents to develop their own idiosyncratic understanding of questions and/or to use biased response styles. In this regard, researchers need to decide whether to rely on item reversals. Item reversal can be achieved in two ways (Baumgartner et al. 2018; Weijters and Baumgartner 2012; Weijters et al. 2009):

- *Negations*: using “not” (“I see myself as someone who is. . .” talkative vs. *not* talkative), affixal negation of adjectives (e.g., *dishonest*), or using negative adjectives or adverbs (e.g., *seldom*) or the negation of a noun with “no.”
- *Polar opposite items*: e.g., “I enjoy taking chances in buying new products” versus “I am very cautious in buying new products” or “talkative” versus “quiet.”

We recommend that researchers avoid negations and use polar opposite items carefully. Reversed items can offer advantages: they can inhibit stylistic responding (e.g., acquiescence bias; section “[Response Styles](#)”), act as cognitive “speed bumps,” and disrupt non-substantive response behaviors. However, and more pressingly, they might also lead to respondent confusion, lowering the quality of the collected data (e.g., Baumgartner et al. 2018; Weijters et al. 2009; Wong et al. 2003).

Simplicity can also be achieved when investigators

- Use rather short sentences and refrain from pyramiding (e.g., through relative clauses), as pyramiding or longer sentences can lead to complexity
- Avoid unnecessary variations in the format and structure of the questions
- Translate scientific language into the participants’ language

- Avoid requiring participants to do computations: for example, instead of directly asking “how much do you spend on average in a focal supermarket per month,” investigators should ask two simpler questions (e.g., “how often do you go to the focal supermarket per month?” and “How much do you spend on average per shopping trip at the focal supermarket?”)

Unambiguousness also avoids biases. Item ambiguity impairs the respondent’s ability to comprehend the question’s meaning and clearly undermines the respondent’s ability to provide accurate answers. When respondents are uncertain about how to respond to the item’s content, systematic response patterns are likely evoked and respondents are more likely to rely on their own implicit theories.

Strategies to avoid item ambiguity include the use of precise and concise language. For instance, universal expressions (e.g., “all,” “always”) or vague quantifiers (e.g., “often,” “many”) mean different things to different respondents. Thus, expressions such as “I read many books” would be less valid than “I read ten books a year” (Johnson 2004). Churchill and Iacobucci (2005, p. 248) summarize further aspects which need to be considered in this regard (e.g., “about” could mean “somewhere near” or “almost,” “like” distracts the attention of participants to the specific examples, and answers to the word “often” largely depend on the respondent’s frame of reference).

Importantly, researchers should use the same words for the same issues. In addition, they should clearly define terms that respondents might interpret differently. For instance, in research on customer reacquisition management in business-to-business markets, in-depth interviews revealed that customer defections can be defined differently. For some companies, only customers who completely stopped purchasing qualified as defected. For other companies, customers that had lowered their purchasing volume represented defectors. Thus, in designing a subsequent questionnaire, the researchers defined customer defection in the beginning of the questionnaire to avoid ambiguity (Homburg et al. 2019a).

Neutrality is also important when formulating questions. To achieve neutrality, researchers should avoid suggestive formulations and anticipations of answers (e.g., “Do you also agree?”). Finally, researchers should avoid loaded words (e.g., market economy vs. capitalism) that are likely to bias participants’ answers.

Decisions about the Question Sequence

Does the Sequence of Questions Foster a Pattern of Certain Answers?

As noted earlier, participants may form their opinions while answering the questionnaire (section “[Psychology of Survey Response](#)”). That is, answers provided to initial questions may activate information to answer questions at later points in the questionnaire (Schwarz 2003). Thus, researchers need to avoid generating survey data that yields only “self-generated validity” (Feldman and Lynch 1988, p. 421).

A prominent example from the field of political science that has been repeatedly affirmed since 1948 illustrates the effects of question sequence. In this example, participants were asked two questions:

- *Question A*: Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?
- *Question B*: Do you think a Communist country like Russia should let American newspaper reporters come in and send back to America the news as they see it?

In randomized experiments, the share of “yes” answers pertaining to *Question A* largely depended on the order of the two questions. The group who saw *Question A* and then *Question B* agreed to question *Question A* in 44% of the cases. However, the group that was asked in reversed sequence (first *Question B* then *Question A*) had higher agreement with *Question A* (70% agreement) (Schuman et al. 1983; Schwarz et al. 1991b report similar examples).

These sequence effects also become important in customer satisfaction surveys. A central decision in customer satisfaction surveys usually is at which point the investigator should ask participants to rate their overall level of satisfaction (e.g., “Overall, please indicate how satisfied you are with the company”). Should the rating be solicited before or after asking specific customer satisfaction questions (e.g., “Please indicate your satisfaction regarding the products,” “Please indicate your satisfaction regarding the after-sales service”)?

Researchers observed that depending on the sequence, different psychological processes set in. If overall customer satisfaction is measured at the beginning of the survey, emotions typically dominate the assessment (e.g., gut feeling). However, if total customer satisfaction is measured at the end of the survey, the same question will trigger cognitive rather than emotional processes. Particularly in this case, the answers to the overall customer satisfaction question will likely depend on the replies to the specific customer satisfaction responses participants provided before.

Which order is the most appropriate largely depends on the research context. Researchers may measure overall customer satisfaction at the beginning (end) of the questionnaire for low (high) involvement products. As emotions (cognitions) are likely to be an important factor when purchasing low (high) involvement products, a more emotional (cognitive) response to overall customer satisfaction might be more insightful.

The literature discusses some measures researchers can employ to reduce the effects of question sequence. One approach could be to randomize the question order, which can easily be achieved in online surveys: researchers present the same questions in different orders to participants. Although this approach might seem intuitively appealing, the downside is that it cannot prevent order effects from occurring on the individual level. In addition, randomization may disrupt the logical order of the questions. An alternative might be to simply tell respondents that the order of presentation is random and therefore of no relevance (Krosnick et al. 1990).

The inclusion of buffer questions might also help to reduce sequence effects. For instance, researchers can include buffer questions between items they expect to evoke a large sequence effect – a tactic that likely lowers artificial correlations among items (Weijters et al. 2009).

Do the Introductory Questions Motivate Respondents to Participate in the Survey? Does the Questionnaire Seem Structured?

In general, we recommend that researchers start with an interesting opening question. This question should motivate participants to take part in the survey, and its importance increases in light of the declining response rates for questionnaires (section “[Non-sampling Bias](#)”). After the opening question, researchers should move from general to specific questions and ask sensitive questions later in the survey. Rapport is established as respondents answer general and nonthreatening questions early in the survey. Finally, when researchers cannot access different sources for evaluating independent and dependent variables, they should first assess the dependent and then the independent variables. Thereby, they lower potential demand effects (section “[Common Method Bias](#)”).

Decisions About the Survey Layout and Pretest of the Questionnaire

Does the Questionnaire Have a Professional Appearance? Does the Format of the Questionnaire Lead to Easy Completion?

The layout of the questionnaire is a critical success factor. Participants perceive the questionnaire to represent the net of the asked questions, layout and design, and logical structure and architecture. All these aspects may affect the effort participants put into answering the survey and thus affect the likelihood of biases to occur (Presser et al. 2004).

In general, the questionnaire “should look as sharp as your résumé” (Iacobucci and Churchill 2010, p. 221). A clean appearance signals interest for the topic and also emphasizes the researcher’s trustworthiness. A questionnaire’s layout may also create the impression of short processing time and thereby increase response rates, lowering the threat of non-response bias (section “[Non-response Bias](#)”). Finally, the instructions should be clearly articulated and the question flow should be evident and supportive.

Pretest of the Questionnaire

Do Respondents Understand the Questions? Do Respondents Engage in Specific Cognitive Patterns?

While textbooks on the development of a survey are plentiful, there are “no silver bullets of questionnaire design” (Schwarz 2003, p. 593), and even experts struggle to anticipate all potential difficulties that arise when participants address the survey questions. Thus, “there is no substitute for in-depth pretesting” (Weijters and Baumgartner 2012, p. 565) and “*all* researchers should use pretests prior to running their main studies” (Hulland et al. 2018, p. 104, emphasis in original).

The goals an investigator may want to achieve when running a pretest are varied and include the following:

- Identification of ambiguous questions and problems of understanding
- Detection of incompleteness of answer categories
- Identification of potential information gaps of respondents
- Verification of the time required to complete the questionnaire

- A preliminary indication of validity and reliability of the measurement (but taking the pretest sample size into account)

Three methods usually pertain to pretests. First, the debriefing method permits pre-testers to complete the questionnaire and then discuss potential issues (for instance, online surveys allow participants to make notes that can be discussed afterwards). Online pretests allow investigators to track pre-tester behaviors, which they can discuss in debriefing meetings. For instance, they could track response time or employ eye-tracking software to identify problematic survey sections (Baumgartner et al. 2018). Second, the protocol method allows pre-testers to raise questions when filling out the questionnaire. Third, the think-aloud method (also referred to as cognitive interviews) require pre-testers to verbalize their thoughts when filling out the questionnaire (Ericsson and Simon 1980; Presser et al. 2004; Weijters et al. 2009).

Data Collection

As our focus is on the design aspect of survey research, we do not discuss in detail various data collection methods but confine our discussion to the importance of sample structure, sample size, and sample response rate. In this handbook, Bornemann and Hattula (chapter ► “Experiments in Market Research”) provide a more detailed discussion of sampling procedures.

What Is the Ideal Structure of the Sample?

Sampling designs can be divided into probability and non-probability samples. Probability samples are random samples. For instance, a simple random sample assumes that observations are selected by chance from the population. While most statistical tests assume probability sampling, applied researchers often rely on non-probability sampling since random samples are often hard to obtain (Short et al. 2002). As already mentioned, in many situations the population of interest is hard to define (section “Non-sampling Bias”), which limits investigators’ ability to draw random samples.

Literature distinguishes four principal ways of non-probability sampling. First, researchers might rely on convenience samples – that is, they select a sample on the basis of accessibility. Convenience samples are the most typical sampling form in marketing and management research (Short, Ketchen, and Palmer (Short et al. 2002) report 42%, and Hulland et al. (2018) report 43%). However, convenience samples are likely to be unrepresentative of the overall population. Importantly, in consumer research debate is ongoing about the use of a particular type of convenience sample – student samples (Peterson 2001).

Second, researchers may apply quota sampling where units are drawn to approximate known proportions found in a population. Third, investigators may rely on snowball sampling, in which researchers identify a few participants from the

population of interest and ask them to forward the questionnaire. This technique is useful when determining the population of interest is highly challenging (e.g., as when market research focuses on extreme sports). Fourth, sampling may be based on typicality. Here, researchers do not focus on representative participants but on participants from whom they expect valuable input. This sampling approach is often used during pretests (section “[Pretest of the Questionnaire](#)”; Iacobucci and Churchill 2010; Short et al. 2002).

How Large Should the Sample Be?

The question of sample size is hard to answer without considering the specific research setting. A rule of thumb for regression analysis is to rely on 10 observations per parameter (Hair et al. 2010). However, in applied research projects, this ratio is usually larger. For instance, the ratio of sample size to variables could be 70.63 (Short et al. 2002).

Small samples can have important disadvantages. Smaller samples can lead to less reliable test statistics, as standard errors may increase. In addition, small samples may limit the types of statistical techniques that can be applied effectively. However, large samples can also be problematic, as negligible effects might become statistically significant owing to high statistical power.

Finally and importantly, no clear relationship exists between sample size and the sample’s representativeness. Opinion polls have revealed that if researchers invest substantial effort in acquiring large samples, representativeness of the final sample might suffer. Therefore, sampling should prioritize representativeness over sample size (Assael and Keon 1982; Krosnick 1999; Rogelberg and Stanton 2007).

How Can we Achieve a High Response Rate?

In section “[Non-response Bias](#),” we discussed several techniques for countering potential non-response bias. As the same techniques should help to increase the response rate (Anseel et al. 2010), we do not describe these actions here. However, we note that the size of a response rate does not indicate the representativeness of a sample. In some instances, lower response rates can lead to more accurate results than higher response rates (Visser et al. 1996). Recent evidence confirms this observation and emphasizes that the assumed positive relationship between response rate and survey quality does not always hold (The American Association for Public Opinion Research 2016).

Measurement Evaluation

To What Extent Do the Survey Questions Measure What they Are Supposed to Measure?

After the questionnaires have been returned to the researcher, the quality of their measurement can be evaluated. In-depth discussions of how survey researchers can demonstrate the reliability and validity of their constructs appear elsewhere in this handbook (chapter ► “[Structural Equation Modeling](#)” by Baumgartner and Weijters).

In the following, we briefly mention which statistics researchers typically report to provide initial guidance. All of the presented techniques require researchers' use of multiple-item scales (section "[Decisions About the Question Content](#)").

Researchers can demonstrate the reliability of their constructs by reporting Cronbach's alpha and/or composite reliability values (both have threshold values of >0.70). Both estimate the squared correlation between a construct and an unweighted sum of its items; composite reliability represents a generalization of Cronbach's alpha because it does not assume equal loading across items.

Through confirmatory factor analysis, investigators can demonstrate convergent and discriminant validity. Convergent validity (which affirms that the items correlate with other concepts they should be related to) can be established for individual items by squaring the individual standardized factor loadings (threshold value: >0.40). At the construct level, researchers calculate the average variance extracted (threshold value: >0.50) (Baumgartner and Weijters (chapter ▶ "[Structural Equation Modeling](#)") offer more flexible cut-off values).

Researchers can rely on the Fornell-Larcker criterion (1981) to establish discriminant validity (which affirms the items are uncorrelated with concepts they are not related to). For each pair of constructs, the square root of the average variance extracted for each construct needs to exceed the correlation between them.

Data Analysis

How Are the Examined Phenomena Related? How Can the Results Be Illustrated?

Analysis of the acquired data is the final step in the typical survey research process. We mention this step only briefly for two reasons. First, important decisions in survey research precede the statistical analyses. Hence, the careful design of the questionnaire is a necessary condition for the empirical analysis to yield valid results. Second, many statistical techniques can be used to analyze survey data and their in-depth discussions appear in designated method chapters of this handbook: most commonly, researchers rely on regression analysis (chapter ▶ "[Regression Analysis](#)" by Skiera et al.), structural equation modeling (Baumgartner and Weijters 2017), partial least squares structural equation modeling (chapter ▶ "[Partial Least Squares Structural Equation Modeling](#)" by Sarstedt et al.), or analysis of variance (chapter ▶ "[Analysis of Variance](#)" Landwehr).

Endogeneity in Survey Research

Addressing endogeneity concerns is typically important when investigators rely on observational data such as survey data (Sande and Ghosh 2018). For instance, the study of Huang and Sudhir (2021) demonstrates the need for survey researchers to address endogeneity concerns. Their study shows that not accounting for endogeneity leads to an underestimation of the true effect of service satisfaction on

customer loyalty. Such underestimation likely leads to less than optimal investment decisions in business practice. While an in-depth discussion of endogeneity is provided elsewhere in this handbook (chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#) by Ebbes et al.), we highlight several aspects that might be particularly important for survey researchers.

Conceptually, endogeneity concerns refer to alternative or rival causal explanations of the findings. Statistically, endogeneity concerns arise if the explanatory variables correlate with the error terms, for example in regression analyses. Literature distinguishes three sources of endogeneity: (1) omitted variables, (2) measurement error, and (3) simultaneity (e.g., Cameron and Trivedi 2005; Kennedy 2008; Rossi 2014).

First, with regard to omitted variables, survey research has an important advantage over secondary data. In principle, survey researchers can rule out omitted variable bias by including control variables (section [“Decision About the Survey Content”](#)). If researchers are able to identify all potentially omitted variables when collecting the data, they can estimate a “rich data model,” through which they directly address endogeneity concerns and meet the “standard recommendation for limiting omitted variable bias” (Rossi 2014, p. 657). However, although these endeavors in theory may lower or rule out concerns of omitted variable bias, the success of a rich data approach depends on the researchers’ ability to identify all relevant control variables. If all potentially omitted variables cannot be identified or if such variables cannot be measured, survey researchers can employ the procedures Ebbes et al. (chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#)) outline.

Second, while measurement error in the dependent variable is absorbed in the residual error term, measurement error in the independent variables represents a principal form of endogeneity (Kennedy 2008; Vomberg and Wies (chapter ► [“Panel Data Analysis: A Non-technical Introduction for Marketing Researchers”](#)) in this handbook). If survey researchers rely on multiple-item measures of their independent variables, they are able to directly evaluate measurement error and through structural equation modeling can directly rule out such concerns (e.g., Grewal et al. 2013; Baumgartner and Weijters (chapter ► [“Structural Equation Modeling”](#)) in this handbook).

Measurement error may also arise in the form of biases in surveys. For instance, CMB (section [“Common Method Bias”](#)) or key informant bias (section [“Key Informant Bias”](#)) also lead to correlation between the independent variables and the residual error term. Therefore, the procedures we outlined in the previous sections can also serve as ways to account for endogeneity (e.g., procedural remedies to deal with CMB in section [“Common Method Bias”](#)).

Third, simultaneity – for instance, in the form of reverse causality – may affect cross-sectional survey data. Potential simultaneity concerns should be considered in the development stage of the questionnaire. In this regard, potential instrumental variables can be included. For instance, Homburg et al. (2012a) studied the influence of comprehensive market performance measurement systems on market knowledge. To alleviate simultaneity concerns (i.e., market knowledge drives companies’ use of

Table 3 Linking biases with steps of the survey research process: In which steps can researchers address potential biases?

Name	Definition of bias Distortion of the samples covariance structure. ...	Reasons for occurrence	Remedies	Steps in survey research process
Common method bias (section “ Common Method Bias ”)	... that arises due to the fact that the same data source was used for measuring both independent and dependent variables	Plenty of explanations, e.g.: implicit theories consistency motif satisficing	Different sources for independent and dependent variables Collect independent and dependent variables at different points in time Increase distance between independent and dependent variables Different scales for independent and dependent variables Inclusion of marker variable or direct measure of source of common method variance	Data collection Data collection Decision about question sequence Decision about the question format Decision about the survey content
Key informant bias (section “ Key Informant Bias ”)	... that arises due to the fact that the data collection has taken place through key informants providing information about a larger social unit	Positional bias Knowledge deficiencies	Adequate selection of key informants Triangulation Avoid demanding evaluations of complex social judgments (if possible in the particular research context) Report average job and company experience of participants	Data collection Decision about survey content Decision about survey content
Social desirability (section “ Social Desirability ”)	... that arises due to the tendency of respondents to reply in a manner that will be viewed favorably by others	Level of awareness: unconscious self-deception versus deliberate impression management Content domain: egoistic responde tendencies versus moralistic response tendencies	Ensuring anonymity of respondents Reinforce that there are no right or wrong answers Encourage honest answers Indirect questioning Distance between respondent and researcher With caution: Inclusion of negatively valenced items Social sensitive questions asked at the end of the questionnaire	Decision about question sequence and Decision about survey layout Decision about question wording Selection of research method Decision about question wording Decision about question sequence

<p>Response styles (section “Response Styles”)</p>	<p>... that arises due to the fact that respondents, regardless of the question content, favor certain response categories</p>	<p>Plenty of explanations, e.g.: Confirmation bias Satisficing Impression management Item ambiguity Personality traits of respondent</p>	<p><i>With caution:</i> Inclusion of balanced scales (i.e., negative and positive items) <i>Potentially:</i> Additional items for deriving measures of response styles</p>	<p>Decision about question wording Decision about research variables</p>
<p>Non-sampling bias (section “Non-sampling Bias”)</p>	<p>... that arises due to the fact that the population is not adequately represented in the original sample</p>	<p>Lack of knowledge about population</p>	<p>Relying on sampling frame Inclusion of variables on which selection likely appear in the questionnaire</p>	<p>Data collection Decision about research variables</p>
<p>Non-response bias (section “Nonresponse Bias”)</p>	<p>... that arises due to the fact that the structure of the final sample does not coincide with the structure of the original sample</p>	<p>Participant declines survey due to (1) personal, (2) organizational, or (3) survey factors Inability to reach participant</p>	<p>Increase opportunity to participate Emphasize importance of the survey Decrease perceived costs of participation Raise perceived utility of participation</p>	<p>Data collection Selection of survey method Cover Letter Decision about survey layout Decision about question content Pretest of the questionnaire Data collection</p>

comprehensive market performance measurement systems), the authors relied on an instrumental variable approach and directly measured an instrumental variable in their questionnaire.

Conclusion

Surveys are flexible and powerful ways to address research questions. Indeed, in many situations it is hard to imagine how topics can be studied without directly asking participants questions. However, effective survey research requires careful survey design. As respondents usually develop answers to surveys in the course of completing a survey, investigators must craft their surveys meticulously to avoid potential biases.

To help researchers construct their surveys, we first outlined the psychology of survey response and discussed important biases. Awareness of these biases helps survey researchers develop surveys that are less susceptible to biases. We have also described the general survey process and discussed important decisions investigators face in each step, and in Table 3, we connect those insights: we delineate the types of biases and link these biases to steps in the survey process where researchers can alleviate these biasing effects.

In closing, we note that many of the issues we discussed arise in other forms of research. For example, decisions regarding question content and data collection are equally important for experimental studies (chapter ▶ [“Experiments in Market Research”](#) by Bornemann and Hattula; chapter ▶ [“Field Experiments”](#) by Valli et al.), and our discussion of CMB is applicable when experimental studies measure mediating variables (chapter ▶ [“Mediation Analysis in Experimental Research”](#) by Koschate-Fischer and Schwillle).

Cross-References

- ▶ [Analysis of Variance](#)
- ▶ [Automated Text Analysis](#)
- ▶ [Challenges in Conducting International Market Research](#)
- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Experiments in Market Research](#)
- ▶ [Measuring Customer Satisfaction and Customer Loyalty](#)
- ▶ [Mediation Analysis in Experimental Research](#)
- ▶ [Partial Least Squares Structural Equation Modeling](#)
- ▶ [Regression Analysis](#)
- ▶ [Structural Equation Modeling](#)

References

- Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research*, 20(1), 139–181.
- Andréß, H. J., Golsch, K., & Schmidt, A. W. (2013). Applied panel data analysis for economic and social surveys. Springer-Verlag Berlin Heidelberg.
- Anseel, F., Lievens, F., Schollaert, E., & Choragwicka, B. (2010). Response rates in organizational science, 1995–2008: A meta-analytic review and guidelines for survey researchers. *Journal of Business and Psychology*, 25(3), 335–349.
- Arora, R. (1982). Validation of an SOR model for situation, enduring, and response components of involvement. *Journal of Marketing Research*, 19, 505–516.
- Assael, H., & Keon, J. (1982). Nonsampling vs. sampling errors in survey research. *Journal of Marketing*, 46, 114–123.
- Baruch, Y. (1999). Response rate in academic studies – A comparative analysis. *Human Relations*, 52(4), 421–438.
- Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 13(2), 139–161.
- Baumgartner, H., & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156.
- Baumgartner, H., & Steenkamp, J. B. E. (2006). Response biases in marketing research. In R. Grover & M. Vriens (Eds.), *The handbook of marketing research. Uses, misuses, and future advances* (pp. 95–109). Thousand Oaks: Sage.
- Baumgartner, H., & Weijters, B. (2017). Measurement models for marketing constructs. In B. Wierenga & R. Van der Lans (Eds.), *Handbook of marketing decision models* (International series in operations research & management science) (Vol. 254). Cham: Springer.
- Baumgartner, H., Weijters, B., & Pieters, R. (2018). Misresponse to survey questions: A conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts. *Journal of Marketing Research*, 55(6), 869–883.
- Bearden, W. O., & Netemeyer, R. G. (1999). *Handbook of marketing scales: Multi-item measures for marketing and consumer behavior research*. Thousand Oaks, Calif: Sage.
- Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44(2), 175–184.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco: Jossey-Bass.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17(4), 665–678.
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, 22(1), 69.
- Bryman, A., & Bell, E. (2015). *Business research methods* (4th ed.). Oxford University Press.
- Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods*, 5(2), 159–172.
- Cabooter, E., Millet, K., Pandelaere, M., & Weijters, B. (2012). The 'I' in extreme responding. Paper presented at the European Marketing Academy Conference, Lisbon.
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics – Methods and applications*. New York: Cambridge University Press.
- Carson, S. J., & Ghosh, M. (2019). An integrated power and efficiency model of contractual channel governance: Theory and empirical evidence. *Journal of Marketing*, 83(4), 101–120.
- Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825.

- Certo, S. T., Busenbark, J. R., Woo, H. S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, *37*(13), 2639–2657.
- Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500–508.
- Chang, S. J., Van Witteloostuijn, A., & Eden, L. (2010). From the editors: Common method variance in international business research. *Journal of International Business Studies*, *41*, 178–184.
- Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, *57*(1), 62–79.
- Churchill, G. A., & Iacobucci, D. (2005). *Marketing research. Methodological foundations* (9th ed.). Mason: South-Western Cengage Learning.
- Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web-or internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836.
- Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research*, *24*, 315–318.
- Cote, J. A., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research*, *14*(4), 579–582.
- Crampton, S. M., & Wagner, J. A., III. (1994). Percept-percept inflation in microorganizational research: An investigation of prevalence and effect. *Journal of Applied Psychology*, *79*(1), 67.
- De Heer, W., & De Leeuw, E. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In *Survey nonresponse* (p. 41). New York: Wiley.
- De Jong, M. G., Steenkamp, J. B. E., Fox, J. P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115.
- De Jong, M. G., Pieters, R., & Fox, J. P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*(1), 14–27.
- De Jong, M. G., Fox, J. P., & Steenkamp, J. B. E. (2015). Quantifying under- and overreporting in surveys through a dual-questioning-technique design. *Journal of Marketing Research*, *52*(6), 737–753.
- De Langhe, B., Puntoni, S., Fernandes, D., & Van Osselaer, S. M. J. (2011). The anchor contraction effect in international marketing research. *Journal of Marketing Research*, *48*(2), 366–380.
- Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*(3), 434–449.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251.
- Evans, M. G. (1985). A Monte Carlo study of the effects of correlated method variance in moderated multiple regression analysis. *Organizational Behavior and Human Decision Processes*, *36*(3), 305–323.
- Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73*(3), 421.
- Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303–315.
- Fornell, C., & Larcker, D. F. (1981). Structural Equation Models with Unobservable Variables and Measurement Error: Algebra and Statistics. *Journal of Marketing Research*, *18*(3):382–388. <https://doi.org/10.1177/002224378101800313>
- Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: nature, purpose, and findings. *Journal of Marketing*, *60*, 7–18.

- Gal, D., & Rucker, D. D. (2011). Answering the unasked question: Response substitution in consumer surveys. *Journal of Marketing Research*, 48(February), 185–195.
- Gannon, M. J., Nothern, J. C., & Carroll, S. J. (1971). Characteristics of nonrespondents among workers. *Journal of Applied Psychology*, 55(6), 586.
- Ghosh, M., & John, G. (2005). Strategic fit in industrial alliances: An empirical test of governance value analysis. *Journal of Marketing Research*, 42(3), 346–357.
- Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, 44(1), 196–210.
- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3), 213–224.
- Grayson, K. (2007). Friendship versus business in marketing relationships. *Journal of Marketing*, 71(4), 121–139.
- Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, 56(3), 328–351.
- Grewal, R., Kumar, A., Mallapragada, G., & Saini, A. (2013). Marketing channels in foreign markets: Control mechanisms and the moderating role of multinational corporation headquarters–subsidiary relationship. *Journal of Marketing Research*, 50(3), 378–398.
- Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 646–675.
- Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. *Public Opinion Quarterly*, 68(1), 2–31.
- Gruner, R. L., Vomberg, A., Homburg, C., & Lukas, B. A. (2019). Supporting new product launches with social media communication and online advertising: Sales volume and profit implications. *Journal of Product Innovation Management*, 36(2), 172–195.
- Gupta, N., Shaw, J. D., & Delery, J. E. (2000). Correlates of response outcomes among organizational key informants. *Organizational Research Methods*, 3(4), 323–347.
- Hagen, L. (2020). Pretty healthy food: How and when aesthetics enhance perceived healthiness. *Journal of Marketing*, forthcoming.
- Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River: Prentice Hall.
- Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, 69(3), 192.
- Heide, J. B., Wathne, K. H., & Rokkan, A. I. (2007). Interfirm monitoring, social contracts, and relationship outcomes. *Journal of Marketing Research*, 44(3), 425–433.
- Heide, J. B., Kumar, A., & Wathne, K. H. (2014). Concurrent sourcing, governance mechanisms, and performance outcomes in industrial value chains. *Strategic Management Journal*, 35(8), 1164–1185.
- Helgeson, J. G., Voss, K. E., & Terpening, W. D. (2002). Determinants of mail-survey response: Survey design factors and respondent factors. *Psychology & Marketing*, 19(3), 303–328.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29.
- Himmelfarb, S., & Lickteig, C. (1982). Social desirability and the randomized response technique. *Journal of Personality and Social Psychology*, 43, 710–717.
- Holbrook, A. L., & Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly*, 74(2), 328–343.
- Homburg, C. (2020). *Marketingmanagement: Strategie – Instrumente – Umsetzung – Unternehmensführung* (7th ed.). Heidelberg: Springer.
- Homburg, C., & Klarmann, M. (2009). Multi informant-designs in der empirischen betriebswirtschaftlichen Forschung. *Die Betriebswirtschaft*, 69(2), 147.

- Homburg, C., & Krohmer, H. (2008). Der Prozess der Marktforschung: Festlegung der Datenerhebungsmethode, Stichprobenbildung und Fragebogengestaltung. In A. Herrmann, C. Homburg, & M. Klarmann (Eds.), *Handbuch Marktforschung*. Heidelberg: Springer Gabler.
- Homburg, C., Jensen, O., & Klarmann, M. (2005). *Die Zusammenarbeit zwischen Marketing und Vertrieb-eine vernachlässigte Schnittstelle* (Vol. 86). Mannheim: Inst. für Marktorientierte Unternehmensführung, Univ. Mannheim.
- Homburg, C., Grozdanovic, M., & Klarmann, M. (2007). Responsiveness to customers and competitors: The role of affective and cognitive organizational systems. *Journal of Marketing*, 71(3), 18–38.
- Homburg, C., Klarmann, M., & Schmitt, J. (2010). Brand awareness in business markets: When is it related to firm performance? *International Journal of Research in Marketing*, 27(3), 201–212.
- Homburg, C., Müller, M., & Klarmann, M. (2011). When should the customer really be king? On the optimum level of salesperson customer orientation in sales encounters. *Journal of Marketing*, 75(2), 55–74.
- Homburg, C., Artz, M., & Wieseke, J. (2012a). Marketing performance measurement systems: Does comprehensiveness really improve performance? *Journal of Marketing*, 76(3), 56–77.
- Homburg, C., Jensen, O., & Hahn, A. (2012b). How to organize pricing? Vertical delegation and horizontal dispersion of pricing authority. *Journal of Marketing*, 76(5), 49–69.
- Homburg, C., Klarmann, M., Reimann, M., & Schilke, O. (2012c). What drives key informant accuracy? *Journal of Marketing Research*, 49(August), 594–608.
- Homburg, C., Schwemmler, M., & Kuehnl, C. (2015a). New product design: Concept, measurement, and consequences. *Journal of Marketing*, 79(3), 41–56.
- Homburg, C., Vomberg, A., Enke, M., & Grimm, P. H. (2015b). The loss of the marketing department's influence: Is it really happening? And why worry? *Journal of the Academy of Marketing Science*, 43(1), 1–13.
- Homburg, C., Gwinner, O., & Vomberg, A. (2019a). *Customer reacquisition in business-to-business contexts*. Working paper.
- Homburg, C., Lauer, K., & Vomberg, A. (2019b). The multichannel pricing dilemma: Do consumers accept higher offline than online prices? *International Journal of Research in Marketing*, 36(4), 597–612.
- Homburg, C., Vomberg, A., & Muehlhaeuser, S. (2020). Design and governance of multichannel sales systems: Financial performance consequences in business-to-business markets. *Journal of Marketing Research*, 57(6), 1113–1134.
- Huang, G., & Sudhir, K. (2021). The Causal Effect of Service Satisfaction on Customer Loyalty. *Management Science*, 67(1), 317–341.
- Hulland, J. (2019). In through the out door. *Journal of the Academy of Marketing Science*, 47(1), 1–3.
- Hulland, J., & Miller, J. (2018). Keep on Turkin? *Journal of the Academy of Marketing Science*, 46(5), 789–794. <https://doi-org.proxy-ub.rug.nl/10.1007/s11747-018-0587-4>
- Hulland, J., Baumgartner, H., & Smith, K. M. (2018). Marketing survey research best practices: Evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science*, 46(1), 92–108.
- Iacobucci, D. (2013). *Marketing models: Multivariate statistics and marketing analytics* (International Edition). South-Western: Cengage Learning.
- Iacobucci, D., & Churchill, G. A. (2010). *Marketing research. Methodological foundations* (10th ed.). Mason: South-Western Cengage Learning.
- Jansen, J. J., Van Den Bosch, F. A., & Volberda, H. W. (2005). Managing potential and realized absorptive capacity: How do organizational antecedents matter? *Academy of Management Journal*, 48(6), 999–1015.
- Jansen, J. J., Van Den Bosch, F. A., & Volberda, H. W. (2006). Exploratory innovation, exploitative innovation, and performance: Effects of organizational antecedents and environmental moderators. *Management science*, 52(11), 1661–1674.

- John, L. K., Loewenstein, G., Acquisti, A., & Vosgerau, J. (2018). When and why randomized response techniques (fail to) elicit the truth. *Organizational Behavior and Human Decision Processes*, 148, 101–123.
- Johnson, J. A. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research*, 39(2), 273–302.
- Johnson, R. E., Rosen, C. C., & Djurdjevic, E. (2011). Assessing the impact of common method variance on higher order multidimensional constructs. *Journal of Applied Psychology*, 96(4), 744.
- Kennedy, P. (2008). *A guide to econometrics* (6th ed.). Cambridge, MA: Wiley-Blackwell.
- Klarmann, M. (2008). *Methodische Problemfelder der Erfolgsfaktorenforschung: Bestandsaufnahme und empirische Analysen* (Doctoral dissertation).
- Knowles, E. S., & Condon, C. A. (1999). Why people say “yes”: A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2), 379.
- Kothandapani, V. (1971). Validation of feeling, belief, and intention to act as three components of attitude and their contribution to prediction of contraceptive behavior. *Journal of Personality and Social Psychology*, 19(3), 321.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50(1), 537–567.
- Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology*, 59(6), 1140.
- Kumar, N., Stern, L. W., & Anderson, J. C. (1993). Conducting interorganizational research using key informants. *Academy of Management Journal*, 36(6), 1633–1651.
- Kumar, N., Scheer, L. K., & Steenkamp, J. B. E. (1995). The effects of supplier fairness on vulnerable resellers. *Journal of Marketing Research*, 32, 54–65.
- Kumar, A., Heide, J. B., & Wathne, K. H. (2011). Performance implications of mismatched governance regimes across external and internal relationships. *Journal of Marketing*, 75(2), 1–17.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815–852.
- Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical Turk samples. *SAGE Open*, 6(1), 2158244016636433.
- Lindell, M. K., & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology*, 86(1), 114.
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433–442.
- Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods*, 4(2), 192.
- MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: Causes, mechanisms, and procedural remedies. *Journal of Retailing*, 88(4), 542–555.
- McElheran, K. (2015). Do market leaders lead in business process innovation? The case (s) of e-business adoption. *Management Science*, 61(6), 1197–1216.
- McKelvey, B. (1975). Guidelines for the empirical classification of organizations. *Administrative Science Quarterly*, 20, 509–525.
- Messick, S. (2012). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 161–200). Hillsdale: Lawrence Erlbaum.
- Mick, D. G. (1996). Are studies of dark side variables confounded by socially desirable responding? The case of materialism. *Journal of Consumer Research*, 23(2), 106–119.
- Mizik, N., & Jacobson, R. (2008). The financial value impact of perceptual brand attributes. *Journal of Marketing Research*, 45(1), 15–32.

- Moosbrugger, H. (2008). Klassische Testtheorie (KTT). In *Testtheorie und Fragebogenkonstruktion* (pp. 99–112). Berlin/Heidelberg: Springer.
- Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77*(1), 261–286.
- Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263–280.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
- Ostroff, C., Kinicki, A. J., & Clark, M. A. (2002). Substantive and operational issues of response bias across levels of analysis: An example of climate-satisfaction relationships. *Journal of Applied Psychology, 87*(2), 355.
- Palmatier, R. W. (2016). Improving and publishing at JAMS: Contribution and positioning. *Journal of the Academy of Marketing Science, 44*(6), 655–659.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Brand, D. N. Jackson, D. E. Wiley, & S. Messick (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah: L. Erlbaum.
- Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66*(6), 1025–1060.
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163.
- Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28*(3), 450–461.
- Phillips, L. W. (1981). Assessing measurement error in key informant reports: A methodological note on organizational analysis in marketing. *Journal of Marketing Research, 18*, 395–415.
- Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management, 12*(4), 531–544.
- Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *The Leadership Quarterly, 1*(2), 107–142.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879.
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.
- Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109–130.
- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15.
- Price, J. L., & Mueller, C. W. (1986). *Handbook of organizational measurements*. Marshfield: Pittman.
- Raval, D. (2020). Whose voice do we hear in the marketplace? Evidence from consumer complaining behavior. *Marketing Science, 39*(1), 168–187.
- Rindfleisch, A., & Heide, J. B. (1997). Transaction cost analysis: Past, present, and future applications. *Journal of marketing, 61*(4), 30–54.
- Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research, 45*(3), 261–279.

- Rogelberg, S. G., & Stanton, J. M. (2007). Introduction: Understanding and Dealing With Organizational Survey Nonresponse. *Organizational Research Methods, 10*(2), 195–209. <https://doi.org/10.1177/1094428106294693>
- Rogelberg, S. G., Fisher, G. G., Maynard, D. C., Hakel, M. D., & Horvath, M. (2001). Attitudes toward surveys: Development of a measure and its relationship to respondent behavior. *Organizational Research Methods, 4*(1), 3–25.
- Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science, 33*(5), 655–672.
- Sa Vinhas, A., & Heide, J. B. (2015). Forms of competition and outcomes in dual distribution channels: The distributor's perspective. *Marketing Science, 34*(1), 160–175.
- Sande, J. B., & Ghosh, M. (2018). Endogeneity in survey research. *International Journal of Research in Marketing, 35*(2), 185–204.
- Schmidt, J., & Bijmolt, T. H. (2019). Accurately measuring willingness to pay for consumer goods: A meta-analysis of the hypothetical bias. *Journal of the Academy of Marketing Science, 48*(3), 499–518.
- Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review, 44*, 692–712.
- Schuman, H., & Presser, S. (1981). The attitude-action connection and the issue of gun control. *The Annals of the American Academy of Political and Social Science, 455*(1), 40–47.
- Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. New York: Sage.
- Schuman, H., Kalton, G., & Ludwig, J. (1983). Context and contiguity in survey questionnaires. *Public Opinion Quarterly, 47*(1), 112–115.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93.
- Schwarz, N. (2003). Self-reports in consumer research: The challenge of comparing cohorts and cultures. *Journal of Consumer Research, 29*(4), 588–594.
- Schwarz, N., & Scheuring, B. (1992). Selbstberichtete Verhaltens- und Symptommhäufigkeiten: Was Befragte aus Antwortvorgaben des Fragebogens lernen. *Zeitschrift für klinische Psychologie*.
- Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991a). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570–582.
- Schwarz, N., Strack, F., & Mai, H. P. (1991b). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55*(1), 3–23.
- Seidler, J. (1974). On using informants: A technique for collecting quantitative data and controlling measurement error in organization analysis. *American Sociological Review, 39*, 816–831.
- Short, J. C., Ketchen, D. J., Jr., & Palmer, T. B. (2002). The role of sampling in strategic management research on performance: A two-study analysis. *Journal of Management, 28*(3), 363–385.
- Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods, 13*(3), 456–476.
- Steenkamp, J. B. E., De Jong, M. G., & Baumgartner, H. (2010). Socially desirable response tendencies in survey research. *Journal of Marketing Research, 47*(2), 199–214.
- Sudman, S., & Blair, E. (1999). Sampling in the twenty-first century. *Journal of the Academy of Marketing Science, 27*(2), 269–277.
- Tellis, G. J., & Chandrasekaran, D. (2010). Extent and impact of response biases in cross-national survey research. *International Journal of Research in Marketing, 27*(4), 329–341.
- The American Association for Public Opinion Research. (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (9th ed.). AAPOR. https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf
- Thompson, L. F., & Surface, E. A. (2007). Employee surveys administered online: Attitudes toward the medium, nonresponse, and data representativeness. *Organizational Research Methods, 10*(2), 241–261.

- Tomaskovic-Devey, D., Leiter, J., & Thompson, S. (1994). Organizational survey nonresponse. *Administrative Science Quarterly*, 39, 439–457.
- Tortolani, R. (1965). Introducing bias intentionally into survey techniques. *Journal of Marketing Research*, 2, 51–55.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
- Van Rosmalen, J., Van Herk, H., & Groenen, P. J. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, 47(1), 157–172.
- Visser, P. S., Krosnick, J. A., Marquette, J., & Curtin, M. (1996). Mail surveys for election forecasting? An evaluation of the Columbus Dispatch poll. *Public Opinion Quarterly*, 60(2), 181–227.
- Vomberg, A., Homburg, C., & Bornemann, T. (2015). Talented people and strong brands: The contribution of human capital and brand equity to firm value. *Strategic Management Journal*, 36(13), 2122–2131.
- Vomberg, A., Homburg, C., & Gwinner, O. (2020). Tolerating and managing failure: An organizational perspective on customer reacquisition management. *Journal of Marketing*, 84(5), 117–136.
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63–69.
- Wathne, K. H., Heide, J. B., Mooi, E. A., & Kumar, A. (2018). Relationship governance dynamics: The roles of partner selection efforts and mutual investments. *Journal of Marketing Research*, 55(5), 704–721.
- Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, 49(5), 737–747.
- Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, 36(3), 409–422.
- Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing*, 26(1), 2–12.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010a). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3), 236–247.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, 34(2), 105–121.
- Weijters, B., Geuens, M., & Baumgartner, H. (2013). The effect of familiarity with the response category labels on item response to Likert scales. *Journal of Consumer Research*, 40(2), 368–381.
- Weijters, B., Millet, K., & Cabooter, E. (2020). Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding. *International Journal of Research in Marketing*. <https://doi.org/10.1016/j.ijresmar.2020.04.002>.
- Wessling, K. S., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, 44(1), 211–230.
- Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: Effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Processes*, 57(2), 185–209.
- Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67(5), 555.

- Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, 30(1), 72–91.
- Yammarino, F. J., Skinner, S. J., & Childers, T. L. (1991). Understanding mail survey response behavior a meta-analysis. *Public Opinion Quarterly*, 55(4), 613–639.
- Yu, J., & Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36–44.
- Zettler, I., Lang, J. W., Hülshager, U. R., & Hilbig, B. E. (2015). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self-and observer reports. *Journal of Personality*, 84(4), 461–472.



Challenges in Conducting International Market Research

Andreas Engelen, Monika Engelen, and C. Samuel Craig

Contents

Introduction	122
Challenges in the Research Process	123
Conceptual Framework (Phase 1)	124
Research Units and Drivers of Differences (Phase 2)	127
International Data Collection (Phase 3)	133
Data Analysis (Phase 4)	138
Interpretation (Phase 5)	140
Summary	141
References	141

Abstract

This chapter explains the need to conduct international market research, identifies the main challenges researchers face when conducting marketing research in more than one country and provides approaches for addressing these challenges. The chapter examines the research process from the conceptual design of the research model to the choice of countries for data collection, the data collection process itself, and the data analysis and interpretation. Challenges identified include differentiating between etic and emic concepts, assembling an adequate research unit, ensuring data collection equivalence, and reducing ethnocentrism

A. Engelen (✉)

TU Dortmund University, Dortmund, Germany
e-mail: andreas.engelen@tu-dortmund.de

M. Engelen

TH Köln, Cologne University of Applied Science, Köln, Germany
e-mail: monika.engelen@th-koeln.de

C. Samuel Craig

New York University, Stern School of Business, New York, NY, USA
e-mail: scraig@stern.nyu.edu

of the research team. We draw on the extant literature to determine methods that address these challenges, such as an adapted etic or linked emic approach, to define the concept of the culti-unit, and to identify prominent approaches to cultural dimensions and collaborative and iterative translation and statistical methods for testing equivalence. This chapter provides researchers with the methods and tools necessary to derive meaningful and sound conclusions from research designed to guide international marketing activities.

Keywords

International research · Cross-cultural research · Emic/etic constructs · National indicators · National culture · Data equivalence · Culti-unit · Ethnocentrism · Back-translation

Introduction

Multinational companies are increasingly finding major opportunities for expansion outside their home markets. The transformation of planned economies into market economies and increasing demand from emerging middle classes in transition countries present new opportunities for firm growth, and rapid advances in technology facilitate access to these markets. As a consequence, many companies generate a large portion of their sales abroad. For example, the US giant Intel generated more than 80% of its overall sales in 2014 outside the US, BMW generated 81% of its sales outside Germany, and Sony generated 72% of its sales outside Japan. In the first quarter of 2012, Porsche sold more cars in China than in its German home market for the first time. Many start-up companies today are even “born global,” generating substantial sales outside their home nations from their founding or soon after (Knight and Cavusgil 2004).

These developments have important implications for marketing science. Practitioners expect advice about whether marketing knowledge and practices that are successful in their home markets (such as how to facilitate a firm’s market orientation, how consumers make purchasing decisions, and how promotion messages work) work in other nations (Katsikeas et al. 2006), as some highly successful companies (e.g., Disney with Disneyland Paris and Walmart in Germany) have experienced problems and even failures in expanding outside their home markets. These blunders have been traced back to failure to understand the new context and to adapt marketing activities to the host country, among other causes (Ghauri and Cateora 2010).

What, then, must marketing science do so it can provide useful recommendations? Single-country studies may be a first step, but they do not permit sound comparisons of phenomena between countries (Craig and Douglas 2005). Instead, multi-country studies that involve data collection in two or more nations are necessary to identify the generalizable similarities and differences in marketing-related insights between countries that marketing managers can use as guidelines for international marketing decisions.

Multi-country studies also contribute to marketing as an academic discipline. Steenkamp (2005) points out that marketing research has traditionally been driven by US researchers, so its constructs, theories, and relationships implicitly reflect the cultural predispositions of these researchers and the respondents in their empirical studies. Steenkamp claims that marketing science has to get out of the “US silo” and either show the cross-national generalizability of marketing phenomena or identify the contingencies that are related to national characteristics. National characteristics are valuable for marketing science since they allow constructs, theories, and relationships to be tested in diverse settings, similar to those that natural scientists create in their experiments (Burgess and Steenkamp 2006). Two nations, such as the US (a highly developed nation) and Cambodia (a developing nation), can provide extreme situations for testing constructs, theories, and relationships that other frequently used external contingency factors, such as a firm’s industry, cannot provide. If constructs, theories, or relationships hold in such diverse conditions as those offered in the US and Cambodia, a high level of generalizability can be assumed (Triandis 1994), and differences can be incorporated into theories and research models to make them more complete.

As a result, increasing numbers of multi-country studies have been published in leading marketing journals like *Journal of Marketing* (e.g., Petersen et al. 2015) and in journals that are dedicated to international marketing topics, such as *Journal of International Marketing* and *International Marketing Review*. While growing in number, multi-country marketing studies are often criticized for how they address the challenges that emerge when data is collected from more than one country (Cadogan 2010; Engelen and Brettel 2011; He et al. 2008; Nakata and Huang 2005). A multi-country research project has much in common with a project that focuses on one country (e.g., in terms of choice between primary and secondary data), but additional challenges render multinational studies more complex. From the conceptual design of the research model to the choice of countries for data collection, the actual data collection process, the data analysis and interpretation, pitfalls must be circumvented, and challenges faced in order to avoid limitations that can render an international marketing study all but useless. When differences in constructs, theories, or relationships between nations emerge, marketing researchers must ensure that they reflect the phenomena of interest and are not artifacts of poor research design and execution. The present article presents an overview of the challenges along the steps of the research process and state-of-the-art approaches to addressing these challenges so international marketing studies provide sound recommendations to practitioners and sound conclusions for marketing science.

Challenges in the Research Process

In order to capture comprehensively and in a structured way the particular challenges of international research projects in marketing, we break the typical research process into five steps, as depicted in Fig. 1: finding a conceptual framework (Phase 1), defining the research unit and identifying the unit’s drivers (Phase 2), conducting the

International research process steps	Major Challenge(s)	Possible approaches
1 Conceptual framework	Differences in how nations understand and apply constructs and frameworks	<ul style="list-style-type: none"> ▪ Conduct a literature review and qualitative fieldwork. ▪ Include researchers from all researched countries (e.g., via an adapted etic or linked emic approach). ▪ Establish construct equivalence (functional equivalence, conceptual equivalence, and category equivalence).
2 Research units and drivers of differences	National borders as units of research: <ul style="list-style-type: none"> ▪ Homogeneity within a country ▪ Cross-border influences Drivers that explain differences among nations	<ul style="list-style-type: none"> ▪ Analyze heterogeneity within nations and define the adequate culti-unit of research. ▪ Research subcultures in nations (e.g., China's nine subcultures). ▪ Choose an appropriate set of cultural dimensions, not only relying on Hofstede's dimensions but also taking Schwartz' and GLOBE's dimensions into account. ▪ Include formal and informal institutions at the national level.
3 International data collection	Equivalence of data across countries in terms of: <ul style="list-style-type: none"> ▪ measurement ▪ data collection 	<ul style="list-style-type: none"> ▪ Collect data in a sufficient number of nations. ▪ Consider alternative explanations for possible nation-level differences. ▪ Avoid mixed-worded, multi-item measures. ▪ Check conversions of measurement units. ▪ Apply collaborative and iterative translation techniques. ▪ Select equivalent samples and data-collection procedures. ▪ Ideally, combine direct-inference and indirect-inference techniques.
4 Data analysis		<ul style="list-style-type: none"> ▪ Correct for different response styles across nations. ▪ Conduct sound statistical analyses of equivalence. ▪ Consider multicollinearity issues in regression models.
5 Interpretation	Influence of the researcher's own culture (ethnocentrism)	<ul style="list-style-type: none"> ▪ Include researchers from all nations being researched.

Fig. 1 Summary of challenges along the international research process; own illustration

data collection in multiple nations (Phase 3), performing the data analyses (Phase 4), and interpreting the findings (Phase 5). The concept of data equivalence, in addition to other issues, assumes a major role in each of these steps, manifesting in terms of various facets of the concept (Hult et al. 2008). Data equivalence ensures that the differences between nations that are identified are actual differences in terms of the phenomena of interest and not artifacts that are due to conceptual and methodological shortcomings that ignore the particularities of multi-country studies.

Conceptual Framework (Phase 1)

A typical starting point of a multi-country study in marketing is the particular constructs, theories, or relationships to be tested. For example, a marketing researcher of Western origin might want to investigate whether the relationship between a firm's market orientation and the firm's performance, measured as the firm's profitability, holds across national contexts. Before collecting data for the research model, the researcher should determine whether the constructs and theories that link them are universal across nations, as the theory that guides the research might not be salient in all research contexts, and even if it is, the constructs might not hold the same meaning in one country as they do in another (Douglas and Craig 2006).

If this challenge is ignored at the beginning of a research process, implications drawn from findings in later phases can be misleading. In our example, while firm profitability might be the primary performance measure in the Western nations,

Asian cultures put more emphasis on group-oriented harmony than on achievement and consider employee satisfaction an important outcome of a market orientation, maybe even at the expense of some degree of profitability (Braun and Warner 2002). Leaving out this effect of a firm's market orientation would lead to an incomplete research model from the Asian perspective.

The degree to which research models, including their constructs, theories, and relationships, allow for country-specific adaptations is captured in the differentiation between an etic and emic approach (Berry 1989). These terms were coined by the linguistic anthropologist Pike (1967), who draws on an analogy with the terms "phonemic" (referring to language-specific sounds) and "phonetic" (referring to sounds that exist in multiple languages). An etic approach in cross-national research assumes that a research model from one national or cultural background can be applied and replicated in other nations and cultures, so it views the elements of the research model as universally valid. The emic view proposes that theories and constructs are specific to the context in which they are applied and are not universal. By using their own domestic situations as frames of reference that are then, without due reflection, applied to and tested in other nations, researchers often implicitly apply an "imposed-etic" approach. While this approach often leads more easily to comparable results ("comparing apples to apples"), the results might be influenced by a pseudo-etic perspective or bias (Triandis 1972) ("apples only relevant in country A, but not in country B") that leads to misguided or prejudiced findings. Schaffer and Riordan (2003) find that more than 80% of all published multi-country studies implicitly take such an "imposed-etics" approach.

Douglas and Craig (2006) propose two alternative approaches to designing the conceptual framework of a research model in a multi-country study: the adapted etic model and the linked emic model, whose differences are illustrated in Fig. 2. The adapted etic model starts with a conceptual model from a base culture and adapts it to other nations, while the linked emic model uses the various nations as a starting point and then incorporates the insights gained from the nations into one overall conceptual framework. Both approaches decenter the research perspective from the researchers' own national perspective by requiring extensive study of local literature on the topic, an international research team, and close consultation with researchers from the other nations.

The "adapted etic model" assumes that the conceptual framework applies to all nations, with some adaptations to local contexts. As a first step, the conceptual framework, its constructs, theories, and relationships are tested in terms of their applicability and relevance to other national contexts. For example, a market orientation may not be relevant in a planned economy. Next, the relevant constructs and hypotheses are checked with support from local researchers. For example, when a researcher is interested in determining whether a particular kind of corporate culture fosters a market orientation across nations (Deshpandé and Farley 2004), it may be necessary to ask local researchers to identify the values (as elements of a corporate culture) that are particularly relevant in their nations. This approach focuses on the similarities among nations, as even when modifications are made, it is likely that

		Adapted etic	Linked emic
Basic assumption		Theory and conceptual frameworks are pancultural (can be adapted to fit other nations)	Theory and conceptual framework are context-and culture-specific (cannot be fit directly into other cultures).
Process	Starting point	One base theory and framework and a uni-national/cultural research team	Several nation-specific frameworks and researchers from (ideally) each nation
	Definition of conceptual model	<ul style="list-style-type: none"> • Explicate the underlying theory and conceptual model in multiple settings. • Examine the relevant constructs and hypotheses and modify them to fit to other national settings. • Involve local researchers (e.g., develop different ways of operationalizing and measuring). 	<ul style="list-style-type: none"> • Agree on a common scope and key research questions. • Individual researchers work on each conceptual model. • Identify similarities and differences between conceptual models. • Define an overarching conceptual model that covers all identified elements and differentiate between emic and etic elements.
	Result	Adaptations of one conceptual framework to local conditions	Culture-specific definition of the conceptual framework that best addresses the key research questions
Emphasis		Similarities across cultures	Differences and the local perspective

Fig. 2 Adapted etic and linked emic research approaches; own illustration based on Douglas and Craig (2006)

the base nation’s perspective dominates, while the unique specifics of other nations may be ignored.

The “linked emic model” addresses this weakness by starting the process of defining a conceptual research model in multiple countries simultaneously, ideally with the support of a host country-based researcher for each setting. As a first step, the researchers from the various nations agree on the scope of the conceptual framework, which serves as input for the subsequent individual work on a conceptual model for each researcher in his or her national setting. Next, the researchers identify similarities among the locally developed models and factors at the national level that explain differences. Ideally, an overarching conceptual model is derived that covers all identified elements and differentiates between emic and etic elements. Nation-specific factors can be integrated into the model as contingencies to capture the nations’ emic particularities. Assuming a researcher is interested in understanding what drives a firm’s market orientation, collaboration among departments may be more important in collectivistic cultures than in individualistic cultures. This process puts a strong emphasis on the local perspective and is facilitated by effective cooperation among researchers from the nations in which the research takes place.

The efforts required in developing the adapted etic or linked emic model are targeted toward ensuring construct equivalence, a major facet of data collection equivalence that refers to whether constructs, theories, and relationships have the same purpose and meaning in all of the nations under investigation. Construct equivalence has three facets (Bensaou et al. 1999): functional equivalence,

conceptual equivalence, and category equivalence. Functional equivalence refers to the degree to which phenomena or objects have the same function across nations. For example, a car provides family transportation in a highly developed country while a motor bike performs the same function in an emerging economy. Conceptual equivalence relates to the degree to which the phenomena are interpreted similarly across nations. For example, in a study of the antecedents of market orientation, local interpretations of the “amount of market-related information transferred between departments” can be tested in exploratory fieldwork. Category equivalence captures the degree to which the same classification scheme can be applied to a phenomenon across nations. For example, a particular market orientation’s primary stakeholder group can be customers in one nation and government institutions in another.

Research Units and Drivers of Differences (Phase 2)

Once the conceptual framework has been established, the next step is to identify a research unit and the drivers of the differences between research units. We investigate the concept of the unit of research (section “[Definition of the Unit of Analysis](#)”) and discuss potential drivers of differences between units of research (section “[Identifying Drivers of Differences Between Nations](#)”).

Definition of the Unit of Analysis

In cross-national research, a unit of analysis must be established that defines the geographic scope and the group of the people or organizations to be examined within it. A good research unit has a high degree of homogeneity in terms of the members’ behaviors and values among the members of this group, which are heterogeneous to other groups and is as free of influence by other groups as possible (Craig and Douglas 2006; Lytle et al. 1995).

Literature reviews on international marketing research indicate (e.g., Engelen and Brettel 2011; Nakata and Huang 2005) that nations are the primary unit of research. Some of the reasons for a focus on nation as the unit are practical and pragmatic. Nations have clear and defined boundaries, and sampling frames are available at the nation level or for defined geographic areas in countries, such as regions and cities. In addition, multinational firms are often organized on a country basis and are interested in formulating strategies in multiple countries, for which they must assess the similarities and differences among countries. This focus carries through to academic researchers’ interest. For example, empirical cross-national marketing research often focuses on comparing phenomena of interest between the Western nations – often the US – and Asian countries – often China (Sheng et al. 2011). However, whether national borders are the best criterion with which to define the unit of research in international marketing may be questioned.

Nations and cultures are increasingly influenced by other nations and cultures. A primary mechanism is the global flows identified by Appadurai (1990). Five primary global flows blur the borders between nations: mediascapes (flow of images and communication, such as by screening US-made films in other countries),

ethnoscapes (flows of tourism, student exchanges, and migrants), ideoscapes (flows of political impacts and ideologies, such as democratization and views of equality), technoscapes (flows of technology and expertise), and finanscapes (flows of capital and money). These global flows, which will only grow because of the increasing ease and decreasing cost of data transfers and travel, lead to multicultural marketplaces (Demangeot et al. 2015), so the view of cultures as localized and determined only by national boundaries has lost much of its validity. These flows can cause changes in a country's cultural elements through cultural contamination (adopting foreign cultural elements), pluralism (individuals in one culture exhibiting features of multiple cultures), and hybridization (fusion of two cultural elements into one new) (Craig and Douglas 2006). For example, the US culture is presented to customers in most other countries via products that are seen as typical of the American lifestyle (e.g., McDonalds, Levis, Marlboro), and it exerts cultural influence via a globally dominant movie industry. Consequently, at least part of many countries' populations adopt these cultural elements and values, leading to the "Americanization" of other nations' cultures.

Further, many nations and their cultures are not homogeneous units but contain subgroups (Cheung and Chow 1999) that may be driven by migration, ethnic heritage (e.g., Chinese in Malaysia, the Dutch heritage of South African immigrants), religious beliefs (e.g., US Jews, Chinese Catholics), or the nation's size (e.g., Russia, China, India). For example, studies find cultural subgroups in Singapore (Chen 2008) and several South American nations (Lenartowicz and Johnson 2002). The frequent presence of such subgroups poses a challenge to international marketing research that seeks specificity in the differences between nations, as findings will depend on the subgroup sampled in such heterogeneous nations.

Given the influences that nations exert on each other and the heterogeneity in most nations, it follows that researchers must examine the homogeneity of the nations or regions they want to analyze. While some countries, such as Belgium, India, and Switzerland, are inherently heterogeneous in terms of behaviors, attitudes, and/or language, some studies provide concrete empirical evidence of this heterogeneity. Based on World Values Survey data, Minkov and Hofstede (2012) identify 299 in-country regions in terms of cultural values in twenty-eight countries that cluster on national borders, even for potentially heterogeneous countries like Malaysia and Indonesia. Intermixtures across borders are rare, even in cases of culturally similar African neighbors, such as Ghana and Burkina Faso. While this study validates empirically that nations can be good approximations for cultures, other studies that focus on a single nation find that there are substantial cultural differences within one nation. For example, Cheung and Chow (1999) use empirical research to find nine distinct subcultures in China, and Dheer et al. (2015) identify nine distinct subcultural regions in India and provide guidance on how these subcultures differ. (For example, the far-eastern and southwestern regions are lower in male dominance than the other parts of India.)

Given these diverse findings, it follows that researchers should consider whether using national borders is an appropriate way to define a unit of research. If there is no homogenous national culture or if there is doubt that homogeneity is present,

researchers should focus instead on subcultures, or “culi-units,” as the units of research (Douglas and Craig 1997). Naroll (1970) introduces the concept of culi-units as the relevant unit for studying cultural phenomena when homogeneous nations cannot be assumed, while a prominent definition is presented by Featherstone (1990; p. 385):

A culi-unit is [...] defined in terms of the racial, ethnic, demographic or social-economic characteristics or specific interests (e.g., ecologically concerned consumers) of its members which provide a common bond and establish a common ethnica, a core of shared memories, myths, values and symbols woven together and sustained in popular consciousness.

A commonly shared ethnica distinguishes the members of one culi-unit from others. This ethnica can be a national culture, a shared religion (e.g., Jewish heritage), or a strong interest (e.g., the hacker community of Anonymous). A major merit of the culi-unit concept is that it incorporates the concept of nation when the nation is sufficiently homogeneous, but it can also be applied to other ethnica. Researchers can benefit from the culi-unit construct since it makes ruling out alternative explanations for a theory or relationship easier than does a broader concept like nations. The ethnica core can be revealed by means of qualitative research. By taking the culi-unit as a starting point in defining the unit of research, researchers are forced to define their units of research unit cautiously and not to use national borders without careful reflection.

Sometimes the country or larger region is the appropriate sampling frame and serves as the culi-unit, particularly when culture has relatively little influence on the product or topic being researched. For example, compared to food and clothing, automobiles and consumer electronics do not have a strong cultural component. However, whenever there is likely to be considerable within-country heterogeneity or when the researcher is interested in understanding culture’s influence on a particular outcome, the researcher should either sample from the culi-unit or be able to identify the various cultural or ethnic groups and conduct analysis to determine their affect. For example, Petruzzellis and Craig (2016) examine the concept of Mediterranean identity across three European countries (Spain, France, and Italy) and find elements of an ethnica core related to Mediterranean identity that transcends national borders.

Research on subcultures within a larger culture illustrates the importance of a culi-unit. For example, Vida et al. (2007) conduct research in Bosnia and Herzegovina, where there are three major cultural/ethnic groups: Croats, Serbs, and Bosnians. The research analyzes responses by cultural group and finds that ethnic identity influences the dependent variables. Studies that examine a particular subculture face challenges in obtaining a sampling frame, but they can view a homogeneous group of respondents. Within-country differences can also be examined geographically. Lenartowicz et al. (2003) find significant within-country and between-country differences among managers on the Rokeach Value Survey, suggesting that using the country as the unit of analysis would mask important within-country variations.

Ultimately, the selection of the unit of analysis will be a function of practical considerations and the research's theoretical underpinnings. If a particular theory is being tested in multiple countries, the respondents in each country must reflect the ethnic core of the culture of interest. Ideally, the research would be able to locate appropriate sampling frames to focus on the specific groups, but if such sampling frames are not available, questions should be included that allow for a fine-grained examination so the entire sample can be analyzed and then broken down by specific cultural/ethnic groups. If there are significant differences between groups, the one (s) most relevant to the research can be examined more closely. A related concern is determining what factors account for the observed differences, whether they are contextual factors like the level of economic development or the influence of other cultures. The use of covariates in the analysis will often help in identifying which factors affect the culture.

Identifying Drivers of Differences Between Nations

When several nations are compared,¹ one of the key questions that arises concerns the underlying drivers between nations that account for the difference and that may even be generalized to explain variations from other nations. Assuming that nations are appropriate units of research for a particular purpose and we find differences (e.g., in the strength of the relationship between market orientation and firm performance between the US and Indonesia). An explanation for these differences can lie in the differing degrees of cultural individualism versus collectivism (Hofstede 2001), but the US and Indonesia also have differences in their economic (e.g., GPD per capita) and development levels (e.g., Human Development Indicator or HDI), which may be the key drivers of the observed differences.

In their review of empirical cross-national and cross-cultural research, Tsui et al. (2007) find that national culture – typically defined as the values and norms that guide a group's behavior (Adler 2002) – is the most frequently investigated driver of differences between nations. National culture can be conceptualized along national cultural dimensions that relate to how societies resolve the problems that all societies face (e.g., whether the individual person is more important than group equality and harmony and how much privacy is granted to individuals). Various schemes of cultural dimensions have been proposed, but the four original dimensions from Hofstede (2001) – power distance, individualism versus collectivism, uncertainty avoidance, and masculinity versus femininity – are the most prominent. Later, Hofstede and colleagues added the dimensions of long-term orientation and indulgence versus restraint. The latter, originally proposed by Minkov (2007), has been identified by means of World Value Survey items (Hofstede et al. 2010). Societies that are strong on indulgence allow free gratification of natural human desires, while

¹For the sake of simplicity, we will subsequently refer to nations as the unit of research, acknowledging that other culti-units may be more appropriate as outlined in section “Conceptual Framework (Phase 1).”

societies that are strong on restraint prefer strict norms that regulate such gratification.

International marketing research focuses on Hofstede's dimensions, as the literature review from Engelen and Brettel (2011) indicates. One might argue that the country scores that Hofstede initially developed at the end of the 1960s/beginning of the 1970s are outdated, but Beugelsdijk et al. (2015) show that cultural change is absolute, rather than relative. By replicating Hofstede's dimensions for two birth cohorts using data from the World Values Survey, Beugelsdijk et al. (2015) find that most countries today score higher on individualism and indulgence and lower on power distance compared to Hofstede's older data, but cultural differences between country pairs are generally stable. Further, to circumvent the threat of using outdated country data, international marketing researchers can apply the updates provided on Hofstede's website (<http://geert-hofstede.com/>). These updated data have been used in some recent cross-national marketing studies, such as Samaha et al. (2014). Other studies, such as the meta-analytical review from Taras et al. (2012), also provide updated country scores for Hofstede's dimensions.

Several authors criticize Hofstede's approach in terms of its theoretical foundation and the limited number of cultural dimensions (Sondergaard 1994). Schwartz (1994) and the GLOBE study address some of these criticisms. Siew Imm et al. (2007) find that the cultural dimensions from Schwartz (1994) are broader than those from Hofstede (2001), as Schwartz (1994) covers all of Hofstede's dimensions and adds the dimensions of egalitarianism and hierarchy. Steenkamp (2001) also highlights Schwartz' (1994) theoretical foundations, concluding that "given its strong theoretical foundations, [Schwartz's approach] offers great potential for international marketing research" (p. 33).

Javidan et al. (2006) point out that the GLOBE study adopts a theory-based procedure and formulate a priori dimensions based on Hofstede (2001) dimensions, values that Kluckhohn (1951) and McClelland (1961) described, and the interpersonal communication literature (Sarros and Woodman 1993). In addition to Hofstede's cultural dimensions of power distance and uncertainty avoidance, the GLOBE study adds performance orientation, assertiveness, future orientation, human orientation, institutional collectivism, in-group collectivism, and gender egalitarianism (House et al. 2001). Some of these novel dimensions are more fine-grained than are Hofstede's (2001) dimensions. For example, the dimensions of assertiveness and gender egalitarianism reflect two major facets of Hofstede's masculinity dimension (Hartog 2004). Cross-cultural marketing studies often neglect or even ignore the potential offered by Schwartz (1994) and the GLOBE study. International marketing researchers should be sure to justify their choices of national cultural dimensions as the most appropriate for their purposes.

A marketing researcher who needs to choose one approach should consider the following thoughts: Hofstede's and GLOBE's dimensions and country scores have been derived theoretically and/or empirically in the workplace setting, so organizational marketing topics might rather build on their dimensions. Schwartz' dimensions have their theoretical origin in psychological research on individual values and have been empirically analyzed by Schwartz in a cross-national sample of teachers

and students. Therefore, these dimensions are rather appropriate when investigating the decisions of private persons across cultures (such as in international consumer studies).

Further, the targeted nations in an international marketing study can lead to the use of the one or other approach. Schwartz generated data in some regions which have not been covered to the same extent in Hofstede's and the GLOBE survey (e.g., some former Eastern European bloc countries and some countries in the Middle East). There are also some countries (e.g., some African countries) which have been covered by GLOBE and not the other approaches. So, the individually targeted countries in a research project may determine the choice of dimensions.

In addition, researchers should take into consideration that the cultural dimensions differ between the approaches. Steenkamp (2001) factor analyzes the dimensions from Hofstede (the original four dimensions) and Schwartz and identifies four factors – three related to both Hofstede's and Schwartz's dimensions and one, a factor related to egalitarianism versus hierarchy that refers to how people coordinate with other people and to what degree they take the other people's interests into account, that emerged in the Schwartz data. Steenkamp (2001) argues that, when a researcher investigates cross-nationally whether the consumption of products that could harm other nonusers is accepted (e.g., cigarettes), this factor is represented in Schwartz's dimensions, not in Hofstede's dimensions, and is highly relevant. Therefore, Schwartz's dimensions might be the best choice. The GLOBE dimensions are also broader than Hofstede's dimensions, breaking down Hofstede's dimension of masculinity versus femininity into gender egalitarianism and assertiveness and differentiating between two versions of Hofstede's individualism versus collectivism dimension (in-group and institutional collectivism), which enables more fine-grained analysis on this dimension. Building on the GLOBE scores, Waldman et al. (2006) differentiate between in-group and institutional collectivism and find that institutional collectivism is positively related to corporate social responsibility in a firm's decision-making, while in-group collectivism has no impact. Using one score for a broader collectivism dimension may have masked these cultural dependencies, so depending on what a marketing researcher wants to examine, the more fine-grained GLOBE dimensions might be more appropriate. Figure 3 provides a summary of the three approaches to cultural dimensions.

In their literature review on cross-national and cross-cultural research, Tsui et al. (2007) conclude that extant research has focused too much on national cultural dimensions while neglecting other drivers of the differences between nations. As a result, the findings of multinational studies that focus only on national cultural dimensions may be misleading. Tsui et al. (2007) and researchers like Glinow et al. (2004) call for a *polycontextualization* of international research in order to accommodate the complexity of the context and avoid misleading conclusions about what drives the differences between nations. Beyond national culture, the physical context (e.g., climate, typology), the historic context (e.g., sovereignty, colonization), the political context (e.g., the political and legal systems), the social context (e.g., religion, family structure), and the economic context (e.g., economic system, technology) may be the reason for differences (Saeed et al. 2014). Sound

	Hofstede (2001, 2010)	Schwartz (1994)	House et al. (2001); GLOBE study
Origin	Deduced from a large-scale survey among IBM employees in 40 countries (1967–1973); extended to wider population in the following decades	Main data collection from 1988 to 1992, primarily among teachers and students in 41 cultures (covering 38 nations)	Survey of more than 17,000 managers in 62 countries on prevalent values and practices from the mid-1990s on
Dimensions	<ul style="list-style-type: none"> • Power Distance • Uncertainty Avoidance • Individualism vs. Collectivism • Femininity vs. Masculinity • Long-Term-Oriented • Indulgence vs. Restraint 	<ul style="list-style-type: none"> • Embeddedness • Intellectual Autonomy • Affective Autonomy • Egalitarianism • Harmony • Hierarchy • Mastery 	<ul style="list-style-type: none"> • Power Distance • Uncertainty Avoidance • In-Group Collectivism • Institutional Collectivism • Performance Orientation • Assertiveness • Gender Egalitarianism • Humane Orientation • Future Orientation
Country scores	<ul style="list-style-type: none"> • Scores from 0 to 120 for 93 nations 	<ul style="list-style-type: none"> • Scores for 76 cultures in 74 nations 	<ul style="list-style-type: none"> • Scores from 1 to 7 for values and practices for 62 nations
Criticism	<ul style="list-style-type: none"> • Originally only IBM employees; later extended, but not representative of the general population • Lack of theoretical foundation 	<ul style="list-style-type: none"> • Not representative of the general population; 85% of respondents are teachers and students 	<ul style="list-style-type: none"> • Not representative of the general population, only managers • Differences of scores on values and practices in some countries are questionable
	<ul style="list-style-type: none"> • Nations as cultural boundaries 	<ul style="list-style-type: none"> • Ethnocentrism of researcher 	<ul style="list-style-type: none"> • Limited number of cultural dimensions

Fig. 3 Comparison of prominent approaches to cultural dimensions; own illustration based on Hofstede (2001), Schwartz (1994), and House et al. (2001)

international marketing research must not neglect these contextual drivers (see Douglas and Craig (2011) for a discussion of the role of contextual factors).

International Data Collection (Phase 3)

After the conceptual framework and the research unit are defined, data collection can begin. A key decision for the researcher is to decide in which and how many nations to collect empirical data. The key challenge is to take steps to ensure that the data are comparable and equivalent across all countries. This is a critical step as sound data provide the foundation for inferences and interpretation. The three pillars that guide data collection relate to the constructs that underlie the research, the actual measurement of the constructs and other variables of interest, and the procedures used to collect the data across multiple countries. These steps are summarized in Fig. 4. In addition, steps need to be taken to ensure translation equivalence, sampling frame equivalence, and data collection procedure equivalence (Hult et al. 2008).

Extant international marketing research is often built on data from only two nations (Cadogan 2010; Engelen and Brettel 2011). However, this approach has serious limitations, particularly since countries typically differ in terms of more than one cultural dimension, as well as in such contextual areas as the macroeconomic development stage or the educational system (Geyskens et al. 2006). As a positive example, Steenkamp et al. (1999) draw on responses from more than 8000

	Type of equivalence	Possible approaches
Con-structs	Functional	Literature review, qualitative fieldwork, and adapted etic or linked emic approach
	Conceptual	
	Category	
Measurement	Translation	Collaborative and iterative translation techniques
	Configural, metric, and scalar	Multiple group confirmatory factor analysis
Data collection	Sampling frame	Selection of equivalent samples across nations
	Data collection procedure	Similar data-collection time and procedures, allowing for different levels of literacy and infrastructure
	Sampling comparability	Statistical control with socio-demographic variables as covariates

Fig. 4 Overview of types of data collection equivalence; own illustration based on Hult et al. (2008)

consumers in 23 countries to isolate the effects of the regulatory system, the moral system (national identity), and the cultural system (degree of individualism) on the perceived value of websites. These effects could not have been separated on the national level with a two-country comparison.

To address what is a meaningful number of nations in which data should be collected and how these nations should be chosen, Sivakumar and Nakata (2001) develop an approach to guide researchers in defining the number of nations for data collection. In their approach, when cultural differences are expected to be due to one cultural dimension (e.g., the degree of power distance), two nations that have strong differences in terms of this dimension and few differences in the other cultural dimensions should be chosen, and when differences are expected to be due to two cultural dimensions, four national cultures should be used to represent all four combinations of the two levels (high and low) for each cultural dimension, while the four national cultures are similar in terms of the remaining cultural dimensions.

While this approach can help researchers determine the appropriate number of nations for identifying the role of cultural dimensions, the procedure does not provide guidance on how to deal with rival and confounding drivers at the national level, such as the stage of macroeconomic development (Ralston et al. 1997). In order to exclude rival explanations for differences between nations, even more nations should be included. For example, Tan (2002) creates a hybrid, quasi-experimental design to determine whether national cultural or contextual effects prevail by drawing on three samples from two subcultures and two countries: mainland Chinese, Chinese Americans, and Caucasian Americans.

In order to identify the roles that national cultural or contextual factors at the national level play, the number of nations for data collection must be extended, as long as identified differences can be traced back to either one of the national cultural or contextual factors, while controlling for alternative explanations at the national level.

Once the national settings are defined but before the data collection begins, three equivalence challenges must be addressed in order to generate sound empirical findings: translation equivalence, sampling frame equivalence, and data collection procedure equivalence (Hult et al. 2008). Collecting data in several countries in which more than one language is involved requires ensuring translation equivalence. Simple back-translation is the dominant approach in international marketing studies, where a questionnaire in the researcher's native language is translated into another language by a bilingual person (Brislin 1980). This translated questionnaire is then back-translated to English by another bilingual person. Only when the researcher compares the original and the back-translated questionnaire and finds no relevant differences can translation equivalence be assumed. While this approach is the most widely applied in international marketing literature, it has some limitations, as it does not necessarily ensure equivalence in meaning in each language (Douglas and Craig 2007). Referring to the "emic versus etic" debate, assuming that a simple translation from the base language that does not take the particularities of the other language into account (e.g., words or idioms that exist in only one language) is inherently etic or even "imposed-etic."

Douglas and Craig (2007) propose a collaborative, iterative approach that finds meanings of the source language in the other languages, thereby integrating emic elements into the questionnaires. Given the complexity of languages, the authors hold that researchers and translators with linguistic skills and skills in questionnaire design collaborate for this purpose. This approach has five major steps, as Fig. 5 shows.

The process starts with the translation, where a questionnaire in one language is translated independently to all target languages by at least two translators. Translators should especially pay attention to items that deal with attitudes since linguistic research indicates that the connotations of words like "happiness" and "mourning" can differ from language to language. The translation of mix-worded multi-item measures – that is, measures that contain positive-worded statements and reverse-worded statements – is a major challenge since empirical studies have found problems with the internal consistency and dimensionality of these measures, which are mostly of US origin, when applied cross-nationally. Wong et al. (2003) identify two reasons for these problems: how languages indicate negation differ such that reverse-worded statements may be difficult or even impossible to translate appropriately, and respondents' cultural predeterminations affect how they respond to reverse-worded statements. When the dominant norm is to be polite and agreeable, as is the case in some Asian cultures (Child and Warner 2003), respondents may tend to agree with any statement, leading to low internal consistency and disruption of the dimensionality of mixed-worded measures. Therefore, Wong et al. (2003) suggest employing only positively worded statements cross-nationally or replacing Likert statements with questions.

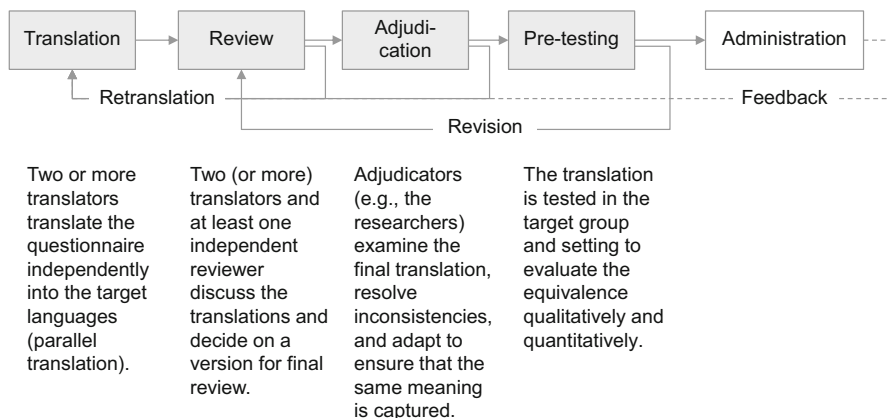


Fig. 5 Collaborative and iterative translation; own illustration based on Douglas and Craig (2007)

The second step of Douglas and Craig's (2007) approach is a review meeting with the translators and an independent researcher in order to agree on a preliminary version of the questionnaire in each language. In the third step, adjudication, inconsistencies are resolved and whether the questionnaire actually measures the same meaning in each country is determined. In the fourth step, the questionnaire is pretested with native respondents in each language to ensure comprehension, and issues are referred to the team of researchers and translators to start a new round of iterations. Finally, in the fifth step, if more than one round of data collection is planned, such as may be the case with longitudinal data collected yearly, insights gained during the initial data collection are reported to the team of researchers and translators so they can improve the questionnaires for successive rounds of data collection (without compromising year-to-year comparability).

Sampling frame equivalence refers to the extent to which the samples drawn from the various nations parallel one another (Hult et al. 2008). For example, if high school students from the middle class in India are compared to public high school students from all social classes in the US, the discrepancy in social class could distort the findings. Researchers must select equivalent samples among the various research units while allowing variations in the sample on the factors to be analyzed. For example, individuals of similar income, education level, and gender (equality of sample) are selected from nations whose cultural values (variation to be researched) differ. Hult et al. (2008) consider sample equivalence a major prerequisite for sound cross-cultural comparisons and recommend that organization-level studies match samples in terms of potentially confounding factors like company age, size, and industry sector. Although sample equivalence does not guarantee that findings can be generalized to the participating countries, it helps to ensure that comparisons are not confounded by other factors and that differences can be traced back to the cultural dimensions or contexts under study (van Vijver and Leung 1997).

Data collection procedure equivalence combines administrative equivalence (e.g., telephone, face-to-face, email) and time equivalence in terms of the time between data collection in the participating countries (Hult et al. 2008). While a completely standardized and parallel data collection procedure in all units of research is ideal, regulations (e.g., national rules and regulations against telephone interviews), cultural norms (e.g., nonconformity of impersonal surveys with the cultural preferences related to personal interactions), and infrastructure (e.g., availability of high-speed internet lines) often prevent data collection procedures from being perfectly equal (Hult et al. 2008). Even so, researchers must seek equivalent data collection procedures and keep unavoidable differences (e.g., different survey settings and different times) in mind as a possible explanation for findings.

Finally, in the international data collection phase, researchers must decide on which level to measure cultural properties. Cultural properties can either be directly measured in the surveyed sample at the individual level of each respondent (“direct value inference”) or captured by means of typically nation-related secondary data (e.g., the Hofstede country scores) according to the surveyed individuals’ or firms’ national identity (“indirect value inference”). Direct value inference measures individuals’ cultural properties and derives cultural properties for data analyses by aggregating these cultural properties to the relevant group level (e.g., the national level). Researchers can directly measure the surveyed individuals’ individual values or ask about their perceptions of their environment’s cultural values. This approach ensures that the actual culture of the surveyed individuals is measured. However, the questions concerning whether the surveyed individual can assess the cultural properties correctly and whether there are any bias in the assessments remain. Indirect value inference assigns existing values on cultural dimensions (e.g., from the Hofstede data) to surveyed individuals according to their group membership – that is, surveyed individuals from Germany receive the score for Germany on the relevant cultural dimension as reported by Hofstede or other researchers. In this case, however, a measurement error might occur since the researcher assumes that the surveyed sample’s cultural properties comply with the cultural properties of the samples used in the earlier studies that report country scores (Soares et al. 2007). Given the benefits and perils of both approaches, Soares et al. (2007) recommend a multi-method approach that combines the indirect and direct value inference approaches.

Whether direct or indirect value inference is pursued, Brewer and Venaik (2014) recommend that researchers that are determining the right level of analysis ensure a fit between the levels at which the constructs (e.g., cultural properties) are theorized and empirically validated. Conceptualization of theories, measurement of constructs, and data collection should be conducted consistent with the underlying research question. Brewer and Venaik (2014) refer to the danger of an ecological fallacy when researchers assume that group-level relationships automatically apply to the individual level. Some recent studies use cultural dimensions explicitly at the individual level (e.g., individual power orientation) to make clear that cultural properties at the level of the surveyed individual are theorized and measured (e.g., Auh et al. 2015; Kirkman et al. 2009). When only group-level data on cultural properties is available,

Brewer and Venaik (2014) recommend that higher-level constructs (e.g., cultural dimensions at the national level) be cross validated with measures at the level at which the construct is theorized (e.g., at the individual level of a consumer).

Data Analysis (Phase 4)

Data analysis starts with analyses which are not specific to international marketing research but have to be done in any data analysis. Such checks and tests include establishing the reliability and validity of the measures and ruling out biases in the survey data like common-method bias, nonresponse bias, and informant bias (Bagozzi et al. 1991; Podsakoff et al. 2003). However, the particularities of international marketing research impose additional challenges related to the data analysis based on how differences in national response styles can affect findings. A response style refers to a person's tendency to respond systematically to questionnaire items on some basis other than what the items are designed to measure (Jong et al. 2008). For example, in the US school grading systems range from A+ (best) to F (worst) or points up to 100 (best) and down to 0 (worst), while the German grading system is the other way around, ranging from 1 (best) to 5 or 6 (worst). Therefore, Germans who are used to "lower is better" might unwittingly answer incorrectly on US surveys that range from 1 (worst) to 5 (best). While these scale definition issues might be resolved easily, national cultural predeterminations based on deeply rooted values and preferences may have more subtle effects on a participant's response style (Clarke III 2001).

Two major response styles have been shown to be subject to the respondent's national culture: An extreme response style (ERS) is the tendency to favor the end points of rating scales, regardless of the item's content, while an acquiescence response style (ARS) refers to the tendency to agree with all items, regardless of the item's content. Chen (2008) reports that US respondents are much more inclined to show an ERS than are respondents from China. In cultural terms, respondents from low-ERS cultures may wish to appear modest and nonjudgmental, whereas members of high-ERS cultures may prefer to demonstrate sincerity and conviction. Regarding ARS, Riordan and Vandenberg (1994) report that a response of 3 on a 5-point Likert-type scale means "no opinion" to American respondents but "mild agreement" to Korean respondents, so a Korean's "3" may be equivalent to an American's "4," and a Korean's "4" may be equivalent to an American's "5." A strong response bias is problematic because whether differences are caused by differences in response styles or differences in the factor of interest remains uncertain. If relationships are compared, differences in response styles lead to variances in the dependent and independent variables that result in unintended and confounding differences in correlations.

Differences in response styles belong to a larger group of issues when measurement models are applied in more than one nation. A major threat to sound multi-national marketing research occurs when respondents do not interpret the constructs that link relevant relationships or that build theoretical frameworks similarly such

that identified differences in relationships are actually due to systematically different interpretations of constructs and measurement models (Mullen 1995). While some countermeasures can be taken in the pre-data collection phase, such as ascertaining translation equivalence, quantitative tests on the actual data collected are necessary.

In particular, measurement equivalence, which relates to whether measurement models are understood similarly across nations, must be established. Steenkamp and Baumgartner (1998) provide a multigroup confirmatory factor analysis approach for reflective multi-item measurement models, which approach consists of configural, metric, and scalar equivalence analyses (for applications, see, e.g., Homburg et al. (2009) and Zhou et al. (2002)). Configural equivalence indicates that respondents from all nations under analysis conceptualize and understand the basic structure of measurement models similarly. Metric equivalence indicates that groups of respondents understand scale intervals similarly. Scalar equivalence indicates that the systematic response style among respondents from the nations under study does not differ.

Configural equivalence, which is tested by running a multigroup confirmatory factor analysis that allows all factor loadings to be free across the national samples, is given when the factor loadings are significant in all samples and the model's fit is satisfactory. Partial metric equivalence is given when at least one item (in addition to a marker item) for each measurement model has equivalent factor loadings across nations. Metric equivalence models must be specified with at least two factor loadings per measurement model that are kept equal across nations while not constraining the remaining factor loadings. Full metric equivalence is given when all factor loadings are equal across groups, although Steenkamp and Baumgartner (1998) indicate that partial metric equivalence is sufficient in most cases. By means of a χ^2 -difference test, this metric equivalence model is compared with a model in which all factor loadings are free across samples, and metric equivalence is confirmed when the two models do not differ significantly. Finally, scalar equivalence, which is tested by comparing means, ensures that differences in observed and latent means between national samples are comparable. The procedure for testing scalar equivalence is the same as the χ^2 -difference test for metric equivalence except that item intercepts are constrained across national samples. Steenkamp and Baumgartner (1998) point out that, in most cross-national comparisons, only partial scalar invariance is realistic.

Since establishing measurement equivalence across a high number of countries would require extremely large sample sizes, some studies have created set of countries with similar cultural and economic conditions between which measurement equivalence is established (e.g., Hohenberg and Homburg 2016; Tellis et al. 2009). For example, Hohenberg and Homburg (2016) cluster their 38 surveyed countries into four categories, differentiating among English-speaking countries, European countries, Asian countries, and Latin American countries.

Once measurement equivalence is established, the relationships of interest can be empirically investigated. When national constructs are integrated as moderators in theoretical frameworks, group comparisons (e.g., in structural equation modeling) and interaction term models (e.g., in regression models) can be applied (Engelen and

Brettel 2011), although some challenges specific to international marketing projects must be considered. Most multinational studies investigate a particular relationship (e.g., the effect of top management's attention on a firm's market orientation) in multiple nations, for which the researchers have a national dependency in mind, such as a particular national cultural dimension (e.g., the degree of national cultural power distance). However, to accommodate the multiplicity of possible drivers at the national level (section "[Identifying Drivers of Differences Between Nations](#)"), researchers should add controls for alternative explanations in their models, an approach that is particularly feasible in regression models (Becker 2005). For example, by integrating a broad set of national cultural dimensions into their model, Samaha et al. (2014) show that national culture has a more multifaceted role in relationship marketing than earlier studies that focus on just one national cultural dimension suggest. Adding several cultural dimensions into a regression model is likely to lead to multicollinearity since the cultural dimensions are often correlated. Individualism versus collectivism and power distance are often strongly correlated. Samaha et al. (2014) circumvent this problem by not adding these two dimensions simultaneously in their regression models, leaving out the individualism versus collectivism dimension in their power distance model and leaving power distance out of all other models.

Interpretation (Phase 5)

While interpretation is an important element in all five steps of the research process – before, during, and after data collection – it manifests particularly at the end of the research process, when the actual findings are available. Of course, interpretation is by no means a particularity of international marketing research projects, but one particular challenge emerges with these kinds of studies. A major assumption of cross-national research projects is that drivers at the national level can lead to differences in marketing relationships, constructs, and theories, and since national drivers, especially national culture, affect everyone living in a nation or culture (Hofstede 2001), the researcher himself or herself is also subject to national or cultural predetermination. Thus, the researcher's cultural values can affect his or her interpretation of the findings (Berry 1980). This bias, called ethnocentrism, occurs when one person's or group's frame of reference is applied in interpreting other groups' responses without adaptation to other national cultures. If, in our example, we find that the attention of top management to market-related issues drives a firm's market orientation more strongly in Asia than in Western nations, coming from a Western perspective, we could easily assume that power distance, which is particularly strong in Asian cultures, is the driving force. However, Asian researchers might relate this finding to particularities of the Confucian teachings. To exclude such an ethnocentric bias, researchers in international studies should build and use cross-national research teams during the entire research process, but especially in the last step of interpreting the findings (Hofstede and Bond 1988), and document nation- or culture-specific interpretations of findings.

Summary

As firms from developed and developing economies continue to expand outside their home markets, marketing research is essential to guide development and execution of marketing strategy. An implicit challenge is for management to appreciate that there are potentially a wide range of differences between their home market and the foreign markets they currently operate in or are planning to enter. Well-constructed research will not only identify differences but also reveal important similarities. Regardless of the specific purpose of the research, it is essential that valid and reliable research be designed and executed. This is the critical challenge and applies whether the research is to guide management decisions or test the applicability of theories and constructs across multiple countries.

However, international marketing research is more complex and time consuming than single country research. Advances in technology, particularly ready access to internet samples in multiple countries has greatly facilitated rapid collection of multi-country data. However, unless careful attention is paid to the design and execution of the research to achieve equivalence on all dimensions across the units studies, the results may be misleading or meaningless. Careful attention to all the steps outlined in this chapter is essential to ensure that the results of international marketing research are reliable and valid and can be used to make meaningful inferences and advance the state of our knowledge about markets outside our own.

References

- Adler, N. (2002). *International dimensions of organizational behavior*. Cincinnati: South-Western College Publishing.
- Appadurai, A. (1990). Disjuncture and difference in the global cultural economy. *Public Culture*, 2, 1–24.
- Auh, S., Menguc, B., Spyropoulou, S., Wang, F. (2015). Service employee burnout and engagement: The moderating role of power distance orientation. *Journal of the Academy of Marketing Science*, 1–20. <https://doi.org/10.1007/s11747-015-0463-4>.
- Bagozzi, R., Yi, Y., & Phillips, L. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly*, 36(3), 421–458.
- Becker, T. E. (2005). Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, 8(3), 274–289.
- Bensaou, M., Coyne, M., & Venkatraman, N. (1999). Testing metric equivalence in cross-national strategy research: An empirical test across the. *Strategic Management Journal*, 20(7), 671–689.
- Berry, J. (1980). Introduction to methodology. In H. Triandis & J. Berry (Eds.), *Handbook of cross-cultural psychology* (pp. 1–28). Boston: Allyn & Bacon.
- Berry, J. (1989). Imposed etics-emics-derived etics: The operationalization of a compelling idea. *International Journal of Psychology*, 24(6), 721–734.
- Beugelsdijk, S., Maseland, R., & van Hoon, A. (2015). Are scores on Hofstede's dimensions of national culture stable over time? A cohort analysis. *Global Strategy Journal*, 5(3), 223–240.
- Braun, W., & Warner, M. (2002). The “Culture-Free” versus “Culture-Specific” management debate. In M. Warner & P. Joynt (Eds.), *Managing across cultures: Issues and perspectives* (pp. 13–25). London: Thomson Learning.

- Brewer, P., & Venaik, S. (2014). The ecological fallacy in national culture research. *Organization Studies*, 35(7), 1063–1086.
- Brislin, R. (1980). Translation and content analysis of oral and written materials. In H. Triandis & J. Berry (Eds.), *Handbook of cross-cultural psychology* (pp. 389–444). Boston: Allyn and Bacon.
- Burgess, S., & Steenkamp, J.-B. (2006). Marketing renaissance: How research in emerging markets advances marketing science and practice. *International Journal of Research in Marketing*, 23(4), 337–356.
- Cadogan, J. (2010). Comparative, cross-cultural, and cross-national research: A comment on good and bad practice. *International Marketing Review*, 27(6), 601–605.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5), 1005–1018.
- Cheung, G. W., & Chow, I. H.-S. (1999). Subcultures in Greater China: A comparison of managerial values in the People's Republic of China. *Asia Pacific Journal of Management*, 16(3), 369–387.
- Child, J., & Warner, M. (2003). Culture and management in China. In M. Warner (Ed.), *Culture and management in Asia* (pp. 24–47). London: Routledge.
- Clarke III, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review*, 18(3), 301–324.
- Craig, C. S., & Douglas, S. P. (2005). *International marketing research* (3rd ed.). Chichester: Wiley.
- Craig, C. S., & Douglas, S. P. (2006). Beyond national culture: Implications of cultural dynamics for consumer research. *International Marketing Review*, 23(3), 322–342.
- Demangeot, C., Broderick, A., & Craig, C. S. (2015). Multicultural marketplaces. *International Marketing Review*, 32(2), 118–140.
- Deshpandé, R., & Farley, J. (2004). Organizational culture, market orientation, innovativeness, and firm performance: An international research odyssey. *International Journal of Research in Marketing*, 21(1), 3–22.
- Dheer, R. J. S., Lenartowicz, T., & Peterson, M. F. (2015). Mapping India's regional subcultures: Implications for international management. *Journal of International Business Studies*, 46(4), 443–467.
- Douglas, S. P., & Craig, C. S. (1997). The changing dynamic of consumer behavior: Implications for cross-cultural research. *International Journal of Research in Marketing*, 14(4), 379–395.
- Douglas, S., & Craig, C. (2006). On improving the conceptual foundations of international marketing research. *Journal of International Marketing*, 14(1), 1–22.
- Douglas, S. P., & Craig, C. S. (2007). Collaborative and iterative translation: An alternative approach to back translation. *Journal of International Marketing*, 15(1), 30–43.
- Douglas, S. P., & Craig, C. S. (2011). The role of context in assessing international marketing opportunities. *International Marketing Review*, 28, 150–162.
- Engelen, A., & Brettel, M. (2011). Assessing cross-cultural marketing theory and research. *Journal of Business Research*, 64(5), 516–523.
- Featherstone, M. (1990). *Global culture: Nationalism, globalism and modernism*. London: Sage.
- Geyskens, I., Steenkamp, J., & Kumar, N. (2006). Make, buy, or ally: A transaction cost theory meta-analysis. *Academy of Management Journal*, 49(3), 519–543.
- Ghauri, P., & Cateora, P. (2010). *International marketing* (3rd ed.). New York: McGraw-Hill.
- von Glinow, M. A., Shapiro, D. L., & Brett, J. M. (2004). Can we talk, and should we? Managing emotional conflict in multicultural teams. *Academy of Management Review*, 29(4), 578–592.
- Hartog, D. (2004). Assertiveness. In R. House, P. Hanges, M. Javidan, P. Dorfman, & V. Gupta (Eds.), *Culture, leadership, and organizations: The GLOBE study of 62 societies* (pp. 395–436). Thousand Oaks: Sage.
- He, Y., Merz, M. A., & Alden, D. L. (2008). Diffusion of measurement invariance assessment in cross-national empirical marketing research: perspectives from the literature and a survey of researchers. *Journal of International Marketing*, 16(2), 64–83.

- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks: Sage.
- Hofstede, G., & Bond, M. (1988). The confucius connection: From cultural roots to economic growth. *Organizational Dynamics*, 16(4), 5–21.
- Hofstede, G., Hofstede, J., & Minkov, M. (2010). *Cultures and organizations – software of the mind: Intercultural cooperation and its importance for survival*. New York: McGraw-Hill.
- Hohenberg, S., & Homburg, C. (2016). Motivating sales reps for innovation selling in different cultures. *Journal of Marketing*, 80(2), 101–120.
- Homburg, C., Cannon, J. P., Krohmer, H., & Kiedaisch, I. (2009). Governance of international business relationships: A cross-cultural study on alternative governance modes. *Journal of International Marketing*, 17(3), 1–20.
- House, R., Javidan, M., & Dorfman, P. (2001). Project GLOBE: An introduction. *Applied Psychology: An International Review*, 50(4), 489–505.
- Hult, T., Ketchen, D., Griffith, D., Finnegan, C., Gonzalez-Padron, T., Harmancioglu, N., et al. (2008). Data equivalence in cross-cultural international business research: Assessment and guidelines. *Journal of International Business Studies*, 39(6), 1027–1044.
- Javidan, M., House, R., Dorfman, P., Hanges, P., & Luque, M. d. (2006). Conceptualizing and measuring culture and their consequences: A comparative review of GLOBE's and Hofstede's approaches. *Journal of International Business Studies*, 37, 897–914.
- de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research (JMR)*, 45(1), 104–115.
- Katsikeas, C. S., Samiee, S., & Theodosiou, M. (2006). Strategy fit and performance consequences of international marketing standardization. *Strategic Management Journal*, 27(9), 867–890.
- Kirkman, B. L., Chen, G., Farh, J.-L., Chen, Z. X., & Lowe, K. B. (2009). Individual power distance orientation and follower reactions to transformational leaders: A cross-level, cross-cultural examination. *Academy of Management Journal*, 52(4), 744–764.
- Kluckhohn, C. (1951). The study of culture. In D. Lerner & H. Lasswell (Eds.), *The policy standard* (pp. 393–404). Stanford: Stanford University Press.
- Knight, G. A., & Cavusgil, S. T. (2004). Innovation, organizational capabilities, and the born-global firm. *Journal of International Business Studies*, 35, 124–141.
- Lenartowicz, T., & Johnson, J. P. (2002). Comparing managerial values in twelve Latin American countries: An exploratory study. *Management International Review (MIR)*, 42(3), 279–307.
- Lenartowicz, T., Johnson, J. P., & White, C. T. (2003). The neglect of intracountry cultural variation in international management research. *Journal of Business Research*, 56(12), 999–1008.
- Lytle, A., Brett, J., Barsness, Z., Tinsley, C., & Janssens, M. (1995). A paradigm for confirmatory cross-cultural research in organizational behavior. *Research in Organizational Behavior*, 17, 167–214.
- McClelland, D. (1961). *The achieving society*. Princeton: Van Nostrand Co.
- Minkov, M. (2007). *What makes us different and similar: A new interpretation of the world values survey and other cross-cultural data*. Sofia: Klasika y Stil Publishing.
- Minkov, M., & Hofstede, G. (2012). Is national culture a meaningful concept? Cultural values delineate homogeneous national clusters of in-country regions. *Cross-Cultural Research*, 46, 133–159.
- Mullen, M. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies*, 26(3), 573–596.
- Nakata, C., & Huang, Y. (2005). Progress and promise: The last decade of international marketing research. *Journal of Business Research*, 58(5), 611–618.
- Naroll, R. (1970). The culture-bearing unit in cross-cultural surveys. In R. Naroll & R. Cohen (Eds.), *A handbook of methods in cultural anthropology* (pp. 721–765). New York: Natural History Press.
- Petersen, J. A., Kushwaha, T., & Kumar, V. (2015). Marketing communication strategies and consumer financial decision making: The role of national culture. *Journal of Marketing*, 79(1), 44–63.

- Petruzzellis, L., & Craig, C. S. (2016). Separate but together: Mediterranean identity in three countries. *Journal of Consumer Marketing*, 33(1), 9–19.
- Pike, K. (1967). *Language in relation to a unified theory of the structure of human behavior*. The Hague: Mouton & Co.
- Podsakoff, P., MacKenzie, C., Lee, J., & Podsakoff, N. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903.
- Ralston, D. A., Holt, D. H., Terpstra, R. H., & Yu, K.-C. (1997). The impact of national culture and economic ideology on managerial work values: A study of the United States, Russia, Japan, and China. *Journal of International Business Studies*, 28(1), 177–207.
- Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20(3), 643–641.
- Saeed, S., Yousafzai, S. Y., & Engelen, A. (2014). On cultural and macroeconomic contingencies of the entrepreneurial orientation–performance relationship. *Entrepreneurship Theory and Practice*, 38(2), 255–290.
- Samaha, S. A., Beck, J. T., & Palmatier, R. W. (2014). The role of culture in international relationship marketing. *Journal of Marketing*, 78(5), 78–98.
- Sarros, J., & Woodman, D. (1993). Leadership in Australia and its organizational outcomes. *Leadership and Organization Development Journal*, 14, 3–9.
- Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizations research: A best-practices approach. *Organizational Research Methods*, 6(2), 169.
- Schwartz, S. (1994). Beyond individualism/collectivism: New cultural dimensions of values. In U. Kim, H. Triandis, C. Kagitcibasi, S. Choi, & G. Yoon (Eds.), *Individualism and collectivism: Theory, methods and applications* (pp. 85–119). Thousand Oaks: Sage.
- Sheng, S., Zhou, K. Z., & Li, J. J. (2011). The effects of business and political ties on firm performance: Evidence from China. *Journal of Marketing*, 75(1), 1–15.
- Siew Imm, N., Lee, J. A., & Soutar, G. N. (2007). Are Hofstede's and Schwartz's value frameworks congruent? *International Marketing Review*, 24(2), 164–180.
- Sivakumar, K., & Nakata, C. (2001). The stampede toward Hofstede's framework: Avoiding the sample design pit in cross-cultural research. *Journal of International Business Studies*, 32(3), 555–574.
- Soares, A., Farhangmehr, M., & Shoham, A. (2007). Hofstede's dimensions of culture in international marketing studies. *Journal of Business Research*, 60, 277–284.
- Sondergaard, M. (1994). Research note: Hofstede's consequences: A study of reviews, citations and replications. *Organization Studies*, 15(3), 447–456.
- Steenkamp, J. (2001). The role of national culture in international marketing research. *International Marketing Review*, 18(1), 30–44.
- Steenkamp, J. (2005). Moving out of the U.S. Silo: A call to arms for conducting international marketing research. *Journal of Marketing*, 69(4), 6–8.
- Steenkamp, J.-B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–90.
- Steenkamp, J., Hofstede, F., & Wedel, M. (1999). A cross-national investigation into the individual and national cultural antecedents of consumer innovativeness. *Journal of Marketing*, 63(2), 55–69.
- Tan, J. (2002). Culture, nation, and entrepreneurial strategic orientations: Implications for an emerging economy. *Entrepreneurship Theory and Practice*, 26(4), 95–111.
- Taras, V., Steel, P., & Kirkman, B. L. (2012). Improving national cultural indices using a longitudinal meta-analysis of Hofstede's dimensions. *Journal of World Business*, 47(3), 329–341.
- Tellis, G. J., Prabhu, J. C., & Chandy, R. K. (2009). Radical innovation across nations: The preeminence of corporate culture. *Journal of Marketing*, 73(1), 3–23.
- Triandis, H. (1972). *The analysis of subjective culture*. New York: Wiley.
- Triandis, H. (1994). *Culture and social behavior*. New York: McGraw-Hill.

- Tsui, A. S., Nifadkar, S. S., & Ou, A. Y. (2007). Cross-national, cross-cultural organizational behavior research: Advances, gaps, and recommendations. *Journal of Management*, 33(3), 426–478.
- Vida, I., Obadia, C., & Kunz, M. (2007). The effects of background music on consumer responses in a high-end supermarket. *The International Review of Retail, Distribution and Consumer Research*, 17(5), 469–482.
- van Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks: Sage.
- Waldman, D. A., de Luque, M. S., Washburn, N., House, R. J., Adetoun, B., Barrasa, A., et al. (2006). Cultural and leadership predictors of corporate social responsibility values of top management: A GLOBE study of 15 countries. *Journal of International Business Studies*, 37(6), 823–837.
- Wong, N., Rindfleisch, A., & Burroughs, J. (2003). Do reverse-worded items confound measures in cross-cultural consumer research?: The case of the material values scale. *Journal of Consumer Research*, 30(1), 72–91.
- Zhou, K. Z., Su, C., & Bao, Y. (2002). A paradox of price-quality and market efficiency: A comparative study of the US and China markets. *International Journal of Research in Marketing*, 19(4), 349–365.



Fusion Modeling

Elea McDonnell Feit and Eric T. Bradlow

Contents

Introduction	148
The Classic Data Fusion Problem in Marketing	148
Mixed Levels of Data Aggregation	151
Developing and Estimating Fusion Models	153
Ex. 1: Fusing Data Using a Multivariate Normal Model	153
Ex. 2: Fusing Data Using a Multivariate Probit Model	163
Summary of the Process for Developing a Fusion Model	165
Summary of Related Literature	167
Literature on Data Fusion	167
Related Missing Data Problems	169
Conclusion	170
Appendix	171
R Code for Generating Synthetic Data and Running Ex. 1 with Stan	171
Stan Model for Ex. 2 (Split Multivariate Probit Data)	175
R Commands for Ex. 2	178
References	179

Abstract

This chapter introduces readers to applications of data fusion in marketing from a Bayesian perspective. We will discuss several applications of data fusion including the classic example of combining data on media viewership for one group of customers with data on category purchases for a different

E. M. Feit (✉)

LeBow College of Business, Drexel University, Philadelphia, PA, USA
e-mail: efeit@drexel.edu

E. T. Bradlow

The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
e-mail: ebradlow@wharton.upenn.edu

group, a very common problem in marketing. While many missing data approaches focus on creating “fused” data sets that can be analyzed by others, we focus on the overall inferential goal, which, for this classic data fusion problem, is to determine which media outlets attract consumers who purchase in a particular category and are therefore good targets for advertising. The approach we describe is based on a common Bayesian approach to missing data, using data augmentation within MCMC estimation routines. As we will discuss, this approach can also be extended to a variety of other data structures including mismatched groups of customers, data at different levels of aggregation, and more general missing data problems that commonly arise in marketing. This chapter provides readers with a step-by-step guide to developing Bayesian data fusion applications, including an example fully worked out in the Stan modeling language. Readers who are unfamiliar with Bayesian analysis and MCMC estimation may benefit by reading the chapter in this handbook on Bayesian Models first.

Keywords

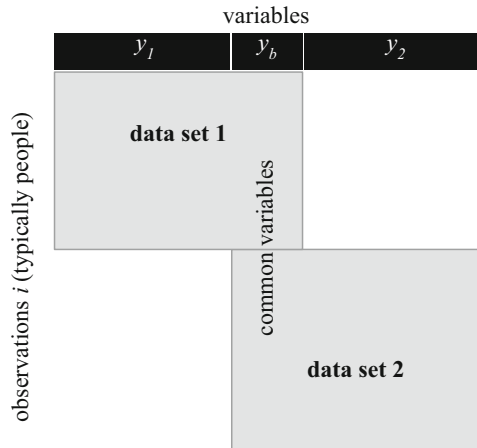
Data fusion · Data augmentation · Missing data · Bayesian · Markov-chain Monte Carlo

Introduction

The Classic Data Fusion Problem in Marketing

Like many other fields, numerous situations arise in marketing where the ideal data for analysis is not readily available. For example, in media planning, marketers want to know whether viewers of a particular media (e.g., television channels or shows, magazines, websites, etc.) purchase a particular product (e.g., breakfast cereal or video games), so that they can decide where to place advertising or estimate the association between exposures to ads and purchases. (See the chapter in this handbook on Return on Media Models for more on marketing response modeling.) Ideally, we would like a data set where the media consumption and purchase behavior are tracked for the same set of customers. However, such data is seldom available. Typically, a media tracking firm (e.g., Comscore, Rentrak) collects data on media usage for one set of consumers, while another firm tracks data on product purchases (e.g., IRI, Niesen, or Dunnhumby for CPG products, Polk for automobiles in the USA, or IMS Health for pharmaceuticals in the USA). Even when media and purchase data are collected by the same firm, it is often impractical to collect that data for the same group of customers, and so firms like Nielsen and Kantar, which collect both purchase and media usage data, typically maintain separate panels for media tracking and purchase tracking. Thus, fusing these separate data sources is a classic problem in marketing analytics (cf. Kamakura and Wedel 1997; Gilula et al. 2006).

Fig. 1 The traditional data fusion problem is to combine two multivariate data sets with different, but overlapping, sets of variables. This data structure occurs in a number of marketing settings and can be addressed as a Bayesian missing data problem



While the goal is to measure the relationship between media usage and product purchase, the data structure that we are faced with is like that shown in Fig. 1, where each row (indexed by i) in data set 1 represents a user in the media panel and the variables in data set 1 describe which content (e.g., channels, shows or websites) user i views. (We let y_{i1} denote the vector of observed variables for each user in data set 1). Data set 2 describes a different set of customers in a purchase panel with observed variables describing which products each consumer has purchased (denoted as y_{i2}). The marketing objective is to estimate what types of products the viewers of some content purchase. The key that makes this possible is a vector of common linking variables which are observed both for customers in data set 1 and customers in data set 2 (y_{ib}), where the subscript b indicates “both.” These variables are often demographics that are collected for users in both data sets. Typically, these demographics are correlated with both media consumption and product purchases (i.e., new parents may be more likely to visit a parenting website and more likely to buy diapers), which enable data fusion.

Beyond the media and purchase data fusion problem, the data structure depicted in Fig. 1 arises in many other contexts in marketing where we observe one set of variables for one group of customers and another set of variables for another group of customers. For example, two retailers considering a co-branding agreement might want to fuse their separate customer lists to estimate how often customers purchase both brands. And even within companies, growing privacy concerns have led firms to avoid maintaining data sets with personal identifiers (e.g., names or addresses) for individual customers. The data fusion methods we describe here can be used to link two data sets which have been de-identified to protect user privacy (Qian and Xie 2014).

In some cases marketing analysts may even plan to create such unmatched data; for instance, in split questionnaire design, marketing researchers minimize the burden on survey respondents by creating a pair of complementary surveys which each can be answered by a subset of the respondents and later fused back together

(Adigüzel and Wedel 2008). Beyond marketing, in educational testing, the data structure in Fig. 1 occurs when data set 1 is students who take the SAT in 2015, data set 2 is those who take it in 2016, and linking variables are test questions (items) that overlap between the two tests.

In all of these examples, the analysis goal is to understand the multivariate relationships between y_{i1} and y_{i2} . The key to linking the two sets of survey data is a set of questions that is common to both surveys, providing the linking variables described in Fig. 1 (y_{ib}).

There are also several closely related data settings where similar methods can be employed including survey subsampling and time sampling (cf. Kamakura and Wedel 2000). In subsampling, some of variables are only collected for a subset of the population, often because those variables are more difficult to collect, e.g., expensive medical tests in population health surveys. In time sampling, a subset of respondents answer a repeated survey at each point in time, avoiding potential respondent fatigue while still collecting data at the desired time interval.

Early approaches to data fusion tackled it as a database integration problem, where in a first stage we match records in data set 1 with records in data set 2 to create a complete database. Statistical modeling and estimation can then proceed as usual with the now-complete data set. For example, the hot deck procedure (Ford 1983), and its more sophisticated variants, can be used to match records in the two data sets using a set of ad hoc rules to match customer i in data set 1 with a customer j in data set 2 who has the same values of y_{ib} . If there are more than one candidate match, the match is selected randomly. If there are a large number of common variables y_{ib} , such that a perfect match to customer i is not always available, then a nearest neighbor approach can be used to match to customers who are similar. In both hot deck and nearest neighbor, once all the customers i in data set 1 are matched to a customer j in data set 2, analysis proceeds as if y_{i1} and y_{j2} were observed from the same customer.

A challenging and often ignored aspect of these two-step imputation-then-analysis approaches is that the uncertainty in the imputation is not propagated forward to the statistical modeling stage (Andridge and Little 2010). In marketing, two-step approaches have become largely superseded by approaches which cast data fusion as a Bayesian missing data problem (Kamakura and Wedel 1997; Gilula et al. 2006; Qian and Xie 2014), which is the approach we will focus on in this chapter.

We focus on analyzing data like that in Fig. 1 as a Bayesian missing data problem: y_{i2} are missing for individuals in data set 1 and y_{i1} are missing for data set 2. Thus, while this chapter resides in the section of this book on data, the approach is more of “a modeling method to handle data that is less than ideal.”

A critical step in analyzing any missing data problem is to consider the process by which the missing data came to be missing. Ideally, data is Missing Completely at Random (MCAR), which means that missingness is unrelated to the observed data or the values of the missing data. When data is MCAR, we can ignore the missing data mechanism in data fusion problems.

In data fusion, this assumption would be violated if one or both of the data sets was a biased sample from the target population. For instance, if a media usage data

set contains mostly lower-income respondents and the relationship between media usage and product usage is different for low- and high-income respondents, then the missing data in Fig. 1 would not be ignorable for overall population-level inferences. This can happen due to poor sampling methods or survey non-response in one or both of the panels. Respondents frequently avoid answering sensitive questions particularly when the true answer is socially undesirable, e.g., viewing content that one might be embarrassed to admit watching. In these instances, the likelihood of a particular survey response being missing depends on the missing response. In these cases, the process that created the missingness can be modeled to avoid bias (Bradlow and Zaslavsky 1999; Ying et al. 2006).

When fusing two data sets that have been carefully sampled from the same target population, we can assume the data is missing by design, which is a special case of Missing Completely at Random (Little and Rubin 2014, Chap. 1). In this case, inference can proceed without explicitly modeling the process that led to the missingness. We are not aware of any published examples of data fusion in marketing where sampling bias or non-response is modeled, although this is a potential area for future research.

The procedure for handling the missing data in Fig. 1 is as follows. If $f(y_i|\theta)$ is the model for the complete vector of responses $y_i = (y_{i1}, y_{i2}, y_{ib})$ with parameters θ , our inference is based on the likelihood of the observed data y^{obs} , which is given by:

$$f(y^{\text{obs}}|\theta) = \int_{y_1^{\text{mis}}} \int_{y_2^{\text{mis}}} \prod_i f(y_i|\theta) dy_2^{\text{mis}} dy_1^{\text{mis}} \quad (1)$$

where y_1^{mis} is the missing observation of y_{i2} in data set 1 and y_2^{mis} is the missing observation of y_{i1} in data set 2.

One way to estimate θ in Eq. 1 is to create a Bayesian MCMC sampler that samples simultaneously from the posterior of θ and the posteriors of the missing data elements y_1^{mis} and y_2^{mis} . This approach is referred to as data augmentation (Tanner and Wong 1987). We will illustrate data augmentation for two alternative specifications of $f(y|\theta)$ in section “[Developing and Estimating Fusion Models](#),” but first we introduce another closely related missing data problem that occurs when merging data from separate sources.

Mixed Levels of Data Aggregation

A second problem that can arise when trying to combine data from two data sources is that the data is provided for individual customers in one data set but is only available in aggregate in another. In analyzing media usage data, this problem occurs because usage of some media channels like websites and mobile apps are easily tracked and linked at the user level, while data on exposure to broadcast media like radio, television, or outdoor signage is only available in aggregate. For example, we might know from a representative panel (e.g., Nielsen People Meter) that approximately 5.3% of a group of users watched a television show, but we do not know

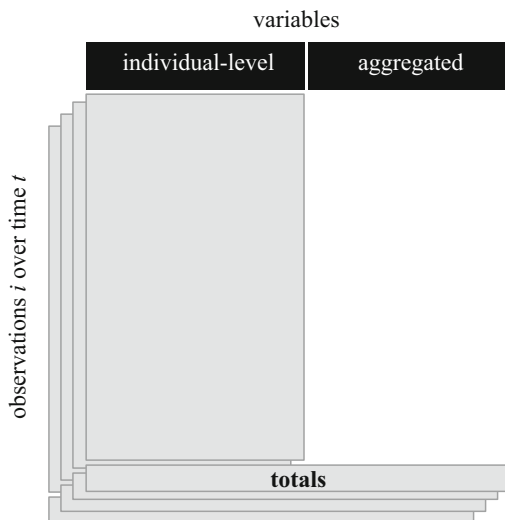
exactly which users those were. Media planners would like to understand the co-usage of media channels – are users who watch some content on TV also likely to watch it on mobile or the web – but we cannot directly observe the co-usage at the consumer level (Feit et al. 2013).

The resulting mixed aggregate-disaggregate data structure is depicted in Fig. 2 where we observe one set of disaggregate variables, y_{i1t} at the individual-level and only totals, Y_{2t} , for a set of aggregate variables. The managerial goal is to infer the correlations across users between the aggregate and the disaggregate variables, which requires repeated observations of y_{i1t} and Y_{2t} over t .

Beyond the media-planning problem described above, this mixed aggregate-disaggregate data structure occurs in other marketing settings, often due to the limitations of tracking systems. For example, a retailer may have detailed customer-level data on visits to an online store, but only aggregate counts of customer visits in physical stores. Even though this data is deficient, one can still use it to infer how many multichannel customers there are and often which customers those are. Similarly, retailers often have customer-level data on coupon redemption (tracked as part of the transaction), but only aggregate data on how many coupons are in circulation. Musalem et al. (2008) show how to use a Bayesian missing data approach with this data to infer “who has the coupon?,” in turn leading to more accurate inference about the effect of coupons on purchases.

Inference for the mixed aggregate-disaggregate data structure described in Fig. 2 can also be viewed as a missing data problem, where the individual-level observations for the aggregated variables, y_{i2t} , are missing; we only observe a total $Y_{2t} = \sum y_{i2t}$ for each period t . As with the traditional data fusion problem, the y_{i2t} is missing by design, and inference can be based on specifying a likelihood for the complete data and then integrating out the missing observations. Specifically, if $f(y_{i1t}, y_{i2t} | \theta)$ is

Fig. 2 In mixed aggregate and disaggregate data, only marginal totals are observed for some variables. Repeated observations of the marginal totals make it possible to identify the individual-level correlations, even when they are not directly observed



the likelihood for the complete individual-level observations that we do not observe, then inference is based on:

$$f(y_{1t}, Y_{2t} | \theta) = \prod_i \int_{y_{i2t}} f(y_{i1t}, y_{i2t} | \theta) dy_{i2t} \quad \text{s.t.} \quad \sum_i y_{i2t} = Y_{2t} \quad (2)$$

The key nuance in Eq. 2 is that the integral over the missing data must conform to the constraint implied by the observed marginal totals $\sum y_{i2t} = Y_{2t}$. By observing covariation between media channels over *time*, and posing a model limiting the covariation structure in the model so that it is driven by user-level behavior, one can estimate the user-level covariation in media usage and make inference about which customers are most likely to have been consuming the aggregated media channel on a given day. By contrast, if we were to aggregate all the data and fit an aggregate time-series model with $\sum y_{i1t} = Y_{1t}$ and $\sum y_{i2t} = Y_{2t}$, it would be impossible to attribute covariation in *aggregate* media usage to co-usage by individual users.

An MCMC sampler can be developed to sample from the posterior of the model in Eq. 2 by developing a way to sample the missing individual-level y_{i2t} such that they conform to the constraint. Thus the method is closely related to the approach used to estimate choice models from aggregate data proposed by Chen and Yang (2007) and Musalem et al. (2008). In fact, in this case, the aggregated data (i.e., constraint) provides information that makes the imputed y_{i2t} even more plausible.

Developing and Estimating Fusion Models

This section provides readers with a step-by-step guide to developing and estimating fusion models by walking the reader through the computation for two examples. These examples are intentionally simplified to allow the reader to focus on the core ideas in data fusion. Our hope is that readers who master these examples will be well-prepared to move on to the more sophisticated examples we discuss in the literature review in section “[Summary of Related Literature.](#)”

Ex. 1: Fusing Data Using a Multivariate Normal Model

We begin with an example of data like that in Fig. 1, where a vector of K_1 variables, y_{i1} , are only observed in the first data set, while another vector of K_2 variables, y_{i2} , are only observed in second data set. As we discussed in the introduction, this is the data structure for split questionnaire designs (where data set 1 and data set 2 represent sub-surveys administered to separate people) and for the classic problem of fusing media consumption data with product purchase data. While creating a complete fused data set (i.e., imputing the missing data in Fig. 1) is often an intermediate step in the analysis, it is important to recognize the ultimate inferential goal in both of these examples is to understand the association between y_{i1} and y_{i2}

despite the fact that those variables are never observed together for the same respondent.

The key to making this inference is that there is a vector of K_b variables that are observed in both data sets, y_{ib} . As we will illustrate, it is vital that these linking variables be correlated with y_{i1} and y_{i2} . If they are independent, then the observed data provide no information about the association between y_{i1} and y_{i2} .

The first step in building a fusion model is to specify a likelihood for the complete observation that we wish we had for all respondents $y_i = (y_{i1}, y_{i2}, y_{ib})$. One simple model for a vector response like this is a multivariate normal distribution:

$$y_i \sim N_{(K_1+K_2+K_b)}(\mu, \Sigma) \quad (3)$$

where N_K denotes the multivariate normal distribution of dimension K and μ and Σ are the mean vector and covariance matrix to be estimated from the data. The multivariate normal model is computationally convenient to work with and so is commonly used in the statistical literature (Little and Rubin 2014; Rässler 2002). It also allows us to estimate correlations between elements of y_i through the covariance matrix Σ . In data fusion problems, we are particularly interested in the correlations between elements of y_{i1} and y_{i2} , which are never observed for the same subject. For example, when combining media consumption and purchase data, these correlations tell us that users who use a particular media channel are likely to purchase a particular product.

While we begin with the simpler multivariate normal model, we should note that the variables we observe in marketing are often binary or discrete, which may not be suitably modeled with a multivariate normal model. In most real cases, an appropriate model is chosen such as a latent cut point model. However, these other models are a relatively straightforward extension of the multivariate model as we will show in Ex. 2.

The core idea of fusion modeling is to estimate the model in (3) using Bayesian methods. Bayesian inference readily handles missing data, including the missing data that is created here due to the fact that some elements of y_i are unobserved for each respondent in the data set. Of course, it would be impossible to follow the approach of estimating the model in (3) using only complete cases, as there are no complete cases.

While our goal is to evaluate the likelihood in Eq. 1, we handle the integral by treating the missing data as unknown and computing the joint posterior of the missing data and the model parameters. In Bayesian inference, all unknown parameters, missing data, and latent variables are treated similarly. Conditional on the model and priors, we compute a posterior distribution for each unknown quantity based on Bayes theorem. Once this joint posterior is obtained, the marginal posteriors of the parameter Σ can be evaluated to understand the correlations in the data. The posteriors of the missing data elements in y_i can also be used to generate a fused data set, if that is desired.

It should be emphasized that by fitting a Bayesian fusion model, we *simultaneously* obtain estimates for both the parameters of interest – in this case μ and

Σ – and impute the missing values in y_i for each respondent, i . Unlike other approaches to missing data which impute in a first stage (often using ad hoc methods) and then estimate parameters in a second stage, the posteriors obtained using Bayesian inference use all the information available in the data and reflect all of the posterior uncertainty resulting from the imputing the missing data.

For all but the most simple Bayesian models, the posterior is obtained by developing an algorithm that will generate random draws from the joint posterior distribution of all unknown parameters. These random samples are then analyzed to estimate the posterior distributions for both the model parameters and the missing data. There are a variety of algorithms for sampling from the posterior distribution that can be adapted to any model including the broad class of Markov-chain Monte Carlo (MCMC) algorithms. When writing MCMC samplers for a fusion model, it is necessary to work out the full conditionals for the missing data and create Gibbs steps to explicitly draw them. An alternative to building the sampler directly is to use a tool like Stan (Carpenter et al. 2016), which allows the user to specify a likelihood using a modeling language and then automatically produces an MCMC algorithm to generate posterior draws from that model. We will illustrate this example using Stan. This code can be run in R, after the Stan software and the `RStan` R package are installed; see the `RStan` Getting Started guide (Stan Core Development Team 2016) for installation instructions.

The first step in estimating the fusion model using Stan is to lay out the Stan model code, which describes the data and the likelihood. We provide this code in Fig. 3.

The `data` block in the code in Fig. 3 tells Stan what data is observed: N_1 observations of a vector of length K_1 called y_1 , N_2 observations of a vector of length K_2 called y_2 , and $N_1 + N_2$ observations of a vector of length K_b called y_b . These correspond to the variables observed only in data set 1, the variables observed only in data set 2, and the common linking variables.

The `parameters` block in Fig. 3 defines the variables for which we want to obtain a posterior. This includes the parameters μ , τ , and Ω , where μ is the mean vector for y_i , τ is a vector of variances, and Ω is the correlation matrix. (This is the preferred parameterization of the multivariate normal in Stan. Note that other MCMC tools like WinBUGS (Spiegelhalter et al. 2003) parameterize the multivariate normal with a precision matrix, the inverse of the covariance matrix.) The `parameters` block also defines the missing elements of y_i : y_{1mis} for the missing variables in data set 1 and y_{2mis} for the missing variables in data set 2. In Stan, the term `parameters` is used for any unknown quantity including both traditional parameters and missing data; following the Bayesian approach to inference, Stan makes no distinction between these two types of unknowns.

In the `transformed parameters` block, there is a bit of code that maps the observed data and the missing data into the full y array. This is simply bookkeeping; the known and unknown elements of y from the `data` and `parameters` blocks are mapped into a single vector. (Note that WinBUGS does not require this step and instead simply assumes that any declared data that is not provided is missing data.)

```

data {
  int<lower=0> N1; //observations in data set 1
  int<lower=0> N2; //observations in data set 2
  int<lower=0> K1;
  int<lower=0> K2;
  int<lower=0> Kb;
  vector[K1] y1[N1];
  vector[K2] y2[N2];
  vector[Kb] yb[N1 + N2];
}
parameters {
  // mean of complete vector
  vector[K1 + K2 + Kb] mu;
  // correlation matrix for complete vector
  corr_matrix[K1 + K2 + Kb] Omega;
  // variance of each variable
  vector<lower=0>[K1 + K2 + Kb] tau;
  // missing elements in data set 1 (observed in y2)
  vector[K2] y1mis[N1];
  // missing elements in data set 2 (observed in y1)
  vector[K1] y2mis[N2];
}
transformed parameters{
  // create the complete data
  vector[K1 + K2 + Kb] y[N1 + N2];
  for (n in 1:N1) {
    for (k in 1:K1) y[n][k] = y1[n][k];
    for (k in 1:K2) y[n][K1 + k] = y1mis[n][k];
    for (k in 1:Kb) y[n][K1 + K2 + k] = yb[n][k];
  }
  for (n in 1:N2) {
    for (k in 1:K1) y[N1+n][k] = y2mis[n][k];
    for (k in 1:K2) y[N1+n][K1+k] = y2[n][k];
    for (k in 1:Kb) y[N1+n][K1+K2+k] = yb[N1+n][k];
  }
}
model {
  //priors
  mu ~ normal(0, 100);
  tau ~ cauchy(0, 2.5);
  Omega ~ lkj_corr(2);
  //likelihood
  y ~ multi_normal(mu, quad_form_diag(Omega, tau));
}

```

Fig. 3 Stan code for multivariate normal data fusion model

Stan requires the user to explicitly declare the observed and missing data and then use the transformed parameters block to define the combined “wished for” data.

In the final `model` block, the model is specified, along with priors for the parameters. The complete y vector is modeled as a multivariate normal with mean vector μ and covariance matrix `quad_form_diag` (Ω , τ), which transforms Ω and τ to the covariance matrix Σ . The prior on μ is the conjugate normal prior, and the priors on τ and Ω are the Cauchy and the LKJ prior for correlation matrices, as recommended by the Stan Modeling Language User's Guide and Reference Manual (Stan Development Team 2017). Note that Stan provides a wide variety of other possible models and priors, and the code in Fig. 3 can be easily modified. More details on how Stan models are specified can be found in the User's Guide.

To estimate the model, the Stan code above is saved in a file and then called using the `stan` function from the `RStan` package in R. (Alternatively, Stan can be called from other languages including Python, MATLAB, Mathematica, and Stata.) In R, the data is passed to Stan as an R list object of elements with the same names and dimensions as defined in the `data` block in the Stan model code, i.e., `d1` is a list with `N1`, `N2`, `K1`, `K2`, `Kb`, `y1`, `y2`, and `yb`. For example, if the data is stored in the R object `d1$data`, its structure would be as follows:

```
> str(d1$data)
List of 8
 $ K1: num 1
 $ K2: num 1
 $ Kb: num 2
 $ N1: num 100
 $ N2: num 100
 $ y1: num [1:100, 1] 1.037 -0.798 0.318 -0.322 0.323 ...
 $ y2: num [1:100, 1] 0.401 -1.821 -1.701 0.726 -0.228 ...
 $ yb: num [1:200, 1:2] 0.607 0.53 1.759 0.49 0.406 ...
```

In this example, the combined vector y_i consists of four variables where the y_{1i} vector is a single variable observed for 100 respondents, y_{2i} is a second single variable observed for a different 100 respondents, and y_{bi} consists of two variables observed for all 200 respondents. Complete code to generate synthetic data and run this example is included in the Appendix and is available online at https://github.com/eleafeit/data_fusion.

If the code above is saved in the file `Data_Fusion.stan` in the working directory of R, then we can obtain draws from the posterior distribution of all unknowns with the following command in R:

```
library(rstan)
m1 <- stan(file="Data_Fusion.stan", data=d1, iter=10000,
           warmup=2000, chains=1)
```

The result is a set of samples from the posterior distribution for μ , τ , Ω , and the missing values of y_i , which are called `y1mis` and `y2mis`. Note the inputs

iter and warmup, which specify that Stan should throw away the first 2000 draws (warmup) and then treat the next $10,000 - 2000 = 8000$ draws as samples from the posterior. See the chapter in this handbook on Bayesian models for more details.

Once this call to Stan from R is completed, the posterior draws are stored in the `m1` object in R. Note the computation may take minutes or hours depending on the size of the data set and the speed of the computer; MCMC algorithms tend to be quite computationally intensive. These draws can be analyzed (typically within R) to make statements about the posteriors of the parameters and the missing data. For instance, a summary of the estimated correlations can be produced with the command:

```
> summary(m1, par=c("Omega"))
```

which results in the output:

```
Summary
      stats
parameter  mean      sd      2.5%      25%      50%      75%      97.5%
Omega[1,1] 1.00000000 0.000000e+00 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
Omega[1,2] 0.41684930 2.061013e-01 0.02423465 0.26268973 0.41374668 0.57872494 0.7835050
Omega[1,3] -0.28081755 7.906684e-02 -0.42740820 -0.33640737 -0.28380772 -0.22730667 -0.1226349
Omega[1,4] 0.64837176 5.226765e-02 0.53581533 0.61582170 0.65217926 0.68441334 0.7424671
Omega[2,1] 0.41684930 2.061013e-01 0.02423465 0.26268973 0.41374668 0.57872494 0.7835050
Omega[2,2] 1.00000000 8.368726e-17 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
Omega[2,3] -0.69550156 4.772961e-02 -0.77889871 -0.72881140 -0.69929899 -0.66573197 -0.5916050
...

```

In this example, the primary inferential goal is to understand the correlation between y_{1i} and y_{2i} , which are the first two elements in the vector y . This corresponds to the `Omega[1,2]` correlation reported in the second row of the above summary. The summary shows that the correlation has a posterior mean of 0.417 with a standard deviation of 0.206, suggesting that the correlation is between 0.024 and 0.784, which is fairly diffuse, but clearly suggests a positive correlation between y_{1i} and y_{2i} . While this correlation is the key parameter of interest in the data fusion problem, posterior summaries of other parameters can be obtained with similar commands. See the R code in the Appendix for details.

The posterior distributions for μ and Σ are shown graphically in Figs. 4 and 5, and code to produce these graphs is included in the Appendix. The posterior distribution of the key correlation between y_{1i} and y_{2i} is shown in the bottom of the plot in Fig. 5, labeled `Omega.12`. Again, the figure clearly shows that the posterior of this correlation is quite diffuse. However, despite the fact that we never observe y_{1i} and y_{2i} for the same individual, we can infer that that y_{1i} and y_{2i} are positively correlated, which was precisely the goal of our data fusion. Since we generated this data synthetically, we happen to know that the true correlation is 0.3, which the model has recovered reasonably well, despite the fact that y_{1i} and y_{2i} are never observed for the same i and the data set is rather small.

Another important feature to notice in Fig. 5 is that the posterior for the correlation between y_{1i} and y_{2i} (labeled `Omega.12`) is much more diffuse than the other correlations. Since the first two elements of y_i are never directly observed together,

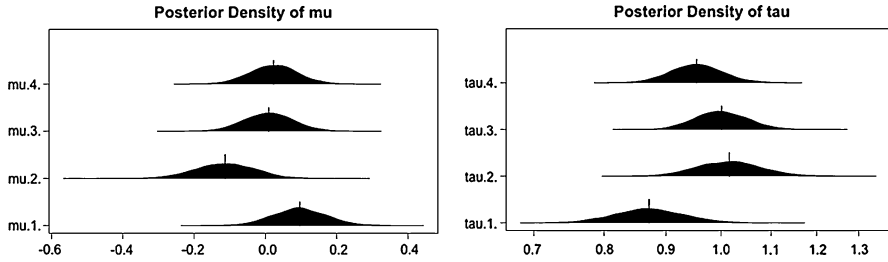
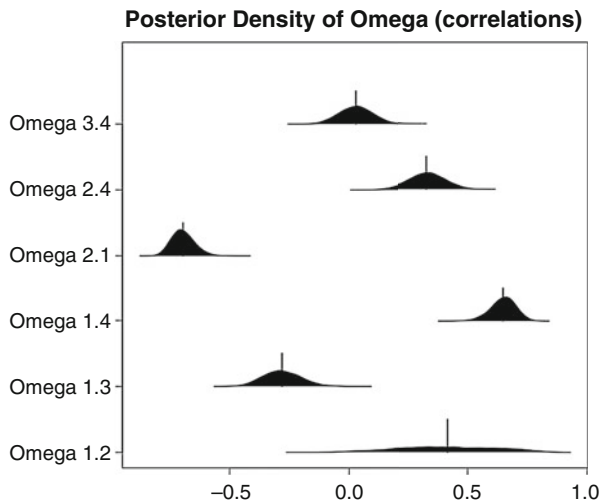


Fig. 4 Posterior distribution of μ (means of multivariate normal for y_i) and τ (variances of y_i) in Ex. 1

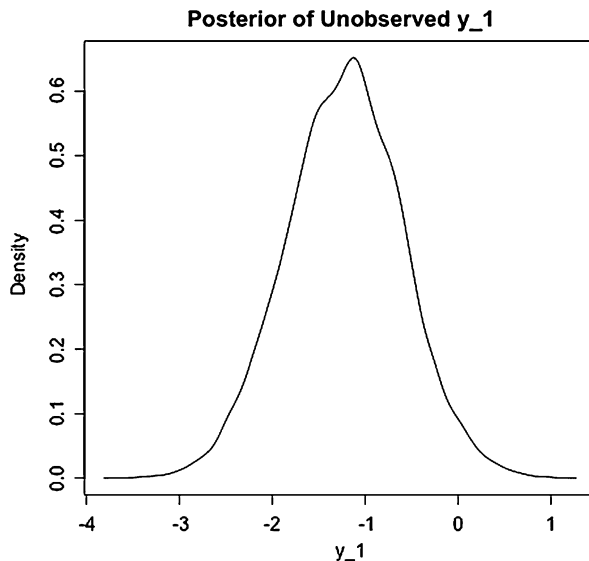
Fig. 5 Posterior distribution of Omega (correlations in the multivariate normal for y_i) in Ex. 1



the data contains only *indirect* information about the correlation. The other correlations are directly observed and therefore better identified resulting in narrower posteriors for the other correlations in Fig. 5. This can be understood by noting that for the correlations where the variables are never observed together, the MCMC sampler is integrating over possible missing values which creates greater diffuseness (appropriately so) in the posterior.

Although our primary goal is to understand the association between y_{i1} and y_{i2} which can be assessed with the posterior of $\text{Omega}.1.2$, the MCMC sampler also produces posterior samples for the missing elements of y_i . An example of one of these posterior distributions is shown in Fig. 6. Although the overall mean of y_{i1} across all respondents (observed and unobserved) is around 0.1 (see Fig. 4), the posterior for this particular respondent is substantially lower and is centered at -1.22 (2.5%-tile = -2.45 , 97.5%-tile = 0.00). Even though we don't observe y_{i1} for this respondent, the posterior for the missing data tells us the likely range of reasonable values of y_{i1} for this respondent, based on his or her observed values for y_{i2} and y_{ib} . The posterior of y_{i1} can be summarized by the mean or median to obtain a "best

Fig. 6 Posterior distribution for one of the unobserved elements of y



estimate” of the missing data for this respondent. These estimates depend on the observed data for individual i , as well as the estimated mean and covariance across the population.

Importantly, because the posterior for the unobserved elements of y_i and the parameters are evaluated simultaneously, the posterior uncertainty in the missing elements of y_i fully accounts for the posterior uncertainty in μ and Σ , and, similarly, the posterior uncertainty in μ and Σ fully accounts for posterior uncertainty in the unobserved elements of y_i . That is, we can say that there is a 95% chance that the missing value of y_{i1} for this respondent is between -2.45 and 0.00 , conditional on the model and our priors.

The MCMC sampler has produced posterior samples for all 100 missing values of y_{i1} and 100 missing values of y_{i2} , and the posterior draws could be summarized to produce a fused data set where the missing values are imputed with posterior means or medians. Although this is unnecessary, if the inferential goal was to measure the correlations in Σ , we can interpret the posteriors for Σ directly. If the goal is to produce a fused data set, then to carry forward the posterior uncertainty into any future analysis, the missing values should be multiply imputed (Rubin 1996), simply by sampling a subset of the posterior draws to create multiple fused data sets. We strongly recommend this approach as opposed to plugging in posterior means or medians (even when appropriately obtained) as biased estimates of nonlinear parameters would occur.

Depending on the context, the imputed individual-level data may also be used to target individual customers. For instance, if y_{2i} represents usage of a particular product, then the imputed values of y_{2i} could be used to target specific customers who are likely to use the product, even if we have never observed those customers’ product usage. This scoring application is useful in any CRM application where the

customers in the data set can be re-targeted, such as fusing the product purchase data between two retailers to identify customers that are good prospects for cross-selling.

To summarize the overall ability of the model to recover the unobserved values in y_{1i} , we plot the posterior medians for all 100 missing observations of y_{1i} against the true value used to generate the data in Fig. 7. (We know the true values because when we generated the data, we drew y_i from a multivariate normal distribution and then removed the “unobserved” elements of y_i .) Figure 7 also shows the posterior uncertainty in the imputation by plotting error bars representing the 2.5 and 97.5%- tiles of the posterior distribution, illustrating the full range of values that the missing data might take. The posterior medians are generally consistent with the true values, and the true value is always contained within the posterior interval. Thus, the fusion model is able to accurately recover the unobserved value of y_{1i} .

Figure 7 shows that the posterior medians for the unobserved y_{1i} tend to be somewhat closer to zero than the true values. (The slope of a best-fit line through the points in Fig. 7 is somewhat less than 1.) This is an example of Bayesian shrinkage, where Bayesian posteriors for individuals tend to be closer to the overall mean and should be expected.

With this example, we have illustrated that fusion modeling is straightforward to execute; however, a word of caution is due. Inference about unobserved values of y_{1i} and y_{2i} and the correlation between them (including non-Bayesian inference) depends critically on their being correlations between the linking variables and y_{1i} and y_{2i} . To illustrate this, we re-estimated the fusion model with two other synthetic data sets that were identical in dimension to the first. One was generated where Σ was diagonal (i.e., no correlations in y_i) and one generated where all correlations in Σ were 0.9. Complete R code for replicating these analyses is included in the Appendix.

When the correlations are high, as shown on the bottom panel in Fig. 8, the missing elements of y_{1i} can be recovered very precisely. The posterior medians are

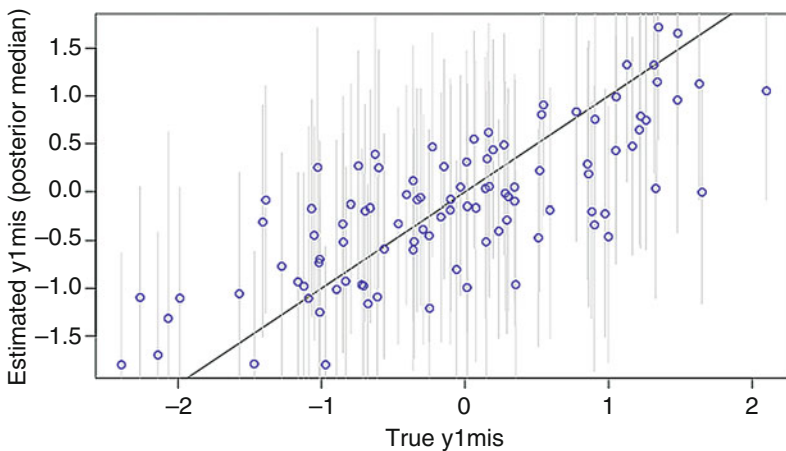


Fig. 7 Posterior estimates of missing elements of y_1 are accurately recovered by the fusion model

close to the true values and the posteriors are quite narrow, which reflects the fact that when there are high correlations the data is more informative about the unobserved elements of y_{1i} than our first example, where the correlations were moderate. However, when the correlations are zero, the data is completely uninformative of the missing observations of y_{1i} . As shown in the top panel of Fig. 8, there is no discernible relationship between the posterior medians and the true values, and the posteriors are so wide that they run off the edges of the plot. This extends to inference about the correlation between y_{i1} and y_{i2} , which has a posterior that is close to uniform between -1 and 1 – which is essentially the same as the prior, reflecting that fact that the data contains no information about the correlation between y_{i1} and y_{i2} . So, even with a substantial number of observations, it is possible

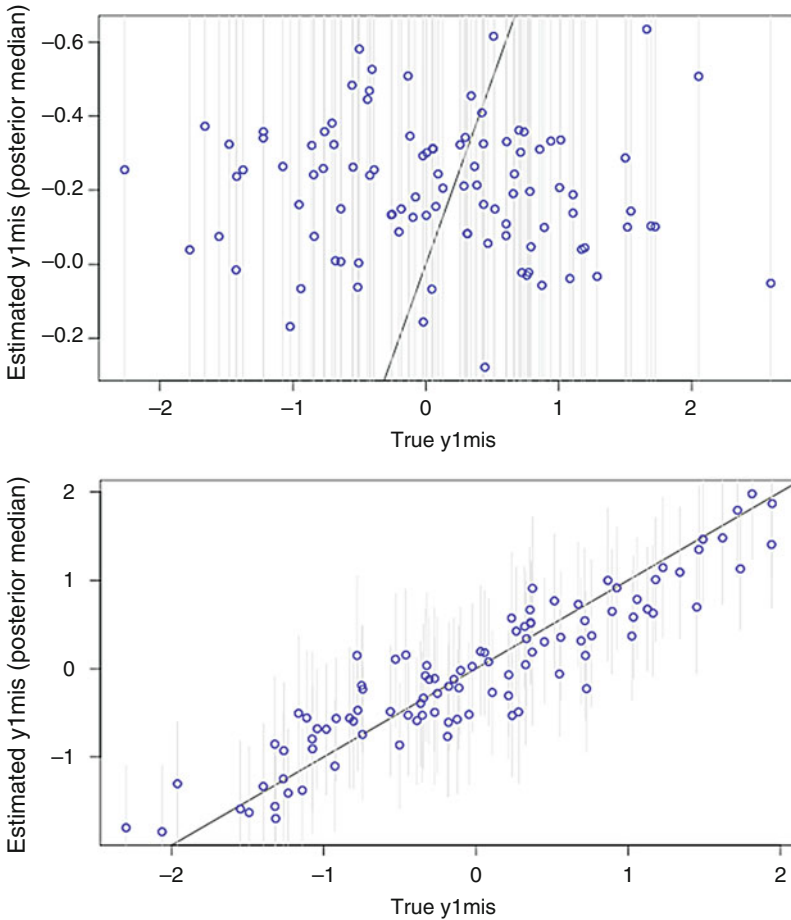


Fig. 8 Inference about missing elements of y_1 depends critically on the correlations between the elements of y . When Σ is diagonal, there is no information in the data about the missing elements of y_{1i} (top). When correlations between elements of y_i are high, unobserved elements of y_{1i} are precisely recovered (bottom)

that the missing elements of y_{i1} and y_{i2} cannot be recovered if the linking variables y_{ib} are not correlated with y_{i1} and y_{i2} .

This discussion relates to the clear distinction between the fraction of missing data and fraction of missing information in the missing data literature (Little and Rubin 2014) where the number of rows that are incomplete may be a poor indicator of how much missing information there is.

The two cases presented in Fig. 8 illustrate why it is extremely important when doing data fusion to carefully choose the linking variables in y_{ib} . For example, in designing a split questionnaire survey, it is important that responses to the questions that are answered by all respondents are correlated with responses to the questions that are only answered by some respondents. (See Adigüzel and Wedel (2008) for more extensive discussion of split questionnaire design.) In fusing media consumption and product purchase data – the classic data fusion example – demographics are usually used as the linking variables. This will work best when demographics are correlated with media consumption (which is very likely but perhaps less so in today’s highly fragmented media landscape) and product purchase (which is likely at the category level but perhaps not at the brand level).

However, even if a poor choice is made for the linking variables, the Bayesian posteriors for the parameters and the missing data will always reflect whatever uncertainty remains. Thus, unlike ad hoc imputation approaches, Bayesian fusion modeling will identify when the linking variables are weak by reporting diffuse posteriors for the individual-level imputations.

Ex. 2: Fusing Data Using a Multivariate Probit Model

As we mentioned earlier, the multivariate normal model is inappropriate for most marketing data, where there are many binary or categorical variables. This is easily accommodated by specifying a latent variable model where an underlying latent vector is normally distributed and then each element of that vector is appropriately transformed to suit the observed data. For example, if the data is binary, which is quite common in marketing, for instance, with “check-all-that-apply”-type questions in a survey or with behavioral variables that track incidence, one can use a multivariate probit model. Assuming that $y_i = (y_{1i}, y_{2i}, y_{bi})$ contains all binary variables, the model for the complete data is:

$$y_{ik} = \begin{cases} 1 & \text{if } z_{ik} > 0 \\ 0 & \text{if } z_{ik} < 0 \end{cases} \tag{4}$$

$$z_i = (z_{1i}, \dots, z_{Ki}) \sim N_K(\mu, \Sigma) \tag{5}$$

where k indexes the elements in y_i from 1 to $K = K_1 + K_2 + K_b$. Complete Stan model code for this model is provided in the Appendix. Note that the variances in Σ are not identified in the multivariate probit model, but associated correlations (Omega) are identified.

We estimated this model using a data set with similar structure as that in Ex. 1. As in the previous example, there is one observed variable in the first data set (y_{i1}), one in the second (y_{i2}), and two linking variables in y_{ib} ; however, these variables are now all binary. As in the previous example, the key inferential goal is to estimate the correlation between y_{i1} and y_{i2} .

Complete R code for running this model is provided in the Appendix; we focus here on the resulting posterior inference. The posterior distribution for the correlations are shown in Fig. 9 which shows that the correlation between y_{i1} and y_{i2} has a posterior mean of 0.368 with a rather wide posterior relative to our first example (2.5%-tile = 0.024, 97.5%-tile = 0.784). This is unsurprising as in the previous example y_{i1} and y_{i2} are never observed for the same subject and, in addition, the binary data used in this example is less informative than the data used in Ex. 1.

For each of the unobserved y_{1i} , we obtain a set of posterior draws that is either 0 or 1. Summarizing these, we can get a probability that a particular missing value is 1. For example, the first missing element of data set 1 is equal to one in 0.296 of the posterior draws, indicating that there is a 0.296 probability that $y_{1,1}$ is one. Consequently, our best estimate of the missing value of $y_{1,1}$ is that it is equal to zero.

We can summarize these best estimates across all individuals in the data set. Comparing these to the true values that generated the data, we get the confusion matrix in Table 1.

So, even with binary data, which is less informative, it is still possible to estimate a fusion model and recover the unobserved variables from each data set reasonably well.

The multivariate probit sampler also produces draws for the underlying continuous normal variables, z_i . In Fig. 10 we plot the posterior means of those estimated

Fig. 9 Posterior distribution of correlations (Omega) from a data fusion model for binary data in Ex. 2

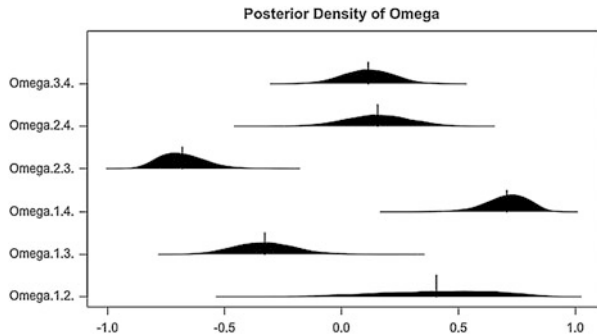


Table 1 Confusion matrix for estimated missing values of y_{1i} in fusion model for binary data

		True value of y_{1i}	
		0	1
Estimated y_{1i}	0	38	17
	1	14	31

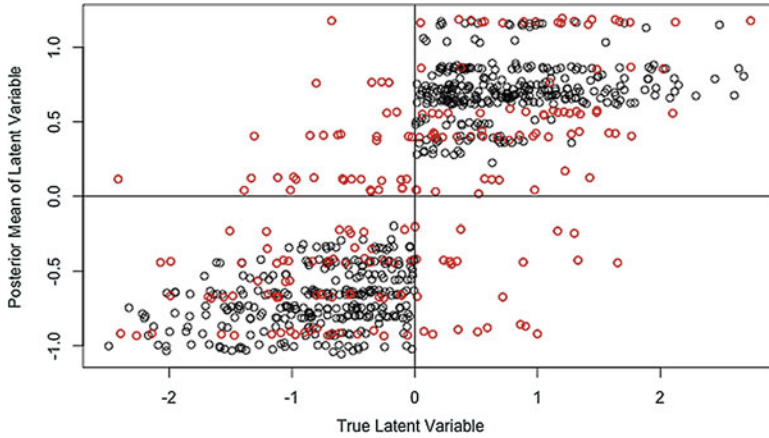


Fig. 10 Posterior means latent variable z for missing binary values y (red) show some confusion where the imputed value is inconsistent with the true value

latent variables. Those z which are associated with an observed binary y are plotted in black, and those z for which the y is missing are plotted in red. In Fig. 10 all of the black points are in the upper right or lower left quadrants, reflecting the fact that when the associated binary variable y is observed, the posterior means for z are consistent with the observed y which is in turn consistent with the sign of the true z . In contrast, the red points appear in all four quadrants, reflecting the same “confusion” we saw in Table 1. However, the points in Fig. 10 do generally follow a diagonal line, reflecting the model’s ability to recover the missing values of y_i for most users and the latent z_i .

Examples 1 and 2 illustrate simple data fusion models for continuous data and binary data. Example 2 uses a continuous normal latent variable to model a binary observation, and this strategy can be extended to allow for ordinal responses, truncated continuous responses or a combination of different variable types. These models can also be extended to allow for mixed levels of aggregation as we discussed in the “Introduction.” Additionally, one could build any number of model structures to relate the data in both data sets. The Bayesian framework and tools like Stan allow analysts the flexibility to build models that reflect the data-generating process.

Summary of the Process for Developing a Fusion Model

To summarize, the general process for developing a fusion model is as follows:

1. Cast the fusion problem as one of missing data.
2. Consider how the missing data came to be missing. In most data fusion problems, the missing data is *missing by design*, which means we do not need to model the process by which the data became missing as in other missing data settings.
3. Specify a parametric model for the complete “wished-for” or “fused” data.

4. Develop an MCMC sampler for the model. In these examples, we have used Stan which automatically produces a MCMC sampler based on a specified model. Programs similar to Stan include WinBUGS and JAGS. One may also code the sampler directly in a statistical language like R, MATLAB, Python, or Gauss.
5. Treat the missing data as unknowns and estimate them using data augmentation. In the above example, we used Stan to define the missing data as a Stan parameter, which resulted in Stan producing a posterior sample for the missing data. When building a Gibbs sampler from scratch, one would find a way to draw the missing parameters from their full conditional distributions based on the model parameters and the observed data.
6. Analyze the posterior samples to make inferences about the model parameters. Often those model parameters correspond directly to the inferential goals of the project.
7. If desired, create a multiply-imputed “fused” data set by taking several random draws from the posterior for the imputed missing data. The fused data can be used as a basis for targeting individual customers.

A point that should be emphasized is that this approach, like all Bayesian inference, conditions on the specified model for the fused data. Our first example used a multivariate normal model and our second used a multivariate probit model. Models based on the multivariate normal are computationally convenient and common in the literature. For instance, in the context of split questionnaires, Raghunathan and Grizzle (1995) and Adigüzel and Wedel (2008) use a cut point model with an underlying multivariate normal distribution. Rässler (2002) also focuses primarily on data fusion with the multivariate normal. However, as with all model-based inference, a model should be chosen that is appropriate for the data and obtaining a posterior based on that model, and the observed data is generally easy to do using modern Bayesian computational methods.

We focused here on methods that propose a model for the joint distribution of the fused data, (y_1, y_2, y_b) , but Gilula et al. (2006) point out that it is actually only necessary to specify the joint distribution of y_1 and y_2 conditional on y_b . They further point out that most of the two-stage matching approaches implicitly assume independence of y_1 and y_2 conditional on y_b , i.e., $f(y_1, y_2|y_b) = f(y_1|y_b)f(y_2|y_b)$. Relying on this assumption, one can specify and estimate models for $f(y_1|y_b)$ and $f(y_2|y_b)$ directly and then integrate over the observations of y_b in the data to find the joint distribution of y_1 and y_2 . This simplifies the modeling task, eliminating the need to specify a model for the linking variables, y_b . The likelihood of $f(y_1|y_b)$ and $f(y_2|y_b)$ can be modeled using off-the-shelf methods such as generalized linear models. Qian and Xie (2014) expand on this approach by proposing an alternative nonparametric model for $f(y_1|y_b)$ and $f(y_2|y_b)$ that is highly flexible and suitable for both continuous and discrete data.

By contrast, the approaches like that illustrated in Exs. 1 and 2 model the full vector (y_1, y_2, y_b) and do not make the assumption of conditional independence directly. Instead, they identify the conditional dependence through the prior

which yields dependence in the marginal distribution. As we illustrated in Ex. 1, the empirical identification of the full joint distribution can be weak, that is, some parameters of the joint distribution are only identified by the prior. The level of identification depends, sometimes in subtle ways, on the data; in our example identification was weak when y_{ib} was not correlated with y_{i1} and y_{i2} . Empirical identification should always be checked by comparing the prior to the posterior uncertainty; if they are the same, then the data has not provided any information.

Summary of Related Literature

We conclude with a brief summary of the literature in marketing on data fusion and then expand to a number of related papers that use Bayesian missing data methods. Our hope is that the examples provided in the previous section will provide a solid base from which students can tackle the more challenging data fusion problems described in the literature.

Literature on Data Fusion

Table 2 organizes several key papers on data fusion into three related problem domains: (1) the classic data fusion problem, (2) split questionnaires, and (3) mixed aggregate-disaggregate data.

The classic problem of fusing media and purchase data (see Fig. 1) was first recognized by Kamakura and Wedel (1997). They cast the problem as a Bayesian missing data problem, recognizing that the missing data mechanism is missing by design and so is ignorable. They propose a joint model for the fused categorical data (y_{i1} , y_{i2} , and y_{ib}) that is a discrete mixture model where incidence is independent across y_i within each latent group. They also show that it is important to account for the uncertainty caused by the data fusion process and propose a multiple imputation approach that is a predecessor to the Bayesian posterior samples we have described in this chapter. Kamakura and Wedel (2000) build on this work by proposing an alternative factor model which can be used in data fusion. They also point out there are a number other related problems where data is missing by design (including subsampling and time sampling, which we discussed in section “[Introduction](#)”) where the same Bayesian missing data approach may be employed.

Gilula et al. (2006) simplified the data fusion problem by making the assumption of conditional independence between the fused variables. If $p(y_{i1}|y_{ib})$ is assumed to be independent of $p(y_{i2}|y_{ib})$, then it becomes unnecessary to specify the full joint distribution of y_{i1} , y_{i2} , and y_{ib} . Instead, the $p(y_{i1}|y_b)$ and $p(y_{i2}|y_b)$ can be estimated separately (using standard models), and then the joint distribution can be approximated by averaging over the observed empirical distribution of y_{ib} , i.e.,

Table 2 Summary of key data fusion papers

	Paper	Fusion	Contribution
Media and purchase	Kamakura and Wedel (1997)	Joint	Recognizing data fusion as a missing data problem and a discrete mixture model for data fusion with categorical variables
	Kamakura and Wedel (2000)	Joint	Factor model for data fusion with continuous and categorical variables
	Gilula et al. (2006)	Direct	Direct approach to data fusion applied with several off-the-shelf models
	Qian and Xie (2014)	Direct or joint	Nonparametric odds ratio model for data fusion with continuous and categorical variables
Split quest.	Raghunathan and Grizzle (1995)	Joint	Split questionnaire as a missing data problem and a model for continuous and discrete data
	Adigüzel and Wedel (2008)	Joint	Method to <i>design</i> a split questionnaire and a normal multivariate cut point model for data fusion
Agg.	Feit et al. (2013)	Joint	Fusion model for mixed aggregate-disaggregate binary data

$$p(y_{i1}, y_{i2} | D) \approx E_{\theta | D} \left[\frac{1}{N} \sum_{obs} p(y_{i1} | y_{ib}, \theta) p(y_{i2} | y_{ib}, \theta) \right] \quad (6)$$

where D is the observed data and N is the number of observations in the complete data set. This so-called direct approach to data fusion reduces the potential for misspecification and is computationally simpler than the joint modeling approach. This approach works well for the standard data fusion problem, yet the joint modeling approach is often desirable when the data fusion problem is embedded within a more complex model (e.g., Musalem et al. 2008; Feit et al. 2010).

Most recently, Qian and Xie (2014) developed a nonparametric odds ratio model, which they show performs better than the parametric models typical of the prior literature and applies this model using both the direct and the joint modeling approaches. They also identify a new application area for data fusion: combining data collected anonymously on a sensitive behavior with data collected non-anonymously on other behaviors. In their specific application, they fuse data on customer's use of counterfeit products with other shopping and product attitudes.

At about the same time the data fusion was recognized as an important problem in marketing, Raghunathan and Grizzle (1995) proposed similar techniques for analyzing split questionnaires in the statistics literature. They propose a model for combined continuous and categorical data and analyze that model with a fully Bayesian approach, using a Gibbs sampler, as we described in section “[Developing and Estimating Fusion Models](#).” Adigüzel and Wedel (2008) extend the work on split questionnaires, focusing on the problem of split questionnaire *design*, using a pilot sample of complete data to determine which questions should be included in each block of the split questionnaire to obtain the most precise posteriors for the

missing (by design) data. They use a normal multivariate cut point model for the data fusion.

Building on this prior work on data fusion, Feit et al. (2013) brought the problem of combining mixed aggregate and disaggregate data into the marketing literature. Their approach involves building a posterior sampler for the complete individual-level data that is constrained to be consistent with the aggregate data.

Related Missing Data Problems

The application of the Bayesian approach to missing data extends far beyond the data fusion problem. Since the reader of this chapter has, by this point, become familiar with the Bayesian approach to missing data, in this section, we provide a brief overview of other applications of this approach. Table 3 lists a few key papers in this area.

Both Feit et al. (2010) and Qian and Xie (2011) propose solutions to the common problem that the analyst wishes to estimate a regression model but has some missing regressors. Regressors are not typically included in the probability model, and so Feit et al. (2010) illustrate how these missing regressors can be handled by including a model specification for them. Inference then proceeds by simulating from the joint posterior for the regression model parameters, regressor model parameters, and the missing regressors. Their work illustrates how, under the Bayesian framework, the posterior for missing regressors is informed by the observed regression outcomes. Specifically, they show that you can impute consumers product needs (typically modeled as a regressor) from their observed choices in a conjoint study. While Feit et al. (2010) use a standard multivariate probit model for the missing regressors, Qian and Xie (2011) propose a more flexible nonparametric model that can handle a variety of missing regressors. Both papers illustrate the point that researchers should specify a model that reflects their beliefs about the data-generating process, whether that model is one of those proposed in the papers in Table 2 specifically for data fusion or a regression model or a more complex structural model. Once the model is specified, model parameters and missing data are estimated simultaneously, rather than treating missing data as a problem that should be handled prior to data analysis.

Table 3 Summary of related work on Bayesian missing data problems

Problem	Paper	Missing data mechanism
Missing regressors	Feit et al. (2010)	Ignorable
	Qian and Xie (2011)	Ignorable
Aggregated regressors	Musalem et al. (2008)	Ignorable
Survey selection	Bradlow and Zaslavsky (1999)	Non-ignorable
	Ying et al. (2006)	Non-ignorable
	Cho et al. (2015)	Non-ignorable
Anonymous visits	Novak et al. (2015)	Non-ignorable

Musalem et al. (2008) develop a model and estimation routine for a similar problem where individual-level regressors are missing but are observed in aggregate. The specific problem they study is the situation where purchase histories are observed for individual customers, and those customers are observed redeeming coupons, but we don't know which customers have a coupon that they chose not to redeem. Instead, we only observe how many coupons were distributed in aggregate. They simultaneously estimate a model that relates the (unobserved) coupon availability to purchases and imputes "who has the coupon" in a way that is consistent with the aggregate observation.

All the previously discussed literature deals with situations where data is missing by design. That is, the data is missing because the researcher planned not to collect it, and the missingness is therefore ignorable. But many missing data problems in marketing address situations where the missingness is stochastic and related to the missing value, which is non-ignorable. The classic example of this is survey non-response. Bradlow and Zaslavsky (1999) impute individual users' missing satisfaction ratings under the assumption that a user will be less likely to answer a satisfaction question when they do not hold a strong opinion. Similar, Ying et al. (2006) study individual users' movie ratings under the assumption that the likelihood that a user will not rate a movie (probably because they didn't watch it) is related to their (unobserved) rating for that movie. Ying et al. (2006) illustrate that when the correlation between a movie being not-rated and the likely rating is ignored, predicted ratings are less accurate, leading to less a less effective movie recommendation system. More recently, Cho et al. (2015) have revisited missing data in customer satisfaction surveys.

Finally, in a recent application of Bayesian missing data methods, Novak et al. (2015) estimate a model of repeat transactions using customer relationship management (CRM) data, which often has the problem that there are a number of visits where the customer is not identified. These transactions may have been made by an existing customer or by a new customer. They show that when there are so-called anonymous visits, a Bayesian missing data approach can be used to impute the missing user ids and identify the customer who made the anonymous visit.

Conclusion

As readers can see, the general problem of missing data in marketing is very broad. The Bayesian framework can be used for missing Ys, missing Xs, data sets where there is individual and aggregate data, and so on. In fact, the broad class of missing data and data fusion problems, we would argue, is one of the most prevalent among practitioners today who want to leverage all the data that they have even when disparate data sources cannot be directly linked. However, as a final note, we warn again that for those who use these sophisticated methods, one should always pay attention to the mechanism (or hopefully lack thereof) that generated the missing data. If the mechanism is non-ignorable, then one would have to build a likelihood for the missing data process, and that process is often hard to observe and verify.

While much research has been done for over 30 years in this area, as new data sets emerge, we expect this area to remain one of high activity going forward.

Acknowledgments We would like to thank the many co-authors with whom we have had discussions while developing and troubleshooting fusion models and other Bayesian missing data methods, especially Andres Musalem, Fred Feinberg, Pengyuan Wang, and Julie Novak.

Appendix

This appendix provides the code used to generate all examples in this chapter. It is also available online at https://github.com/eleafeit/data_fusion. Note that the results in the chapter were obtained with Stan 2.17. If you use a different version of Stan, you may obtain slightly different results even when using the same random number seed.

R Code for Generating Synthetic Data and Running Ex. 1 with Stan

R Commands for Ex. 1 (Requires Utility Functions Below to Be Sourced First)

```
library(MASS)
library(coda)
library(beanplot)
library(rstan)

# Example 1a: MVN =====
# Generate synthetic data
set.seed(20030601)
Sigma <- matrix(c(1, 0.3, -0.2, 0.7, 0.3, 1, -0.6, 0.4, -0.2,
                 -0.6, 1, 0.1, 0.7, 0.4, 0.1, 1), nrow=4)
d1 <- data.mvn.split(K1=1, K2=1, Kb=2, N1=100, N2=100,
                    mu=rep(0,4), Sigma=Sigma)

str(d1$data)
# Call to Stan to generate posterior draws
m1 <- stan(file="Data_Fusion_MVN.stan", data=d1$data,
           iter=10000, warmup=2000, chains=1, seed=12)
# Summaries of posterior draws for population-level parameters
summary(m1, par=c("mu"))
summary(m1, par=c("tau"))
summary(m1, par=c("Omega"))
plot.post.density(m1, pars=c("mu", "tau"), prefix="Ex1",
                  true=list(d1$true$mu, sqrt(diag(d1$true$Sigma))),
                  returncov2cor(d1$true$Sigma))
```

```

draws <- As.mcmc.list(m1, pars=c("Omega"))
png(filename="Ex1PostOmega.png", width=600, height=600)
beanplot(data.frame(draws[[1]][,c(2:4, 7:8, 12)]),
          horizontal=TRUE, las=1, what=c(0, 1, 1, 0),
          side="second", main=paste("Posterior Density of Omega
          (correlations)", log=""), cex.axis=0.5)
dev.off()
# Summaries of posterior draws for missing data
summary(extract(m1, par=c("y1mis"))$y1mis[,3,])
png("Exly13mis.png")
plot(density(extract(m1, par=c("y1mis"))$y1mis[,3,]),
      main="Posterior of Unobserved y_1", xlab="y_1")
dev.off()
summary(m1, par=c("y")) # posteriors of observed data place a
point mass at the observed value
plot.true.v.est(m1, pars=c("y1mis", "y2mis"), prefix="Ex1",
                true=list(d1$true$y1mis, d1$true$y2mis))

# Example 1b: MVN with zero correlations =====
# Generate synthetic data
set.seed(20030601)
Sigma <- matrix(0, nrow=4, ncol=4)
diag(Sigma) <- 1
# Call to Stan to generate posterior draws
d2 <- data.mvn.split(K1=1, K2=1, Kb=2, N1=100, N2=100,
                    mu=rep(0,4), Sigma=Sigma)
m2 <- stan(file="Data_Fusion_MVN.stan", data=d2$data,
           iter=10000, warmup=2000, chains=1, seed=12)
# Summarize posteriors of population-level parameters
summary(m2, par=c("mu"))
summary(m2, par=c("tau"))
summary(m2, par=c("Omega"))
plot.post.density(m2, pars=c("mu", "tau"), prefix="Ex2",
                  true=list(d1$true$mu, sqrt(diag(d1$true$Sigma)),
                            cov2cor(d1$true$Sigma)))
draws <- As.mcmc.list(m2, pars=c("Omega"))
png(filename="Ex2PostOmega.png", width=600, height=400)
beanplot(data.frame(draws[[1]][,c(2:4, 7:8, 12)]),
          horizontal=TRUE, las=1, what=c(0, 1, 1, 0), side="second",
          main=paste("Posterior Density of Omega", log=""),
          cex.axis=0.5)
dev.off()
# Summaries of posterior draws for missing data
plot.true.v.est(m2, pars=c("y1mis", "y2mis"), prefix="Ex2",
                true=list(d2$true$y1mis, d2$true$y2mis))
# Example 1c: MVN with strong positive correlations =====

```

```

# Generate synthetic data
set.seed(20030601)
Sigma <- matrix(0.9, nrow=4, ncol=4)
diag(Sigma) <- 1
# Call to Stan to generate posterior draws
d3 <- data.mvn.split(K1=1, K2=1, Kb=2, N1=100, N2=100,
                    mu=rep(0,4), Sigma=Sigma)
m3 <- stan(file="Data_Fusion_MVN.stan", data=d3$data,
           iter=10000, warmup=2000, chains=1, seed=12)
# Summaries of population-level parameters
summary(m3, par=c("mu"))
summary(m3, par=c("tau"))
summary(m3, par=c("Omega"))
plot.post.density(m3, pars=c("mu", "tau"), prefix="Ex3",
                  true=list(d1$true$mu, sqrt(diag(d1$true$Sigma))))
draws <- As.mcmc.list(m3, pars=c("Omega"))
png(filename="Ex3PostOmega.png", width=600, height=400)
beanplot(data.frame(draws[[1]][,c(2:4, 7:8, 12)]),
          horizontal=TRUE, las=1, what=c(0, 1, 1, 0), side="second",
          main=paste("Posterior Density of Omega", log=""))
dev.off()
# Summaries of posterior draws for missing data
plot.true.v.est(m3, pars=c("y1mis", "y2mis"), prefix="Ex3",
                true=list(d3$true$y1mis, d3$true$y2mis))

```

Utility Functions for Ex. 1

```

data.mvn.split <- function(K1=2, K2=2, Kb=3, N1=100, N2=100,
                           mu=rep(0, K1+K2+Kb),
                           Sigma=diag(1, K1+K2+Kb))
{
  y <- mvrnorm(n=N1+N2, mu=mu, Sigma=Sigma)
  list(data=list(K1=K1, K2=K2, Kb=Kb, N1=N1, N2=N2,
                y1=as.matrix(y[1:N1, 1:K1], col=K1),
                y2=as.matrix(y[N1+1:N2, K1+1:K2], col=K2),
                yb=as.matrix(y[,K1+K2+1:Kb], col=Kb)),
        true=list(mu=mu, Sigma=Sigma,
                  y1mis=y[1:N1, K1+1:K2],
                  y2mis=y[N1+1:N2, 1:K1]))
}
data.mvp.split <- function(K1=2, K2=2, Kb=3, N1=100, N2=100,
                           mu=rep(0, K1+K2+Kb),
                           Sigma=diag(1, K1+K2+Kb))

```

```

{
  z <- mvrnorm(n=N1+N2, mu=mu, Sigma=Sigma)
  y <- z
  y[y>0] <- 1
  y[y<0] <- 0
  y1mis <- y[1:N1, K1+1:K2]
  y2mis <- y[N1+1:N2, 1:K1]
  y[1:N1, K1+1:K2] <- NA
  y[N1+1:N2, 1:K1] <- NA
  true=list(mu=mu, Sigma=Sigma, z=z, y=y, y1mis=y1mis,
            y2mis=y2mis)
  y[is.na(y)] <- 0
  data=list(K1=K1, K2=K2, Kb=Kb, N1=N1, N2=N2, y=y)
  list(data=data, true=true)
}

plot.post.density <- function(m.stan, pars, true, prefix=NULL){
  for (i in 1:length(pars)) {
    draws <- As.mcmc.list(m.stan, pars=pars[i])
    if (!is.null(prefix)) {
      filename <- paste(prefix, "Post", pars[i], ".png", sep="")
      png(filename=filename, width=600, height=400)
    }
    beanplot(data.frame(draws[[1]]),
              horizontal=TRUE, las=1, what=c(0, 1, 1, 0),
              side="second", main=paste("Posterior Density of",
                                         pars[[i]]))
    if (!is.null(prefix)) dev.off()
  }
}

plot.true.v.est <- function(m.stan, pars, true, prefix=NULL){
  for (i in 1:length(pars)) {
    draws <- As.mcmc.list(m.stan, pars=pars[i])
    est <- summary(draws)
    if (!is.null(prefix)) {
      filename <- paste(prefix, "TrueVEst", pars[i], ".png", sep="")
      png(filename=filename, width=600, height=400)
    }
    plot(true[[i]], est$quantiles[,3], col="blue",
          xlab=paste("True", pars[i]),
          ylab=paste("Estiamted", pars[i], "(posterior median)"))
    abline(a=0, b=1)
    arrows(true[[i]], est$quantiles[,3], true[[i]],
           est$quantiles[,1], col="gray90", length=0)
  }
}

```

```

    arrows(true[[i]], est$quantiles[,3], true[[i]],
           est$quantiles[,5], col="gray90", length=0)
    points(true[[i]], est$quantiles[,3], col="blue")
    if (!is.null(prefix)) dev.off()
  }
}

```

Stan Model for Ex. 2 (Split Multivariate Probit Data)

```

functions {
  int mysum(int[, ] a) {
    int s;
    s = 0;
    for (i in 1:size(a))
      s = s + sum(a[i]);
    return s;
  }
}

data {
  int<lower=0> K1;    // number of vars only observed in data set 1
  int<lower=0> K2;    // number of vars only observed in data set 2
  int<lower=0> Kb;    // number of vars observed in both data sets
  int<lower=0> N1;    // number of observations in data set 1
  int<lower=0> N2;    // number of observations in data set 2
  int<lower=0, upper=2> y[N1+N2, K1+K2+Kb]; // should contain
    zeros in missing positions
}

transformed data {
  int<lower=1, upper=N1+N2> n_pos[mysum(y)];
  int<lower=1, upper=K1+K2+Kb> k_pos[size(n_pos)];
  int<lower=1, upper=N1+N2> n_neg[(N1+N2)*(K1+K2+Kb) - K2*N1
    - K1*N2 - mysum(y)];
  int<lower=1, upper=K1+K2+Kb> k_neg[size(n_neg)];
  int<lower=0> N_pos;
  int<lower=0> N_neg;
  N_pos = size(n_pos);
  N_neg = size(n_neg);
  {
    int i;
    int j;

```

```

i = 1;
j = 1;
for (n in 1:N1) {
    //positions in observed y1
    for (k in 1:K1) {
        if (y[n,k] == 1) {
            n_pos[i] = n;
            k_pos[i] = k;
            i = i + 1;
        } else {
            n_neg[j] = n;
            k_neg[j] = k;
            j = j + 1;
        }
    }
}
for (k in (K1+K2+1):(K1+K2+Kb)) {
    if (y[n,k] == 1) {
        n_pos[i] = n;
        k_pos[i] = k;
        i = i + 1;
    } else {
        n_neg[j] = n;
        k_neg[j] = k;
        j = j + 1;
    }
}
}
for (n in (N1+1):(N1+N2)) { //positions in observed y2
    for (k in (K1+1):(K1+K2+Kb)) {
        if (y[n,k] == 1) {
            n_pos[i] = n;
            k_pos[i] = k;
            i = i + 1;
        } else {
            n_neg[j] = n;
            k_neg[j] = k;
            j = j + 1;
        }
    }
}
}
}
}
}

```

```

parameters {
  vector[K1 + K2 + Kb] mu;
  corr_matrix[K1 + K2 + Kb] Omega;
  vector<lower=0>[N_pos] z_pos;
  vector<upper=0>[N_neg] z_neg;
  vector[K2] z1mis[N1];
  vector[K1] z2mis[N2];
}
transformed parameters{
  vector[K1 + K2 + Kb] z[N1 + N2];
  vector[K2] y1mis[N1];
  vector[K1] y2mis[N2];
  for (i in 1:N_pos)
    z[n_pos[i], k_pos[i]] = z_pos[i];
  for (i in 1:N_neg)
    z[n_neg[i], k_neg[i]] = z_neg[i];
  for (n in 1:N1) {
    for (k in 1:K2) {
      z[n, K1 + k] = z1mis[n, k];
      if (z1mis[n, k] > 0)
        y1mis[n, k] = 1;
      if (z1mis[n, k] < 0)
        y1mis[n, k] = 0;
    }
  }
  for (n in 1:N2) {
    for (k in 1:K1) {
      z[N1 + n, k] = z2mis[n, k];
      if (z2mis[n, k] > 0)
        y2mis[n, k] = 1;
      if (z2mis[n, k] < 0)
        y2mis[n, k] = 0;
    }
  }
}
model {
  mu ~ normal(0, 3);
  Omega ~ lkj_corr(1);
  z ~ multi_normal(mu, Omega);
}

```

R Commands for Ex. 2

```

# Generate synthetic data
set.seed(20030601)
Sigma <- matrix(c(1, 0.3, -0.2, 0.7, 0.3, 1, -0.6, 0.4, -0.2,
                 -0.6, 1, 0.1, 0.7, 0.4, 0.1, 1), nrow=4)
d1 <- data.mvp.split(K1=1, K2=1, Kb=2, N1=100, N2=100, mu=rep
(0,4), Sigma=Sigma)
# Call to Stan to generate posterior draws
m1 <- stan(file="Data_Fusion_MVP.stan", data=d1$data,
           iter=10000, warmup=2000, chains=1, seed=35)
# Summaries of posteriors of population-level parameters
summary(m1, par=c("mu", "Omega"))
plot.post.density(m1, pars=c("mu"), prefix="Ex1MVP", true=list
(d1$true$mu))
png(filename="Ex1MVPPostOmega.png", width=600, height=400)
draws <- As.mcmc.list(m1, pars=c("Omega"))
beanplot(data.frame(draws[[1]][,c(2:4, 7:8, 12)]), horizontal=TRUE,
          las=1, what=c(0, 1, 1, 0), side="second",
          main=paste("Posterior Density of Omega", log=""))
dev.off()
# Summarize posteriors for one of missing values
y1mis.draws <- extract(m1, par=c("y1mis"))[[1]][,1,1] # draws for
  third respondent
mean(y1mis.draws > 0)
# Confusion matrix for missing data
y1mis.est <- summary(m1, par=c("y1mis"))$summary[, "50%"]>0
xtabs(y1mis.est + (d1$true$y1mis>0))
y2mis.est <- summary(m1, par=c("y1mis"))$summary[, "50%"]>0
xtabs(y2mis.est + (d1$true$y2mis>0))
z.est <- data.frame(z.true=as.vector(t(d1$true$z)),
                   y=as.vector(t(d1$true$y)),
                   z.postmed=summary(m1, pars=c("z"))
                   $summary[, "50%"])
png(filename="Ex1MVPTrueVEstz.png", width=600, height=400)
plot(z.est[,c(1,3)], xlab="True Latent Variable",
     ylab="Posterior Mean of Latent Variable")
points(z.est[is.na(z.est$y), c(1,3)], col="red")
abline(h=0, v=0)
dev.off()

```


References

- Adigüzel, F., & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, 45(5), 608–617.
- Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78(1), 40–64.
- Bradlow, E. T., & Zaslavsky, A. M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with no answer responses. *Journal of the American Statistical Association*, 94(445), 43–52.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76.
- Chen, Y., & Yang, S. (2007). Estimating disaggregate models using aggregate data through augmentation of individual choice. *Journal of Marketing Research*, 44(4), 613–621.
- Cho, J., Aribarg, A., & Manchanda, P. (2015). *The value of measuring customer satisfaction*. Available at SSRN 2630898.
- Feit, E. M., Beltramo, M. A., & Feinberg, F. M. (2010). Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science*, 56(5), 785–800.
- Feit, E. M., Wang, P., Bradlow, E. T., & Fader, P. S. (2013). Fusing aggregate and disaggregate data with an application to multiplatform media consumption. *Journal of Marketing Research*, 50(3), 348–364.
- Ford, B. L. (1983). An overview of hot-deck procedures. *Incomplete Data in Sample Surveys*, 2(Part IV), 185–207.
- Gilula, Z., McCulloch, R. E., & Rossi, P. E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, 43(1), 73–83.
- Kamakura, W. A., & Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 34, 485–498.
- Kamakura, W. A., & Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, 37(4), 490–498.
- Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. Hoboken: Wiley.
- Musalem, A., Bradlow, E. T., & Raju, J. S. (2008). Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *Journal of Marketing Research*, 45(6), 715–730.
- Novak, J., Feit, E. M., Jensen, S., & Bradlow, E. (2015). *Bayesian imputation for anonymous visits in crm data*. Available at SSRN 2700347.
- Qian, Y., & Xie, H. (2011). No customer left behind: A distribution-free bayesian approach to accounting for missing xs in marketing models. *Marketing Science*, 30(4), 717–736.
- Qian, Y., & Xie, H. (2014). Which brand purchasers are lost to counterfeiters? An application of new data fusion approaches. *Marketing Science*, 33(3), 437–448.
- Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches* (Vol. 168). New York: Springer Science & Business Media.
- Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429), 54–63.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473–489.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS User Manual Version 1.4, January 2003 at <https://faculty.washington.edu/jmiamot/p548/spiegelhalter%20winbugs%20user%20manual.pdf>.

-
- Stan Development Team. (2017). *Stan modeling language user's guide and reference manual, version 2.17.0*. <http://mc-stan.org>
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.
- Stan Development Team (2016). *Rstan getting started*. <https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started>
- Ying, Y., Feinberg, F., & Wedel, M. (2006). Leveraging missing ratings to improve online recommendation systems. *Journal of Marketing Research*, *43*(3), 355–365.



Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers

P. Ebbes, D. Papies, and H. J. van Heerde

Contents

Introduction	182
What Is Endogeneity?	183
Why and When Does Endogeneity Matter?	187
Price Endogeneity	188
Advertising Endogeneity	189
Detailing Endogeneity	190
Firm Strategies	190
CMO Presence	190
Digital Piracy	190
Summary	191
How to Address Endogeneity in a Regression Model	192
Implementing IV Estimation	199
What Happens in an IV Regression When Using Poor IVs?	205
Extensions of the Basic IV Approach	207
Control Function	207
Multiple Endogenous Regressors	209
Interaction Terms	210
The Benefit of Panel Data	211
Conclusions	214
References	216

P. Ebbes (✉)
HEC Paris, Jouy-en-Josas, France
e-mail: ebbes@hec.fr

D. Papies
School of Business and Economics, University of Tübingen, Tübingen, Germany
e-mail: dominik.papies@uni-tuebingen.de

H. J. van Heerde
School of Communication, Journalism and Marketing, Massey University, Auckland, New Zealand
e-mail: heerde@massey.ac.nz

Abstract

This chapter provides a nontechnical summary of how to deal with endogeneity in regression models for marketing research applications. When researchers want to make causal inference of a marketing variable (e.g., price) on an outcome variable (e.g., sales), using observational data and a regression approach, they need the marketing variable to be exogenous. If the marketing variable is driven by factors unobserved by the researcher, such as the weather or other factors, then the assumption that the marketing variable is exogenous is not tenable, and the estimated effect of the marketing variable on the outcome variable may be biased. This is the essence of the endogeneity problem in regression models. The classical approach to address endogeneity is based on instrumental variables (IVs). IVs are variables that isolate the exogenous variation in the marketing variable. However, finding IVs of good quality is challenging. We discuss good practice in finding IVs, and we examine common IV estimation approaches, such as the two-stage least squares approach and the control function approach. Furthermore, we consider other implementation challenges, such as dealing with endogeneity when there is an interaction term in the regression model. Importantly, we also discuss when endogeneity matters and when it does not matter, as the “cure” to the problem can be worse than the “disease.”

Keywords

Endogeneity · Bias · Regression · Instrumental variables · IV · 2SLS · Omitted variables · Causal inference

Introduction

Suppose a firm sells a product, and at some point the firm anticipates that the product will be in higher demand. For example, the product is likely to be in higher demand due to events such as seasonality, promotions, or free publicity. To benefit from this anticipated positive “demand shock,” the firm decides to raise its price. Despite the rise in price, demand is so strong that there is an increase in sales. A researcher who is examining price and sales data from this firm now observes price increases going together with sales increases. If the researcher is unaware of the demand shock, then (s)he may falsely conclude that an increase in price causes an increase in sales. That is, when the researcher attempts to estimate a price elasticity in a regression model, but does not control for the demand shock in the model, the estimated elasticity of price will be biased. In such circumstances, price is said to be endogenous, and the subsequent optimization of the price level results in suboptimal decision-making.

More formally, endogeneity problems arise when the independent variables in a regression model are correlated with the error term in the model. In the example, the unobserved (to the researcher) demand shock is part of the model’s error term, which is now correlated with the independent variable price. Endogeneity problems are

common in marketing studies that use observational data. Observational data are data where the researcher just records or observes what happens in the marketplace, without interference or experimentation. Observational data includes transaction data (e.g., scanner data, online purchase data) and survey data. In observational data, it is not unlikely that there is some unobserved factor that is part of the model's error term that is correlated with the marketing variable of interest, which is the essence of the endogeneity problem. To address the problem, good practice calls for instrumental variable (IV) estimation techniques. However, to avoid an endogeneity problem altogether, the best approach would be to use experimental data, where the researcher experimentally manipulates the marketing variable.

This chapter provides a nontechnical summary of dealing with endogeneity in market research applications via instrumental variable (IV) estimation. IV methods were developed to overcome the endogeneity problem, but finding suitable IVs is challenging. We discuss good practice in finding instrumental variables and in using these to estimate the model, such as at the two-stage least squares approach and the control function approach. Furthermore, we discuss other implementation challenges, such as dealing with endogeneity when there is an interaction term in the regression model. Importantly, we discuss when endogeneity matters and when not, as the "cure" to the problem can be worse than the "disease."

What Is Endogeneity?

Regression modelling in marketing often centers around the estimation of the effects of marketing activities, such as price or advertising, on a performance metric (e.g., sales or profit). However, managers are strategic in their use of marketing activities and adapt these activities in response to factors that are related to demand, but that are often unobserved by and unknown to the researcher. Regression models in marketing that seek to estimate the causal effect of marketing instruments need to account for such deliberate planning of marketing activities or otherwise may suffer from an endogeneity problem, leading to biased estimates of the effects of the marketing activities on performance.

To illustrate the problem with a simple example, we consider an ice-cream vendor who is selling ice creams on the beach. She is the only ice-cream vendor in the near vicinity. Her main decision is centered on pricing of the ice creams. She knows that when the weather is warm, there are more people on the beach, and they are willing to pay more for the ice creams. To take advantage of this, she increases prices on days with higher temperature and sets prices lower on days with lower temperature.

She asks a researcher to estimate a linear demand model which would help her for her decision-making (e.g., pricing, purchasing ingredients). While she kept daily records of sales and prices for about 2 years, she did not inform the researcher about her price setting strategy using temperature. Hence, the researcher observes a data set consisting of daily dates, prices, and sales for the (let's say) 500 days of observations.

The researcher may now fit a simple linear¹ regression model of the following form:

$$Y_i = \beta_0 + \beta_1 P_i + \varepsilon_i \quad (1)$$

Here, Y_i indicates the i -th sales observation (e.g., number of ice creams sold on day i), P_i indicates the price on day i , and ε_i is the model error term capturing (among other things) all unobserved factors that also affect sales². The coefficient β_1 is the coefficient of interest: the effect of price on sales.

When a researcher estimates model (1) using the ordinary least squares (OLS) approach, (s)he is unlikely to estimate the causal effect of price on sales, i.e., β_1 . The reason is that by using OLS, we implicitly assume that price is exogenous. Price is “exogenous” in this regression model when it does not correlate with the error term ε_i . In other words, OLS estimation requires that the covariance between price and error is zero: i.e., $\text{cov}(P_i, \varepsilon_i) = 0$.

However, in the ice-cream example, we have a problem: the temperature that the ice-cream vendor used to set prices also affects sales. Therefore, temperature is part of the model error term ε_i . Because she used temperature information to set prices, prices and temperature are correlated, i.e., $\text{cov}(P_i, \varepsilon_i) \neq 0$, and prices P_i are said to be “endogenous.” In fact, as she increases price with higher temperature, we tend to see in the data that price increases go together with sales increases as higher temperature also leads to higher sales. In the absence of data on temperature, the increase in price is positively associated with the increase in sales, which could make the estimated price coefficient less negative or even positive.

This is the essence of the endogeneity problem: the estimated effect of a marketing variable (e.g., price) on the dependent variable (e.g., sales) is distorted (biased), because there is a correlation between one or more independent variables (price in the example) and one or more unobserved factors that are part of the regression model’s error term (temperature in the example).

We now consider the problem of endogeneity in the above model a bit more formally. For illustration sake, we assume that prices can be described by a normal distribution with mean μ_p and variance σ_p^2 . We can write $P_i = \mu_p + \nu_i$, where ν_i is normally distributed with mean 0 and variance σ_p^2 . We also assume that the error term ε_i has a normal distribution with mean 0 and variance σ_ε^2 . Because prices are

¹Many demand models in marketing are nonlinear. At the end of this chapter, we briefly discuss nonlinear models. A popular nonlinear demand model to estimate price elasticities is the log-log demand model, where both the dependent and independent variables are the natural logs of the original variables, which can be estimated using standard approaches for linear regression models. Log-log models are also prone to an endogeneity problem.

²We use the cross-sectional setup in Eq. 1 as the leading example in this chapter. A similar logic applies to a time series setup (e.g., when we would view Eq. 1 as a time series model). However, this would require an additional discussion of dealing with potential autocorrelation in the model error terms, which is beyond the scope of this chapter. Therefore, we assume that the error terms ε_i are independent and identically distributed in this chapter.

correlated with the errors, ε_i and v_i have a non-zero covariance, $\text{cov}(\varepsilon_i, v_i) = E(\varepsilon_i v_i) = \sigma_{\varepsilon p}$. Viewing the distribution of Y_i and P_i as a bivariate normal distribution, the conditional mean and variance of Y_i given $P_i = p_i$ are (e.g., Lindgren 1993, p. 423)

$$\begin{aligned} E(Y_i | P_i = p_i) &= \beta_0 + \beta_1 p_i + \frac{\sigma_{\varepsilon p}}{\sigma_p^2} (p_i - \mu_p) \\ &= \left(\beta_0 - \frac{\sigma_{\varepsilon p}}{\sigma_p^2} \mu_p \right) + \left(\beta_1 + \frac{\sigma_{\varepsilon p}}{\sigma_p^2} \right) p_i \end{aligned} \quad (2)$$

$$\text{Var}(Y_i | P_i = p_i) = \sigma_\varepsilon^2 (1 - \rho_{\varepsilon p}^2) = \sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon p}^2}{\sigma_p^2} \quad (3)$$

where $\rho_{\varepsilon p} = \frac{\sigma_{\varepsilon p}}{\sigma_\varepsilon \sigma_p}$ is the correlation between prices and the error term. If the endogeneity in prices is ignored, then standard OLS produces the coefficients $\beta_0^r = \left(\beta_0 - \frac{\sigma_{\varepsilon p}}{\sigma_p^2} \mu_p \right)$ and $\beta_1^r = \left(\beta_1 + \frac{\sigma_{\varepsilon p}}{\sigma_p^2} \right)$ instead of the true parameters β_0 and β_1 . That is, in the previous ice-cream example, where prices were positively correlated with the error term (because she increases prices when temperature is higher), the estimated price coefficient is higher than the true value because $\frac{\sigma_{\varepsilon p}}{\sigma_p^2}$ is positive. As we may expect that the true value β_1 is negative, the OLS estimated price coefficient $\hat{\beta}_1^r$ is “less negative” or potentially positive depending on the magnitude of β_1 and $\frac{\sigma_{\varepsilon p}}{\sigma_p^2}$.

Furthermore, from Eq. 3, we can see that the conditional variance of Y_i given price is less than the true unobserved variance σ_ε^2 . In other words, OLS produces an estimate of the residual variance that is smaller than the actual variance. Hence, using OLS, we are led to believe that the model “fits” better than it actually does. We return to this below when it comes to predictions in the presence of endogenous regressors.

Figure 1 shows a scatterplot for 500 hypothetical daily observations of price and sales for the ice-cream seller from the example above. The solid black line represents the incorrectly estimated demand curve by OLS, whereas the dashed line is the true demand curve (in this case the “curve” is a straight line given that we use linear regression). Indeed, as the equations above suggest, the OLS line is less “steep” than the true line. We discuss how to estimate the correct demand curve using an instrumental variable approach below.

In this stylized example, we have an endogeneity problem because the temperature variable was omitted from the model. If the researcher had known the price setting behavior and had observed the temperature variable in the data set, then (s)he should have included this variable as a covariate in the model in Eq. 1, and this would have taken care of this particular endogeneity problem. Then an OLS regression using both prices and temperature as covariates would have estimated the correct price effect. Unfortunately, in many real-world applications, it is impossible to enumerate all relevant demand drivers, measure them, and include them in the model. Thus, we often cannot fully address the endogeneity problem by just including a set of control

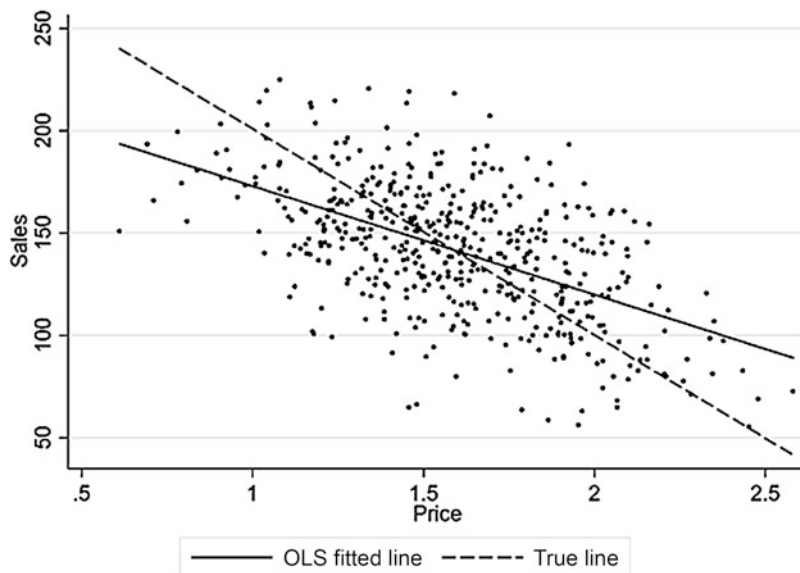


Fig. 1 Scatterplot of sales versus price in the presence of an omitted variable

variables. Nevertheless, we do recommend to always include a rather complete set of control variables as they do make the endogeneity problem less severe.

In sum, if models of consumer demand do not account for marketing instruments (e.g., price, advertising, sales force) that are set strategically, an endogeneity bias is likely present in the estimated regression coefficients (e.g., Petrin and Train 2010; Villas-Boas and Winer 1999). This leads to flawed decisions in determining optimal levels of the marketing instruments, as we show next.

Before we proceed, we would like to highlight two additional aspects. First, the stylized ice-cream vendor example above, as well as most examples that we discuss below, considers transaction data in which endogeneity arises because managers or consumers exhibit strategic (nonrandom) behavior, which is not captured by the regression model. We would like to emphasize, however, that endogeneity concerns are also relevant for survey research. Unobserved respondent characteristics often correlate with the dependent as well as the independent variables in regression models that are estimated using survey data. Hence, causal inferences in cross-sectional survey analyses are only possible if we can rule out these unobserved components. This problem is similar to the common-method bias that is often presented as a serious concern for survey research (Podsakoff et al. 2003).

Second, in marketing models, a regressor may be correlated with the error term, not only because of omitted variables (e.g., temperature) but also because of measurement error. Measurement error in econometrics refers to situations where one or more regressors cannot be measured exactly and are observed with an error. Another cause for an endogeneity problem is when price and demand are determined simultaneously, as is the case, e.g., in an auction for commodities.

In the case of the ice-cream seller example, measurement error and simultaneity problems may happen, for instance, in the following case. Suppose that the ice-cream seller has the goal of selling all ice cream she has in her cart on any given day, e.g., because the ice cream is less fresh the next day. If at the same time, she is afraid of losing customers during a stock-out situation, she may use price to control supply and demand during the day, e.g., increase prices when she observes long queues and decrease prices when there are no queues. If she now reported daily average price, the OLS approach would not capture that price and demand are formed simultaneously. Besides, the average daily price would be a proxy for actual price charged, leading to potential measurement error in price. Similar to omitted variables (e.g., temperature), both measurement error and simultaneity may result in regressor-error dependencies, such that $\text{cov}(P_i, \varepsilon_i) \neq 0$, and standard OLS suffers from an endogeneity problem. For technical details on measurement error and simultaneity, see, e.g., Verbeek (2012) or Greene (2011).

Why and When Does Endogeneity Matter?

From a practical perspective, it may be tempting to dismiss the problem of endogeneity as an academic exercise that is of little relevance to managers. Indeed, many applied textbooks on empirical methods rarely touch upon the issue of endogeneity at all, which may seem to support this argument. We, however, argue the opposite: the problem of endogeneity is of high managerial relevance because obtaining correct effect estimates is essential.

Bijmolt et al. (2005) analyze 1851 estimated price elasticities at the brand level that were published across 40 years in 81 articles. They find that the estimated price elasticity when endogeneity is controlled for is, on average, -3.74 . In contrast, the estimated price elasticity when endogeneity is ignored is, on average, -2.47 , which is quite a strong difference. In other words, when we do not control for endogeneity, the price elasticity estimate is biased toward zero (less negative), similar to the ice-cream vendor example above.

Moreover, there can be endogeneity in other important marketing variables as well. Sethuraman et al. (2011) investigate in a meta-analysis of advertising effectiveness the potential endogeneity bias in the estimated advertising elasticity. They find that the estimated advertising elasticity is lower when endogeneity is not accounted for than when it is accounted for. Albers et al. (2010) analyze 506 estimated personal selling elasticities from 75 articles and find that the estimated elasticity when endogeneity is not taken into account is, on average, 0.37 , while it is, on average, 0.28 when endogeneity is accounted for. Hence, the personal selling elasticity is overestimated when endogeneity is not incorporated in model estimation.

Are these empirical findings on endogeneity biases something that managers need to worry about? Yes, we believe so, and here is why. Going back to the ice-cream vendor example, suppose that the ice-cream vendor wanted to change her price strategy from using temperature to profit maximization. She asks the researcher to

calculate the optimal price, given a marginal cost (c) for ice cream of, say, €1. The researcher estimates a demand model using the data described above with a standard estimation approach (e.g., OLS) and finds that the estimated price elasticity of ice cream is -2 . Following the Amoroso-Robinson theorem³, we find that the optimal price would be €2. This optimal price is not correct, because the true price elasticity is underestimated. Suppose that the true price elasticity is -3 . Using this number, we find that the optimal price is €1.5. Hence, in this example, ignoring endogeneity in estimating price elasticity leads to an “optimal” price of €2 instead of €1.5. Clearly, this is suboptimal. We therefore believe that it is of critical importance to managers and decision-makers to be aware of potential endogeneity problems in estimating marketing regression models.

What are typical situations in which researchers and managers must be aware of potential endogeneity problems? We now consider several examples in detail.

Price Endogeneity

Many retailers such as supermarkets have to decide which items to include and which items to exclude from their assortments. To offer guidance for these decisions, Rooderkerk et al. (2013) develop a model to optimize retailer assortment of laundry detergents. As part of the analysis, they estimate demand models using supermarket scanner data to understand the effect of price on demand, while controlling for price endogeneity. In essence, they estimate a regression model for sales of an SKU (stock keeping unit) in week t as a function of its price. Why may price be endogenous in this case? In the case of laundry detergents, there may be time-varying unobserved demand shocks such as promotional (e.g., coupons) or advertising activities. If these effects are not included in the model, an endogeneity bias in the estimated effect of price is likely.

The same is true if there are brand-specific time-varying shocks in brand popularity and managers use this information to adjust prices. These shocks may arise, for instance, from online buzz or media coverage. One product category in which these variations seem quite natural is experiential goods, such as music or movies.

In cross-sectional analyses, unobserved product characteristics pose a problem. Managers may set prices based on product characteristics such as style, quality, durability, status, service levels, or brand strength. If these factors are not observed by the researcher (which is often the case), an endogeneity bias may occur in the estimated regression parameters (Berry 1994). On the other hand, if we observe variation across brands and across time as in standard panel data applications, these

³For the sake of simplicity in this example, we assume that the researcher estimates a constant elasticity model (e.g., using a log-log regression). The optimal price can then be computed as $p^* = c(\beta/(\beta + 1))$, where c is marginal cost and β is the estimated price elasticity (Amoroso-Robinson theorem, e.g., Homburg et al. 2009, p. 181).

unobserved factors will be less of a concern as they can potentially be controlled for using panel data estimation strategies (as we discuss below).

Advertising Endogeneity

Similar arguments apply to non-price marketing instruments. Consider the following example in Dinner et al. (2014). They address the question of whether advertising in one channel (e.g., online) affects sales in the other channel (e.g., offline or brick-and-mortar sales). They are in particular interested in the direction and magnitude of these effects. Dinner et al. (2014) analyze retailer data to test for the presence of “own-channel” effects (e.g., online display advertising on online sales) versus “cross-channel effects” (e.g., online display advertising on store sales). They estimate the following regression models for online and offline sales in week t :⁴

$$\begin{aligned} \ln \text{ Online Sales}_t &= \beta_0 + \beta_1 \ln \text{ Online Advertising}_t \\ &+ \beta_2 \ln \text{ Offline Advertising}_t + \varepsilon_{1t} \end{aligned} \quad (4)$$

$$\begin{aligned} \ln \text{ Offline Sales}_t &= \gamma_0 + \gamma_1 \ln \text{ Online Advertising}_t \\ &+ \gamma_2 \ln \text{ Offline Advertising}_t + \varepsilon_{2t} \end{aligned} \quad (5)$$

Online advertising includes expenditures on online banner ads and offline advertising expenditures on TV, print, or radio advertising. Since Eqs. 4 and 5 are log-log models, the response parameters can be interpreted as advertising elasticities. For example, β_1 is the own-channel online advertising elasticity (the percentage increase in online sales due to a 1% increase in online advertising), and β_2 is the cross-channel elasticity for the offline advertising on online sales. An endogeneity problem arises when the managers use unobserved demand shocks to adjust their advertising budgets. For instance, a manager may anticipate seasonal shocks in demand. If the manager then allocates her advertising budget accordingly, then this may lead to an overestimation of the advertising elasticity using a standard estimation approach. Dinner et al. (2014) control for endogeneity in model estimation.

As another example, consider an artist who sells records and downloads. If this artist experiences a surge in popularity (e.g., because of online word of mouth or a TV show), the artist may decide to cut down on advertising because she feels that the product does not need the advertising. This would potentially lead to an underestimation of the advertising elasticity if endogeneity is not accounted for, because we tend to observe high demand with relatively low advertising expenditures.

⁴We simplified their model here for sake of exposition. The full model of Dinner et al. (2014) splits online advertising into search advertising and banner ads, allows for advertising carryover effects, and for the effects of other covariates.

Detailing Endogeneity.

Pharmaceutical companies typically spend a substantial amount of their marketing budget on detailing, i.e., sales reps visiting doctors to influence the doctors' prescription behavior. A well-recognized pattern in this domain is that high-volume doctors (i.e., those that often prescribe the drug of the focal brand) receive more detailing. When this is not considered in the regression model, the estimation approach may produce a positive estimated detailing effect, even in the complete absence of a causal effect of detailing on prescription behavior (Manchanda et al. 2004). Accordingly, the meta-analysis by Albers et al. (2010) on personal selling elasticities (that includes detailing) finds that estimation approaches that do not account for endogeneity in detailing studies overestimate the personal selling elasticity.

Firm Strategies

“Firms choose strategies based on their attributes and industry conditions” (Shaver 1998). These attributes may represent a general strategic orientation, management background, or other characteristics that affect firm performance, but are generally difficult to observe for the researcher. Furthermore, firms deliberately choose those strategies that are most likely to increase profit. Hence, most empirical models, in which firm performance is modelled as a function of strategic choices made by firms, are likely to suffer from an endogeneity problem (Shaver 1998).

CMO Presence

Germann et al. (2015) estimate the effect of the presence of a chief marketing officer (CMO) on firm performance. Similar to the point made by Shaver (1998), Germann et al. (2015) argue that the presence or absence of a CMO is likely an endogenous regressor in a regression model because firms do not randomly decide to have a CMO. Rather, the decision to employ a CMO is a strategic choice that will be related to other firm characteristics (e.g., management style, strategic orientation, beliefs about the effect of a CMO). Hence, the variable that captures the presence and absence of the CMO will be correlated with unobserved firm characteristics that also drive firm performance, which will potentially induce an endogeneity bias in the estimated effect of CMO on firm performance, which Germann et al. (2015) address in their analyses.

Digital Piracy

Since the rise of digital distribution of media products, a strong debate has developed on the extent to which digital piracy hurts music sales (see Liebowitz 2016 for an overview). A naïve approach would be to use OLS to estimate a regression model

with artist's sales as dependent variable and piracy (e.g., the number of illegal downloads) as an independent variable. However, in such a regression model, the popularity of the artist would be an omitted variable. We expect that an artist's popularity affects both the piracy level as well as music sales. As both piracy and sales are likely positively correlated with popularity, the OLS estimated effect of piracy on sales will be biased (probably less negative than the true effect of piracy on sales).

An endogeneity problem may also occur when *survey data* are used to estimate the effect of piracy on music sales. We could, for instance, survey a random sample of 10,000 respondents and measure the extent to which they illegally download music as well as their expenditures for music. Respondents highly involved with music are probably more likely to download music illegally, whereas at the same time, they are also more likely to purchase music. Hence, if we would correlate the survey responses of the illegal download and music expenditure questions, we probably would find a rather positive and significant correlation. This estimated positive correlation between music expenditures and illegal download activity is likely to be spurious and should not be interpreted as causal, because the estimated correlation is largely driven by unobserved involvement of the respondents with music.

Summary

Researchers and managers must be aware of the problem of endogeneity whenever they are interested in the causal effect of a marketing (or other) variable on an outcome variable. And, very often managers are interested in causal effects. For instance, the statement "If you change the marketing variable by 1%, the performance changes by $\beta_1\%$ " requires an estimate of the causal effect of the marketing variable on performance. Such a statement is useful for making predictions about the consequences of changing a marketing policy or the value of a marketing instrument. It would tell the manager what would happen in an alternative scenario (Angrist and Pischke 2008).

To avoid endogeneity all together, we would need to run field experiments. In the ice-cream vendor example, for instance, we could run a field experiment where we set prices randomly every day for a period of time. In such an experiment, the random assignment of prices would likely guarantee that there is no correlation between price and other drivers of demand that are unobserved by the researcher. For instance, because of the randomization, we would sometimes observe high prices with high temperatures, but also sometimes observe high prices with low temperatures (and vice versa). Prices are no longer strategically set. This would allow us to use a straightforward (OLS) regression of sales on price to unbiasedly estimate the causal effect of price. Similarly, we could randomly allocate advertising budgets across brands, or we could randomly assign some firms a CMO, while other firms would have to work without a CMO, and observe the effect on sales or firm performance.

Unfortunately, in most marketing applications, such field experiments are not feasible, and we would have to resort to nonexperimental approaches that can be used with observational data. Fortunately, in studies with observational data, we can often develop an identification strategy to estimate the effect of interest (Angrist and Pischke 2008; Germann et al. 2015), as we discuss next.

How to Address Endogeneity in a Regression Model

The goal of a regression analysis is to estimate the causal effect of the regressor (e.g., price) on the dependent variable (e.g., demand). For the reasons discussed above, there is often reason to believe that price and other regressors may be correlated with the error term in the regression equation (e.g., ε_i in Eq. 1). To address the endogeneity problem, a popular approach is to find one or more additional variables, called instrumental variables (IVs), which correlate with the price variable but not with the unobserved determinants of sales (that are part of the error term). This IV approach is the classical approach to address endogeneity in linear regression models (e.g., Greene 2011; Wooldridge 2010). To identify potential IVs, the researcher must have a deep understanding of the practical context of the study, because IVs must meet two requirements (e.g., Angrist and Pischke 2008):

1. The *relevance criterion*, i.e., the IVs must have a strong relation with the endogenous regressors.
2. The *exclusion restriction*, i.e., the IVs must be unrelated to the error of the main Eq. 1.

If one of these two criteria is *not* fulfilled, the IV approach, which we outline in more detail next, *will fail*.

The idea behind IV estimation is to use exogenous variation in the independent variable to estimate the causal effect of the independent variable on the dependent variable. Returning to the ice-cream example, we have to augment the demand equation in (1) with an auxiliary equation for price, where price is modelled as:

$$P_i = \gamma_0 + \gamma_1 Z_i + \nu_i \quad (6)$$

Here Z_i is the IV that must be uncorrelated with ε_i in Eq. 1. Thus, P_i is partitioned into a part $\gamma_0 + \gamma_1 Z_i$ that is exogenous (i.e., uncorrelated with ε_i in Eq. 1) and a random part ν_i that is endogenous (i.e., correlated with ε_i in Eq. 1). The exogenous part is used to estimate the regression parameters of Eq. 1, as we argue below. Thus, the IV enables us to “partition the variation [in prices] into that which can be regarded as clean or as though generated via experimental methods, and that which is contaminated and could result in an endogeneity bias” (Rossi 2014, p. 655). This decomposition of price in an exogenous and endogenous part is schematically represented in Fig. 2.

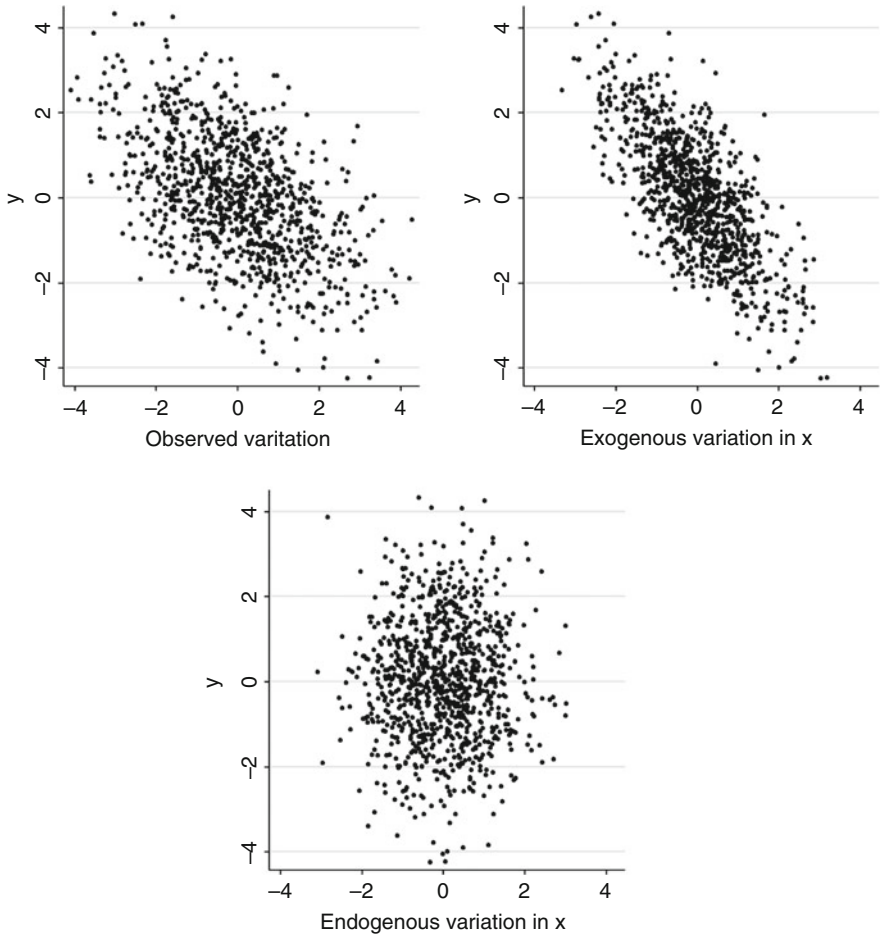


Fig. 2 Decomposing the observed variation of an endogenous regressor in exogenous ($\gamma_0 + \gamma_1 Z_i$) and endogenous (v_i) variation

How can we use the exogenous and “clean” variation in prices to estimate β_0 and β_1 in Eq. 1? The answer is to isolate the exogenous variation through the auxiliary regression model (6), which is typically referred to as the first-stage regression. From this first-stage regression, we obtain the predicted prices, given values for Z . Then, the predicted prices are used instead of the original prices as regressor in Eq. 1. More specifically, this stepwise approach goes as follows:

1. Estimate Eq. 6 using OLS to obtain the estimated parameters $\hat{\gamma}_0$ and $\hat{\gamma}_1$. Compute the predicted values for prices using the estimated parameters, i.e., compute $\hat{P}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_i$.

2. Use the predicted values \hat{P}_i in Eq. 1, instead of observed prices P_i , and estimate the resulting equation with OLS. That is, estimate the following equation with OLS:

$$Y_i = \beta_0 + \beta_1 \hat{P}_i + \varepsilon_i \quad (7)$$

The resulting estimate for β_1 is now the causal effect of prices on sales, and this estimation approach results in a “consistent estimate for β_1 .” A consistent estimate means the estimate converges (in probability) to the true value as the sample size tends to infinity. This stepwise approach is called the two-stage least squares (2SLS) approach.

Many software packages implement the 2SLS approach, and we would caution against doing these steps manually in a research study (see also next section). However, from this stepwise approach, we can get a better intuition as to why IV estimation works. We can see that in step 1, we predict prices using only the exogenous information contained in Z_i . Thus, by construction, \hat{P}_i is exogenous (under the assumption that Z_i is exogenous, i.e., uncorrelated with ε_i). Subsequently, in step 2, the endogenous variable P_i is replaced by its exogenous “alter ego” \hat{P}_i , which is not correlated with the error term ε_i . Thus, we can simply apply OLS in step 2 to estimate Eq. 7 because all regressors are now exogenous.

Importantly, from this stepwise approach, we can also get a better intuition for what criteria an IV must satisfy for this to work. Firstly, Z_i needs to be able to predict the endogenous variable P_i well, i.e., it needs to satisfy the relevance criterion. This criterion can be tested, as we discuss below. When the IV has little explanatory power, the IV is a “weak” IV (Bound et al. 1995), and the predicted values from step 2 above are bad “alter egos” that have little to do with the original price variable. Estimating the resulting Eq. 7 with OLS is not going to give a very precise estimate for β_1 . In the worst case, there is no explanatory power of Z_i at all, i.e. $\gamma_1 = 0$, and the predicted values for P_i would all be the same (and equal to the average value of prices) for $i = 1, 2, \dots, N$, and the price effect β_1 cannot be estimated.

The second requirement is the exclusion restriction. This means that Z_i must be exogenous: uncorrelated with the error term ε_i from the main Eq. 1. Otherwise \hat{P}_i is also endogenous, and we have not solved the endogeneity problem in P_i . This requirement implies that Z_i must be unrelated to all unobserved factors that drive demand and that may be correlated with price. Going back to the example of the ice-cream seller, the Z_i must be uncorrelated with weather and other unobserved factors that are part of the model’s error term. Unfortunately, and this is often the biggest challenge in the implementation of IV, the exogeneity assumption for Z cannot be tested directly. Hence, researchers must rely on the knowledge of the empirical context (i.e., the data generating process), and theory, to argue that their IV meets this criterion.

In the ice-cream example, the cost of ingredients (e.g., milk) may serve as an IV, because these costs will influence consumer prices, but they are unrelated with other unobserved factors that drive consumer demand. In the CMO example above

(Germann et al. 2015), we need a variable that predicts the presence of a CMO, but that is not related to unobserved firm characteristics that drive the decision to appoint a CMO and affect firm performance. For the piracy example, we need a variable that is correlated with an artist's piracy level, but unrelated to shocks in the artist's popularity. We discuss possible IVs for these contexts below.

Hence, two conditions are central to IV estimation: (1) Z_i correlates with P_i , but (2) is uncorrelated with the errors ε_i . Without these two conditions, the IV estimate for β_1 is not consistent, and we cannot interpret the IV estimate as the causal effect of prices on sales. Good IVs can be difficult to find (e.g., Germann et al. 2015, p. 8). The reason is that the two conditions are very hard to meet simultaneously: when the IV is very strongly correlated with the endogenous regressor, it is often hard to argue that the IV is not a direct driver of demand (i.e., it does not satisfy the exclusion restriction). When the IV is completely exogenous (it does satisfy the exclusion restriction), it often is a rather weak IV.

Finding IVs requires "a combination of institutional knowledge and ideas about processes determining the variable of interest" (Angrist and Pischke 2008, p. 117). Likewise, Rossi (2014) notes that "good IVs need to be justified using institutional knowledge because there is no true test for the quality of IVs." Thus, substantive knowledge about the marketing context is needed to identify and argue which variables may be proper candidates for an IV.

To assess whether a candidate IV (or a set of candidate IVs; more on that below) is an appropriate IV, the researcher is advised to perform the following two tasks (see also Germann et al. 2015, pp. 8–9):

1. *Demonstrate* relevance of the IV (i.e., that the IV is not weak).
2. *Argue* that the IV meets the exclusion restriction (i.e., the IV is exogenous).

To perform the first task, the researcher needs to make the case that the IV correlates with the endogenous variable. The arguments should provide a prediction of why and how the IV affects the endogenous variable. This task also includes a discussion of the first-stage regression estimation results (Eq. 6) and assessing whether the estimates make sense in the light of the theoretical context. For instance, do the magnitude and signs of the estimated effects for the IVs (e.g., γ_1) make sense (see also Angrist and Pischke 2008, p. 173)? If these do not make sense, perhaps the hypothesized mechanism for the IVs is not correct or is incomplete. In addition, the researcher should report the R^2 or the F-statistic of the excluded IVs. That is, the researcher needs to run *two* first-stage regressions. The first one includes the IVs and other exogenous variables in the main regression equation, and the second one only includes the exogenous variables but excludes the IVs (note that in the ice-cream vendor example, there are no other exogenous regressors in Eq. 1). Then, the change in R^2 and value of the F-statistic for the comparison of the two models is indicative of the strength of the IVs and should routinely be reported with an IV regression. The bigger the change in R^2 and the higher the F-statistic, the stronger the IVs are.

The second task involves providing arguments for the exclusion restriction of the IV. That is, why is the IV uncorrelated with the omitted variables that affect the dependent variable? Unfortunately, this assumption cannot be tested for, which is arguably the biggest drawback of IV estimation. Therefore, it is important to develop valid theoretical arguments that support this assumption, as the consistency of the IV approach depends on whether this assumption is met or not. Thus, any IV analysis must be accompanied by such a discussion.

How could we develop such arguments? Often it is useful to think of an endogeneity problem in a marketing model as an omitted variable problem, as in the ice-cream vendor example where temperature was an omitted variable. Here, the omitted variable explains variation in both the dependent variable and the endogenous independent variable. For example, temperature drives sales, but also prices, as the merchant used temperature to set prices. When we try to argue for an IV in a marketing regression model, the argument has to make a case that the IV is uncorrelated with the key unobserved factors driving demand. For example, the cost of ingredients (e.g., milk) as an IV is unlikely to be related to the unobserved factors that drive demand for ice creams on the beach, such as the temperature that day, and hence the IV may be valid in the sense that it is uncorrelated with the error term.

Formally speaking, the second task examines the assumption $E(Z_i \varepsilon_i) = 0$. When the researcher has provided arguments in favor of this assumption, and also argued and demonstrated that the IV is not weak (first task), then we may continue with estimating the regression parameters in model (1) using an IV approach. The key to IV estimation is thus to decompose the endogenous regressor into an exogenous part, which is independent of the model error term ε , and an endogenous part v , which is correlated with the error in the regression model for Y . The exogenous part is used to estimate the regression parameters of Eq. 1. As should be clear from this discussion, while there are some empirical checks, the validity of the IV(s) is an assumption that ultimately cannot be tested for.

We would like to add a note on the relation between the criteria of IV strength and IV exogeneity. Consider the following thought experiment. Imagine an IV Z_i that leads to an R^2 in the first-stage regression in Eq. 6 of 0.98. Under the assumption that the IV is truly exogenous, this implies that the endogeneity bias cannot be large in the first place because almost all variation in P_i is exogenous given this very high R^2 . If, in contrast, theory suggests that there is a sizeable endogeneity problem, then such a high R^2 in a first-stage regression makes it implausible that the IV is uncorrelated with the error (Rossi 2014). In general, stronger IVs are less likely to be exogenous, and vice versa, which is unfortunate because an IV needs to be both at the same time.

We have one more remark regarding the underlying assumptions of IVs. Occasionally one can hear or read the statement that “a valid, exogenous IV must be unrelated with the dependent variable Y .” For a demand model with endogenous price, this statement would mean that we need an IV “that is correlated with price but uncorrelated with demand.” These statements are generally *not correct*. Here is why. Consider Eqs. 1 and 6. If we substitute Eq. 6 into Eq. 1, we obtain:

$$\begin{aligned}
 Y_i &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 Z_i + \nu_i) + \varepsilon_i \\
 &= (\beta_0 + \beta_1 \gamma_0) + \beta_1 \gamma_1 Z_i + \beta_1 \nu_i + \varepsilon_i = \pi_0 + \pi_1 Z_i + u_i
 \end{aligned}
 \tag{8}$$

where π_0 , π_1 , and u_i are defined accordingly. For explanation sake, imagine we estimate an OLS regression with Z_i as the independent variable and Y_i as the dependent variable (we note that this regression is typically not performed in practice). We would expect that π_1 is *non-zero*, because π_1 is zero only if $\gamma_1 = 0$, if $\beta_1 = 0$, or if both are 0. In a demand model, we would expect that $\beta_1 < 0$, in general. Furthermore, as argued above, γ_1 has to be non-zero for an IV approach to be valid. Hence, in general, when using an IV approach for a demand model, we are actually expecting π_1 to be non-zero. In other words, the IV likely has an effect on demand in Eq. 8. But, this effect is an indirect effect through the endogenous regressor P_i . For example, assuming that cost is a strong and an exogenous IV (i.e., it meets the exclusion restriction), it must affect demand in (8) ($\pi_1 \neq 0$), because prices are affected by cost ($\gamma_1 \neq 0$) and we expect that prices affect demand ($\beta_1 \neq 0$). In fact, we would expect $\pi_1 < 0$. Hence, it is important to recognize that the assumption for an IV to be exogenous is *not* that the IV is unrelated to demand. Rather, the assumption is that the IV is uncorrelated with the *error term* from the demand equation in (1) (i.e., it meets the exclusion restriction)⁵.

To conclude the discussion on endogeneity and potential remedies, Table 1 gives several examples of endogeneity problems in academic marketing studies. For each example, we point the reader to some ideas of where to find potential good IVs. We briefly outline the rationale for the choice of some of these IVs.

In line with the discussion about the endogenous price of ice cream, a considerable number of studies treat price as endogenous. In many cases, the researchers manage to obtain cost data that they use as an IV. For example, the study by Rooderkerk et al. (2013) in Table 1 uses the costs of ingredients of liquid laundry detergents and transportation costs as IVs. The idea is that these costs should affect the price of the consumer product, but they should be unrelated to brand-specific unobserved demand shocks.

Nevo (2001) proposes to use the prices in other cities (markets) as an IV for price in the focal market. The argument is that these prices (e.g., on the US west coast) capture common cost shocks, but they are unrelated to specific demand shock in the focal city (e.g., on the US east coast). Nevo (2001) highlights potential limitations of these IVs and discusses in great detail the relevant assumptions underlying their validity. Readers interested in an example of how to build a case supporting the validity of an IV should turn to Nevo (2001).

A similarly detailed discussion to build a case for an IV can be found in Germann et al. (2015) regarding the choice of CMO prevalence among peer firms as an instrument for CMO presence in a focal firm. The argument is that the CMO prevalence captures the extent to which it is common among a group of firms to have a CMO, but CMO

⁵This discussion is similar in spirit to the rationale behind mediation analysis. We refer the reader to chapter ► “[Mediation Analysis in Experimental Research](#)” in this handbook for more details.

Table 1 Examples of endogeneity problems and potential IVs

	Dependent variable	Potentially endogenous regressor	Potential causes for endogeneity	Potential IVs (the pages that are cited offer arguments for the suitability of the IVs)
Managers' behavior is source of endogeneity	Demand	Price	Managers set prices based on factors that are related to demand but are unobserved by the researcher: <ul style="list-style-type: none"> • Promotional activities • Changes in popularity • Product characteristics (for cross-sectional analyses) 	Costs of ingredients and transportation (Roederkerk et al. 2013, p. 706); prices in other markets (Nevo 2001, p. 320)
	Demand	Advertising	Managers set advertising based on factors that are related to demand but are unobserved by the researcher: <ul style="list-style-type: none"> • Price changes • Changes in preferences or popularity • (Expected) sales volume 	Advertising expenditures of other (noncompeting) firms (Dinner et al. 2014, p. 534); Advertising cost
	Demand	Detailing	Managers set advertising based on factors that are related to demand but are unobserved by the researcher, e.g., large accounts receive more detailing than small accounts	Costs for detailing (e.g., wages; Chintagunta and Desiraju 2005, p. 74)
Consumers' behavior is source of endogeneity	Firm performance	CMO presence	Firms strategically choose to have a CMO, and this choice depends on unobserved firm characteristics as well as expected performance effects	CMO prevalence among peer firms (Germann et al. 2015, p. 8)
	Demand	Piracy	Unobserved factors (e.g., popularity) affect both a music album's sales as well as piracy	RIAA announcement to start legal actions against file sharing (Bhattacharjee et al. 2007, p. 1364)
	Demand offline channel	Demand online channel	Preferences for channels are correlated	Ease of access to online channel (Gentzkow 2007, p. 726)
	Consumer share of wallet	Membership in loyalty program	Consumers with high intrinsic preference for a firm have a higher share of wallet and are more likely to join the firm's loyalty program	General inclination to join a loyalty program (Leenheer et al. 2007, p. 36)

prevalence among peers is likely unrelated to the focal firm's optimization considerations. In that sense, their argument for the IV's exclusion restriction is similar to the argument provided by Nevo (2001) for prices in demand models.

As a last example, consider the case of piracy and demand for music. Here, unobserved factors such as an artist's popularity will drive both sales and piracy. Bhattacharjee et al. (2007) investigate whether illegal downloads hurt album sales. As a proxy for album sales, they measure how long the album stays in the billboard top charts. They use an announcement of the Recording Industry Association of America (RIAA) to start legal actions against file sharers as the IV for illegal downloads. That is, their IV is a dummy variable that varies for albums observed before and after the announcement. Their motivation for this IV is that the perceived legal cost of file sharing and illegal downloading increases substantially as a result of this announcement. Consequently, we would expect that the IV has a negative effect on illegal downloads. At the same time, it can be argued that the RIAA announcement is unrelated to an album's specific unobserved factors, such as the artist's popularity.

Naturally, researchers must develop arguments whether their IVs fulfill the criteria of relevance and exogeneity. We highlight these studies as examples, as the authors go in great length to build a theoretical case for the validity of their IVs. Without such a discussion of instrument validity, we would not be able to judge whether the IV analyses have merit. We include in Table 1 several references to studies (including the page numbers) that offer arguments for the suitability of the IVs mentioned in Table 1.

We next discuss practical matters of how to implement IV estimation using Stata.

Implementing IV Estimation

We assume that the researcher has one or more valid IVs for the endogenous regressor (we briefly discuss the case of multiple endogenous regressors in the next section). We also assume that the researcher has provided the arguments to demonstrate (i) relevance of the IVs, and (ii) that the IVs meet the exclusion restriction, i.e., there is strong theoretical evidence that the IVs are uncorrelated with the error from the main Eq. 1.

The standard approach to estimate Eq. 1 using IV(s) is to use the two-stage least squares (2SLS) approach. This approach is available in most statistical software packages. For instance, in Stata, researchers can use the `ivreg` or the (higher optioned) `ivreg2` command (Baum et al. 2007, 2015). In R, researchers may use, for instance, the packages `ivmodel`, `ivpack`, or `sem` (`tsls`).

We suggest that the researcher who uses 2SLS, in addition to providing the theoretical argumentation to support the IVs that she uses, discusses the outcomes of the following six empirical tasks (see also Angrist and Pischke 2008):

1. Report both the standard OLS results (i.e., ignoring endogeneity) and the 2SLS results. At the minimum, the estimated coefficients and standard errors from both approaches should be reported. Furthermore, the direction of the bias in OLS as

suggested by 2SLS should be in the expected direction as predicted by theory. For instance, in the ice-cream example, we expected OLS to be biased upward, and 2SLS should therefore give a more negative price effect estimate.

2. Report the complete results of the first-stage regression. The first-stage regression is basically the OLS regression of Eq. 6, including all IVs Z and all other exogenous regressors that are in Eq. 1 (note that in the ice-cream vendor example, there are no other exogenous regressors in Eq. 1). As in the previous step, both the estimated coefficients and the standard errors should be reported. Do the estimated coefficients have the correct sign and magnitude, particularly the coefficient(s) of Z ?
3. Report the R^2 of the first-stage regression and report the R^2 of the same regression excluding all IVs. How big is the difference? The smaller the difference, the worse, because it means that the IVs have little incremental explanatory power. Some researchers suggest reporting the incremental F -statistic of the excluded IVs and argue that this number should be at least 10 (e.g., Stock et al. 2002, Rossi 2014). However, this number should not be seen as standalone threshold that needs to be passed. Rossi (2014) argues that this number may be seen as an absolute minimum requirement. While it is important to discuss the R^2 (or F -statistic) of the first-stage regression Eq. 6, however, as we discuss below, we caution against a comparison of the R^2 of OLS to the R^2 of 2SLS of the second-stage regression (the main Eq. 1) as this comparison is usually meaningless (e.g., Ebbes et al. 2011).
4. When there is more than one IV in the case of one endogenous regressor, we suggest the researcher estimates the model with the best IV (based on theory) and compares the results to the model with all IVs included. Are the results stable in the sense that the main conclusions stay the same? If so, that is good news (more on that below when we discuss Sargan's test). If not, what can explain the difference? The theoretical arguments of IV validity may have to be revisited.
5. Report the results for a test for the presence of endogeneity. This test formally compares the OLS estimates to the 2SLS estimates and tests whether OLS is consistent or not (the latter is the null hypothesis). If the null hypothesis is rejected, then the 2SLS estimates should be used for inferences. In contrast, if the test does not reject the null hypothesis, then the OLS results should be used for inference. This test can be carried out in a few different ways. One way is through a Hausman test (e.g., Verbeek 2012, p. 152; Wooldridge 2010, p. 130). The other way is by estimating the IV regression model through a control function approach (see next section). Regardless of the outcome of the test, we recommend that all previous tasks (1–6) are reported in an IV regression analysis.

We note that sometimes researchers may add an additional task to this list. This additional task involves carrying out a Sargan test, which attempts to test whether the IVs that we use are exogenous. This test is an overidentification test which can only be used if there are more IVs than endogenous regressors. The idea behind this test is as follows. In case we have more IVs than endogenous variables (i.e., we have overidentification), we can regress the fitted residuals from Eq. 7 on the IVs. That is, we run an OLS regression with the fitted residuals as dependent variable

and all IVs (as well as all other exogenous regressors in the regression model, if present) as independent variables. If the set of IVs is exogenous, then they should be jointly unrelated to the residuals, i.e., the R^2 from this regression should be close to zero. If that is not the case, then we have to reject the *entire* set of IVs. The reason is that this test cannot tell which of the IV or IVs are problematic – we can only assess the set of IVs as a whole. The test is usually referred to as the Sargan test (Wooldridge 2010, p. 135), and it is available in packages such as `ivreg2` in Stata. When there are more IVs than endogenous regressors, this test can be used as *additional* evidence besides the theoretical arguments that we develop to choose the IVs. We advise against this test as a substitute for theoretical arguments. For more information on this test, we refer to Wooldridge (2010, p. 135) or Basile (2008).

We now demonstrate the above six tasks for the ice-cream example. We provide a simulated dataset that contains ice-cream sales and prices for 500 days. We generate the data with a true price coefficient of -100 and a true intercept of 300 . In generating the data, we create a correlation between price and the error term such that we have an endogeneity problem. Table 2 contains descriptive statistics for this hypothetical dataset.

The average sales is about 131 servings of ice cream per day, and the average daily price is €1.69. Table 3 contains the estimation results of a simple OLS regression that does not account for endogeneity. The estimated coefficient for price is -62.88 , while the true coefficient in the data generating process (DGP) is -100 . Hence, it becomes apparent that OLS severely underestimates the true price sensitivity of consumers, which is in line with previous research (e.g., Bijmolt et al. 2005). We note that these numbers are the marginal effects of a €1 price change on sales; these are not price elasticities.

We now attempt to solve the endogeneity bias using an IV approach. The IV Z_1 is the sum across the costs for all ingredients that are used for the ice-cream production (e.g., milk, sugar, fruits). The costs do not vary every day as the ice-cream seller often makes bulk purchases for several days. Because the ice-cream seller does not use local produce, the prices of the ingredients that she uses do not depend on local short-term temperature fluctuations. We therefore can accept the assumption that costs are exogenous. In other words, using costs in this example as the IV results in

Table 2 Descriptive statistics of example data set

	Mean	Std. dev.	Min	Max
Sales (units)	130.69	34.15	35.60	226.19
Price (€)	1.69	0.33	0.72	2.63
Instrument 1 (Z_1)	0.99	0.29	0.50	2.00

Table 3 OLS estimation results of example data set

Sales	True β	Estimate for β	<i>se</i>	<i>t</i>	<i>p</i>
Price	-100	-62.88	3.71	-16.96	0.00
Intercept	300	236.99	6.39	37.12	0.00

an IV that meets the exclusion restriction. Table 4 contains the first-stage regression of price on Z_1 .

Here, we expect that the IV (cost) has a positive effect on price, which is the case as the estimate is 0.50 and significant. Such an empirical examination of the sign and magnitude of the estimated effect of the IV on the endogenous regressor is important, as it provides face validity for the theoretical arguments supporting the IV. The R^2 of this regression is 0.20. Because Z is the only regressor in the first-stage regression model, we can conclude that 20% of the variance in price is explained by the IV. Table 4 also reports an F -test, which assesses the joint significance of the IVs. In this case the F -test for the excluded IV is 125.07 and highly significant ($p < 0.0001$). This provides support for the strength of the cost variable as IV.

We now use the IV to estimate Eq. 1 with 2SLS, which we could do by estimating Eq. 7 with OLS, but it is better to rely on a command from the shelf (e.g., `ivreg` in Stata), because otherwise there is a risk that the standard errors of the second stage are not correct (more on that below). Table 5 summarizes the IV results from the 2SLS procedure.

The results indicate that 2SLS accurately estimates the true regression coefficient (-100). Note, however, the large standard error (9.05) compared to the OLS standard error in Table 3 (3.71), which is a typical outcome of an IV regression, i.e., IV is known to be less efficient (have a higher asymptotic variance) than OLS.

In Stata, rather than using `ivreg`, we can also use `ivreg2`, which provides a comprehensive set of additional statistics and diagnostics that are important to examine. For example, the Stata command would be:

```
ivreg2 sales (price = Z1), first.
```

Please note that running this command for this example does not give the Sargan test for the exogeneity of the IV. This makes sense because the Sargan test cannot be conducted if there is exactly the same number of IVs (one here) as endogenous regressors (one here).

We now turn to the Hausman test for endogeneity, which compares the IV estimate to the potentially inconsistent OLS estimate. In case of systematic differences between the IV and OLS estimates, the null hypothesis of no difference between the two will be rejected, and we would conclude that endogeneity is present. In Stata, we can implement this test by first running an OLS regression, storing the

Table 4 First-stage regression results with the IV Z_1

Price	Estimate for γ	<i>se</i>	<i>t</i>	<i>p</i>
Z_1	0.40	0.04	8.95	0.00
Intercept	1.19	0.05	25.78	0.00
F -test excluded IV	df = 498, $F = 125.07$, $p < 0.0001$			

Table 5 2SLS estimation results of example data set

Sales	True β	Estimate for β	<i>se</i>	<i>t</i>	<i>p</i>
Price	-100	-99.99	9.05	-11.05	0.00
Intercept	300	299.73	15.36	19.51	0.00

estimation results, then running an IV regression, and again storing the estimation results. After that we can run the Hausman test. We then tell Stata to compare both estimates:

```
reg sales price
estimates store ols
ivreg sales (price = Z1), first
estimates store iv
hausman iv ols, sigmamore6
```

For the ice-cream example, we find a Hausman χ^2 test statistic of 25.17 (d.f. = 1), which leads to a strong rejection of the null hypothesis of no endogeneity ($p < 0.001$). We therefore conclude that endogeneity is present in the demand model for ice cream and the 2SLS results should be used for interpretation of the effect of price on sales.

It is important to recognize that the validity of the Hausman test depends on the assumption that we used valid IVs. If we use weak or endogenous IVs, the results of the test are likely to be meaningless. Second, we recommend researchers to complement this test with an assessment of the managerial and economic relevance of the difference between the estimated coefficients of OLS and IV. For instance, there may be cases where the test indicates a statistically significant difference, but the difference in estimated coefficients is managerially or economically not relevant.

Recall that for the ice-cream example, we argued that the endogeneity issue arises because the ice-cream seller sets prices based on weather, which was unobserved by the researcher. Let us now assume that we managed to collect information on this variable. We may add the variable that captures weather into the main Eq. 1 as an additional covariate. Table 6 summarizes the results of the OLS regression with both price and temperature as covariates in the model.

The results indicate that controlling for the previously omitted variable eliminates the endogeneity problem. The standard error of the estimated price effect (1.15) is also *much* smaller than in the OLS model where temperature is omitted (3.71), as well as in the IV model where price endogeneity is corrected for using 2SLS (9.05). Hence, it is our advice to develop a regression model that controls for relevant covariates, such that no important covariate (e.g., temperature) that correlates with both the dependent (e.g. sales) and key independent variables (e.g., price) is omitted.

Table 6 OLS estimation results of example data set

Sales	True β	Estimate for β	<i>se</i>	<i>t</i>	<i>p</i>
Price	-100	-100.62	1.15	-87.78	0.00
Temperature	30	29.98	0.39	76.74	0.00
Intercept	300	300.88	1.97	152.89	0.00

⁶The option sigmamore specifies that the covariance matrices are based on the estimated error variance from the efficient OLS estimator. Stata's online help provides more information ("help Hausman").

This approach is also discussed in Germann et al. (2015, p. 4), who refer to such an approach as the “data-rich approach” to endogeneity correction.

Are there ways to assess whether the endogeneity correction was successful by examining in-sample model fit criteria or carrying out holdout-sample validation? For many advances in market response models, (e.g., unobserved parameter heterogeneity, nonlinear functional forms), we can use model fit criteria to assess how well the model performs. Unfortunately, however, this is *not* possible in general for endogeneity correction. The reason is that any endogeneity correction in a linear regression model will tilt the fitted line away from the best fitting OLS line, both in and out of sample. We illustrate this point in Fig. 3, which displays the same scatterplot of price and sales as before, along with the fitted regression lines for IV (dashed) and OLS (solid).

Figure 3 shows that the OLS line fits right through the scatter of sales and price observations. The IV line, however, has a more negative slope because the IV approach estimates the correct (more negative) price effect compared to OLS. However, this does mean that the IV fitted line is tilted away from the OLS fitted line. The OLS line minimizes the sum of squared residuals, and thus the OLS fitted line has the better fit to the observed data, as shown in Fig. 3.

The same principle holds both in an estimation sample and in a holdout sample, as long as the underlying data generating process in the holdout sample has not changed from the data generation process in the estimation sample. In other words, the OLS fitted line will predict better in a holdout sample than the IV fitted line, even though the OLS fitted line is based on biased parameter estimates. For instance, going back to the ice-cream example, suppose that the ice-cream vendor

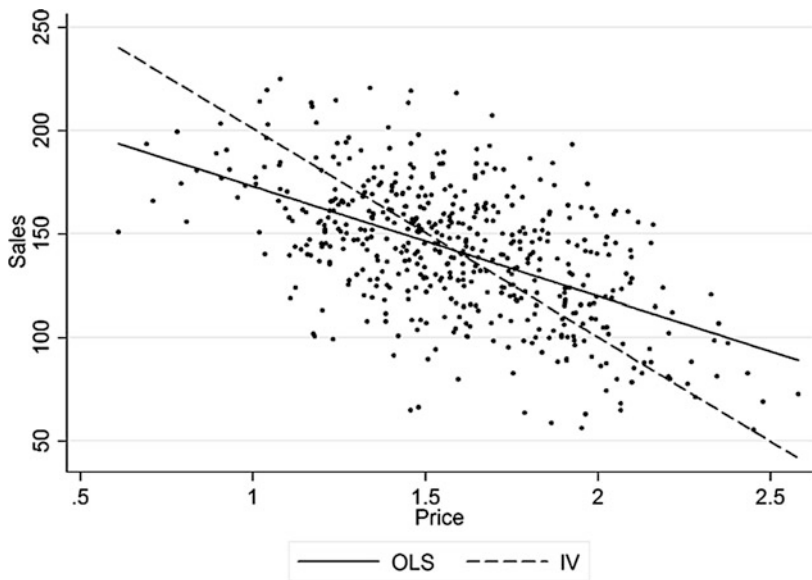


Fig. 3 OLS and IV regression fitted lines for the ice-cream example

wants to predict tomorrow's demand. Suppose also that tomorrow's temperature can be predicted quite accurately and that she is planning to set her prices using tomorrow's temperature information, as is usual practice for her. What would be the best prediction for tomorrow's sales? Looking at the lines in Fig. 3, we are better off using the OLS fitted line (solid line) than the "true" IV fitted line describing the causal effect of price on sales (dashed line) to predict tomorrow's sales, given price.

Ebbes et al. (2011) formalize this discussion and conclude with two main recommendations. First, if the model has a descriptive or normative purpose, consistent estimates for the regression effects are key, and an estimation approach that corrects for endogeneity is required. Descriptive models are developed to provide statements about the effectiveness of marketing instruments (Franses 2005), and normative models are developed to offer a recommended course of action (Leeflang et al. 2000). For either descriptive or normative models, comparisons of in- and out-of-sample fit of OLS versus 2SLS, or any other method that corrects for endogeneity, *are not useful*. Second, if prediction is the primary model objective, Ebbes et al. (2011) recommend to *not* correct for endogeneity. Particularly, if the data generating process in the estimation or holdout samples are the same, then the fitted line obtained from estimating Eq. 1 with OLS will predict as well or better as a model that corrects for endogeneity.

Hence, when considering potential endogeneity in a regression model, the researcher should first decide on the model objective, before attempting to correct for endogeneity. Importantly, we should not use the standard R^2 measure for model fit to assess the success of an endogeneity correction compared to OLS: OLS will typically perform better, despite its biased estimates.

What Happens in an IV Regression When Using Poor IVs?

As we saw above, the IV approach can be an effective way of addressing an endogeneity problem in a linear regression model *if the IVs used are strong and exogenous*. However, the quality of an IV regression deteriorates quickly when one or both of these key requirements are violated. To demonstrate this, we run two IV regressions, one with an endogenous IV and another one with a weak IV, continuing from the earlier hypothetical example.

Suppose we have two additional IVs available, Z_2 and Z_3 , where Z_2 is not exogenous and Z_3 is weak. What would be an example for Z_2 such that it is an endogenous IV? Rossi (2014) argues that lagged marketing variables, such as lagged prices, are often invalid IVs, as they are likely to be endogenous, and should thus not be used as IVs. For instance, in the ice-cream vendor example, using lagged price as IV (e.g., price of the previous day, or P_{t-1}) would likely result in an endogenous IV, because today's temperature is likely correlated with yesterday's temperature, while yesterday's price (the proposed IV) is also correlated with yesterday's temperature, because of the ice-cream vendor's price setting behavior. Hence, $\text{corr}(\varepsilon_t, P_{t-1}) \neq 0$, making the IV $Z_{2t} = P_{t-1}$ endogenous and thus invalid.

An example for a weak IV (Z_3) in the ice-cream vendor case would be, for instance, gasoline prices. The price of gasoline is potentially an exogenous IV because gasoline prices are unlikely to be correlated with today’s temperature or other demand shocks. However, while gasoline prices are part of the cost of producing ice creams (e.g., transporting ingredients), these prices will probably be only a small fraction of total cost and thus are expected to only weakly correlate with the price of ice creams. Hence, we would expect $\text{corr}(P, Z_3) \approx 0$, making the IV Z_3 a weak IV and thus inappropriate.

Table 7 contains the first-stage OLS estimation results using the (invalid) instrument Z_2 . Table 8 shows the IV results for the main Eq. 1, using only Z_2 as an IV.

From Table 7, we can see that the instrument is clearly not weak, as it explains a significant portion of the variation in prices as indicated by the F -test of the excluded IV. However, the results in Table 8 show that the 2SLS estimate for the price effect is severely biased. Interestingly the IV estimate (-47.95) is similarly biased compared to the true value (-100) as the OLS estimate (-62.88 , Table 3), while the IV estimate has a much larger standard error. Applying the Hausman test here to test for the presence of endogeneity would give us χ^2 test statistic of 1.42 (d.f. = 1) and a p -value of 0.23. Hence, the researcher would believe that there is no endogeneity problem in estimating Eq. 1 with OLS, which is clearly a wrong conclusion.

We now turn to the case of using the weak IV Z_3 . Table 9 contains the first-stage estimation results using OLS, and Table 10 displays the results for the main equation using only Z_3 as a weak IV.

From the first-stage regression results in Table 9, we can conclude that Z_3 is a weak IV, because it is not a significant predictor for price. Consequently, the F -test statistic of the excluded IV is very low ($F = 1.54$). As the results in Table 10 show,

Table 7 First-stage regression with the strong but endogenous instrument Z_2

Price	True γ	Estimate for γ	<i>se</i>	<i>t</i>	<i>p</i>
Z_2 : strong but endogenous IV	0.3	0.32	0.05	6.60	0.00
Intercept	0	1.36	0.05	26.10	0.00
F -test excluded IV	df = 498, $F = 43.50$, $p < 0.001$				

Table 8 2SLS estimation results with the strong but endogenous instrument Z_2

Sales	True β	Estimate for β	<i>se</i>	<i>t</i>	<i>p</i>
Price	-100	-47.95	13.27	-3.61	0.00
Intercept	300	211.76	22.46	9.43	0.00

Table 9 First-stage regression with the exogenous but weak IV Z_3

Price	True β	Estimate for γ	<i>se</i>	<i>t</i>	<i>p</i>
Z_3 : exogenous but weak IV	.3	-0.02	0.01	-1.24	0.22
Intercept	0	1.69	0.01	115.11	0.00
F -test excluded IV	df = 498, $F = 1.54$, $p = 0.22$				

Table 10 2SLS estimation results with the exogenous but weak IV Z_3

Sales	True β	Estimate for β	<i>se</i>	<i>t</i>	<i>p</i>
Price	-100	-19.65	75.31	-0.26	0.79
Intercept	300	163.90	127.33	1.29	0.20

the estimated coefficients from the main equation using the IV approach with only Z_3 as IV are now severely biased and have huge standard errors.

Unlike using Z_2 as IV, which is an endogenous IV, we could have identified that Z_3 is inappropriate as an IV because of its weakness, by examining the first-stage regression (Bound et al. 1995). This example illustrates that it is important that researchers assess the strength of the IV by examining the first-stage regression. When the IV is not strong, we should resist the temptation to interpret and use the IV estimation results, as the ones presented in Table 10. Unfortunately, for the endogenous IV Z_2 , an examination of the first-stage regression (Table 7) did not reveal any problems with this IV, and only theoretical arguments would have served to dismiss Z_2 as an appropriate IV.

In sum, when the assumptions underlying the validity of IVs are violated, as was the case for Z_2 and Z_3 , the IV estimates are potentially severely biased, and we would be better off to not use an IV approach to correct for endogeneity, as the “cure” to the problem is worse than the “disease.”

Extensions of the Basic IV Approach

Many empirical regression applications in marketing cannot be addressed by a simple linear market response model with just one independent variable. More likely, we encounter applications where there are multiple endogenous regressors, other regressors (covariates) that are not endogenous, interaction terms with endogenous regressors, or there is endogeneity in the presence of binary dependent or independent variables. We now provide a discussion of several common extensions of the basic IV approach, which may be useful to address an endogeneity problem in a marketing application. We continue using a linear regression model for our discussion; if the original model is a model that can be linearized (e.g., a multiplicative model using the log transform), then the extensions that we discuss next will still apply.

Control Function

There is an alternative way of estimating Eq. 1 using IVs. Instead of using the 2SLS approach, we could use the control function (CF) approach (e.g., Petrin and Train 2010; Ebbes et al. 2011; Wooldridge 2015). Recall that in the 2SLS approach, we replace the observed values of the endogenous regressor by its predictions obtained from the first-stage regression. In contrast, in the control function approach, we add

the *residuals* of the first-stage regression as an *additional regressor* into the main Eq. 1.

We can implement the control function approach by using the same first-stage regression (6) as in 2SLS. After estimating the first-stage regression with OLS, we compute the fitted residuals

$$\hat{\nu}_i = P_i - \hat{P}_i \quad (9)$$

Then, we include these fitted residuals *as an additional regressor in the main regression* Eq. 1 *for the dependent variable*, resulting in

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 \hat{\nu}_i + \varepsilon_i \quad (10)$$

Subsequently, this “augmented” regression equation can be estimated using standard estimation approaches (e.g., OLS). It can be shown that 2SLS and CF give identical estimates for the regression parameters in the linear regression model (e.g., Verbeek 2012). Note that the original P_i is used in Eq. 10 and not \hat{P}_i . The idea is that the term $\hat{\nu}_i$ captures the “omitted” variables that make P_i endogenous. By including this term in Eq. 1, we “control” for endogeneity. Interestingly, a standard t-test for the significance of β_2 would be a fairly straightforward way to test for the presence of endogeneity, i.e., it is a computationally easy version of the Hausman test (step 6 in the previous section).

One note of caution regarding the use of the CF approach is the following. While in linear models the estimated coefficients using the CF approach are identical to the 2SLS estimates, the standard errors when estimating Eq. 10 using OLS are incorrect, because $\hat{\nu}_i$ is an estimated quantity. One way to address this is to use bootstrapping techniques to estimate the correct standard errors. The procedure to compute the correct standard errors using bootstrapping is given in Karaca-Mandic and Train (2003) and Papies et al. (2017).

We illustrate the control function approach using the same data as above. We first estimate Eq. 6, then store the residuals (e.g., in Stata by using the command `predict new_variable, res`), and use these residuals as an additional regressor in the main regression equation. The estimated coefficient (−99.99) using the CF approach is identical to the one obtained from 2SLS (Table 5). Importantly, the control function requires the exact same conditions for the IVs as before. That is, the IVs must be strong and exogenous. Thus, in the linear regression model, there is little reason to use the CF approach. In fact, it is harder to implement the CF approach than the 2SLS approach, as we need to bootstrap to obtain the correct standard errors.

However, the CF approach is often the more straightforward way of correcting for endogeneity when the dependent variable in Eq. 1 is *not continuous*. We may have a dependent variable that is, e.g., binary (e.g., purchased an ice cream or not), a discrete variable (e.g., how many scoops of ice creams were purchased), or a choice variable (e.g., which ice-cream flavor is chosen); see, e.g., Petrin and Train (2010),

Ebbes et al. (2011), and Andrews and Ebbes (2014). In those cases, the model in Eq. 1 would not be a linear regression model, but rather a binomial logit or probit model for a binary dependent variable, a Poisson regression model for a discrete count dependent variable, or a multinomial regression model for a nominal (choice) dependent variable (Andrews and Ebbes 2014; Petrin and Train 2010). We can control for endogeneity in these models by including the control function term $\hat{\nu}_i$ as an additional regressor and bootstrapping the standard errors of the estimated regression coefficients.

Multiple Endogenous Regressors

Many marketing applications have more than one potentially endogenous regressor. Suppose we have two regressors, price and advertising, in a regression model that we suspect are both endogenous because of strategic planning. In addition, we have one other regressor X_i , which we believe is exogenous. That is, let us consider the following extension of Eq. 1:

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 A_i + \beta_3 X_i + \varepsilon_i \quad (11)$$

Here, A_i is the endogenous advertising variable, and X_i is the exogenous regressor. We now need at least two IVs to correct for endogeneity in Eq. 11. In general, when there are K endogenous regressors, we need $L \geq K$ IVs. For each IV, the researcher needs to develop theoretical arguments to argue that (1) the IV is relevant (i.e., is not weak) and (2) the IV meets the exclusion restriction.

Once we have identified appropriate IVs, we could then think of the IV approach as estimating a separate first-stage regression for each endogenous regressor. But, there is an important practical matter: each first-stage regression needs to have the *same set of right-hand-side variables*. That is, all available exogenous information (IVs and exogenous regressors) belong to the right-hand side of each first-stage regression equation, as in the following two first-stage regression equations for price P and advertising A :

$$P_i = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \gamma_3 X_i + \nu_i^P \quad (12)$$

$$A_i = \eta_0 + \eta_1 Z_{1i} + \eta_2 Z_{2i} + \eta_3 X_i + \nu_i^A \quad (13)$$

where ν_i^P and ν_i^A are the error terms of the first-stage regression equations. Another important matter concerns the theoretical development of the IVs: we must ensure that *each* endogenous variable is identified by *at least one unique IV*. That is, we only address the endogeneity problem adequately if at least one IV is related to P_i and the other IV is related to A_i . We cannot have that Z_1 is correlated with both P and A , while Z_2 is correlated with none. We can also not have that Z_1 and Z_2 are both correlated with P and neither is correlated with A .

To test for the strength of the instruments in case of multiple endogenous variables, we can use a multivariate F -test (e.g., the Sanderson-Windmeijer F -test). This test is implemented in Stata's `ivreg2` (Baum et al. 2007, 2015; Sanderson and Windmeijer 2016).

In case we want to address an endogeneity problem with multiple endogenous regressors using the CF approach, the procedure is quite similar as before. Using the first-stage regressions (12) and (13), we can compute the fitted OLS residuals from Eqs. 12 and 13 and include both as two additional regressors in the main Eq. 11. Subsequently, the augmented Eq. 11 may be estimated by OLS. As before, we need to bootstrap to obtain the correct standard errors of the estimated regression coefficients.

Interaction Terms

In many marketing applications, the effect of one independent variable on the dependent variable depends on the level of a second independent variable. In that case, we need an interaction term in the regression equation, i.e., the product of the two independent variables enters the regression as an additional covariate. Consider the following example in which the effect of the endogenous regressor price depends on the level of the exogenous variable X_i :

$$Y_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \beta_3 P_i X_i + \varepsilon_i \quad (14)$$

To use 2SLS, we must treat the interaction as a separate endogenous regressor that needs its own IV(s) and its own first-stage regression equation. As a second IV, we could use the interaction $X_i Z_i$ between the exogenous regressor X_i and Z_i .⁷

$$P_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i Z_i + \gamma_3 X_i + \nu_i^p \quad (15)$$

$$P_i X_i = \eta_0 + \eta_1 Z_i + \eta_2 X_i Z_i + \eta_3 X_i + \nu_i^{px} \quad (16)$$

Here, ν_i^p and ν_i^{px} are the error terms of the two first-stage regression equations. The regression coefficients in Eq. 14 should be estimated with 2SLS. As before, we would need to support the IV analyses with the six tasks discussed above, including a discussion of the first-stage regressions and the strength of the instruments (which now includes the constructed instrumental variable $X_i Z_i$).

A potentially more straightforward and parsimonious way to address an endogeneity problem in an interaction term is through the CF approach (Wooldridge

⁷In case the variables are mean centered before they enter the product, i.e., $(P_i - \bar{P})(X_i - \bar{X})$, we need to use the mean-centered interaction term on the left-hand side of (16) and on the right-hand side of (14), instead of $P_i X_i$. The IVs do not require mean centering as the first-stage predictions will not be affected by mean centering.

2015). It is sufficient to estimate the following first-stage regression and add the fitted residuals as an additional regressor to (14):

$$P_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i + \nu_i^p \quad (17)$$

By including the residuals from Eq. 17 in the main Eq. 14, we can directly address the endogeneity problem, without having to specify a second first-stage equation for the interaction term (Wooldridge 2015, p. 428). However, a potential downside is that the standard errors need to be computed using a bootstrapping approach.

The Benefit of Panel Data

As the discussion above highlights, correcting for potential endogeneity in a market response model is important but not straightforward. In fact, when the IVs are poor (e.g., weak, endogenous, or both), the IV estimator is potentially more biased than the OLS estimator. Hence, the “cure” to the problem can be worse than the “disease” (see also Rossi 2014). Therefore, the researcher should only resort to IV estimation if there is a serious concern of an endogeneity problem and when it cannot be solved by adding other covariates (e.g., Germann et al.’s (2015) rich data approach).

However, another opportunity to correct for endogeneity arises when the researcher has panel data (e.g., Wooldridge 2010; Verbeek 2012). Panel data means multiple observations per response unit across time, such as tracking the sales of a set of stores over time. In some cases, the panel structure of the data can be leveraged to correct for endogeneity. In the literature, these models are often labeled “unobserved effects models” (Wooldridge 2010; Germann et al. 2015). The idea behind such models is that they control for omitted variables by using fixed effects dummy variables as control variables. To illustrate this idea, suppose we have daily data not only for one ice-cream vendor but for multiple vendors. We extend model (1) for panel data as follows:

$$Y_{it} = \beta_0 + \beta_1 P_{it} + \alpha_i + \lambda_t + \varepsilon_{it} \quad (18)$$

Here $i = 1, \dots, N$ indicates ice-cream vendor i , and $t = 1, \dots, T$ indicates day t . In addition, there are two new terms in the model: α_i and λ_t . The first term, α_i , is a term specific to ice-cream vendor i and does not vary over time. This term represents all factors that we cannot observe that are particular to this ice-cream vendor and which do not change during the observation window. As examples we could think of the location of the vendor or the quality of the vendor’s ice cream. If they are not accounted for in model estimation, these unobserved, time-constant effects could lead to an endogeneity problem: if the ice-cream vendors realize the potential of their location, they may be tempted to charge higher prices on premium locations, regardless of which day it is. When this price setting behavior is unobserved to the researcher, P_{it} will correlate with the (composite) error term $u_{it} = \alpha_i + \varepsilon_{it}$ through α_i , and we have an endogeneity problem.

The second term, λ_t , is a time-specific term that affects all ice-cream vendors in the same way. Here, we could think of, say, weather (assuming the same weather applies to all ice-cream vendors), industry-wide changes, economic cycles, government policy, etc. Following similar reasoning, when ice-cream vendors take these “time” shocks into consideration for setting their prices, then we may have an endogeneity problem if the model estimation does not account for these time shocks. Hence, when the factors α_i and λ_t are not explicitly accounted for in the estimation, they will be part of the composite error term $u_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$, and price (or other regressors) may be correlated with u_{it} , leading to biased estimates using standard estimation approaches such as OLS.

Fortunately, the panel structure of the data allows us to eliminate these two unobserved components α_i and λ_t and any endogeneity problem arising from these two components, *without needing IVs*. This is done by leveraging a two-way fixed effects model (Baltagi 2013, p. 39) that uses fixed effects to control for systematic differences between cross-sectional units (e.g., ice-cream vendors) and for factors that are common to all cross-sectional units but vary by time period.

As fixed effects for the cross-sectional units, we could include a set of dummy variables for each ice-cream vendor in the model. However, this can potentially lead to many more parameters to estimate. Instead, we could calculate the average demand and average price across t for each i , by averaging (18) across time, resulting in:

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{P}_i + \alpha_i + \bar{\lambda} + \bar{\varepsilon}_i \quad (19)$$

with $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^T Y_{it}$, and the other averages \bar{P}_i , $\bar{\lambda}$, and $\bar{\varepsilon}_i$ defined similarly. Then, subtracting Eq. 19 from Eq. 18 results in the following regression equation:

$$(Y_{it} - \bar{Y}_i) = \beta_1 (P_{it} - \bar{P}_i) + (\lambda_t - \bar{\lambda}) + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (20)$$

Examining (20), we see that the unobserved time constant effect α_i dropped out of the model, and the concern about it inducing an endogeneity problem is gone. We may estimate (20) with OLS, using $Y_{it} - \bar{Y}_i$ as the “new” dependent variable and $P_{it} - \bar{P}_i$ as the “new” independent variable and including time-fixed effects dummies $\tilde{\lambda}_1, \dots, \tilde{\lambda}_T$ (alternatively, we can include $T - 1$ time-fixed effects dummies and an intercept). This estimator for β_1 is called the “within estimator” because it only uses within-cross-section variation to estimate β_1 .

However, as Germann et al. (2015, p. 4) note, we do need the assumption that prices are uncorrelated with the error term ε_{it} across all time periods (this assumption is sometimes called “strict exogeneity”). Thus, the identifying assumptions underlying the fixed effects model (20) are that (i) the omitted variable(s) is (are) time invariant (i.e., the individual-specific intercept captures the omitted variable(s)) and (ii) there is enough variance in the dependent variable as well as the focal endogenous variable within one specific individual (ice-cream vendor) to allow for the estimation of its effect (the endogenous variable is identified only through the within-individual

variation). The effects of time-invariant independent variables (e.g., quality of the location of the ice-cream vendor) cannot be estimated in the fixed effects approach and are thus eliminated from the fixed effects model during estimation.

How about the time unobserved effects? In estimating Eq. 20 with OLS, we already suggested to include time-fixed effects dummies. These dummies capture any unobserved time-specific effects. Thus, any endogeneity concern arising from unobserved time-specific shocks that are common to all ice-cream vendors are now no longer present.

However, if we have daily data, this would lead to the inclusion of many time dummies in (20), which may complicate the estimation of Eq. 20 with OLS. As an alternative, we could subtract the average across cross section (per time period) from both the dependent and independent variables. That is, we average (18) across cross-sectional units, for each time period $t = 1, 2, \dots, T$, resulting in

$$\bar{Y}_t = \beta_0 + \beta_1 \bar{P}_t + \bar{\alpha} + \lambda_t + \bar{\varepsilon}_t \quad (21)$$

with $\bar{Y}_t = \frac{1}{N} \sum_{i=1}^N Y_{it}$, and the other averages \bar{P}_t , $\bar{\alpha}$, and $\bar{\varepsilon}_t$ defined similarly. In addition, we average (18) across cross section and time, giving

$$\bar{Y} = \beta_0 + \beta_1 \bar{P} + \bar{\alpha} + \bar{\lambda} + \bar{\varepsilon} \quad (22)$$

with $\bar{Y} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N Y_{it}$, and the other averages \bar{P} , $\bar{\alpha}$, $\bar{\lambda}$, and $\bar{\varepsilon}$ defined similarly. Now we subtract Eqs. 19 and 21 from Eq. 18 and add Eq. 22 (Baltagi 2013, p. 40), to obtain

$$(Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}) = \beta_1 (P_{it} - \bar{P}_i - \bar{P}_t + \bar{P}) + (\varepsilon_{it} - \bar{\varepsilon}_i - \bar{\varepsilon}_t + \bar{\varepsilon}) \quad (23)$$

We can now run a standard OLS regression on (23) to estimate β_1 as both α_i and λ_t are dropped out of the model. Here, we would *not* include an intercept and use as dependent variable the “new” variable $Y_{it} - \bar{Y}_i - \bar{Y}_t + \bar{Y}$ and as independent variable the “new” variable $P_{it} - \bar{P}_i - \bar{P}_t + \bar{P}$. There would be neither time dummies nor vendor-specific dummies in this regression. We believe that in many panel data applications in marketing, such an approach can already address most of the endogeneity problems and should routinely be carried out.

We note that another approach commonly discussed in panel applications is the random effects estimator that treats the unobserved intercepts α_i in (18) as random variables. We would like to stress that this approach does *not* account for endogeneity, and it has even slightly stronger exogeneity assumptions regarding the identification of the regression effects in (18) than OLS. Hence, if the researcher believes that there is an endogeneity problem arising from unobserved, time-constant effects, then including random intercepts in the model, as in a random effects regression approach, does *not* address the endogeneity problem (see also Ebbes et al. (2004) for a discussion). A discussion on the main identifying assumptions of panel data model applications in marketing is given by Germann et al. (2015, Table 2).

Conclusions

Many marketing research professionals and consultants are interested in estimating regression models to understand the effect of one variable (e.g., price or a marketing investment) on another (e.g., sales or market share). If the goal is to find the causal effect based on observational data such as transaction data or survey data, endogeneity is a serious challenge in achieving this goal. Endogeneity occurs if the independent variable is set deliberately and strategically by managers in order to capitalize on factors causing shocks in demand that are observable to the manager but not to the researcher. Similarly, consumers may be strategic and make decisions based on factors not observable to the researcher. As a result, these unobserved factor(s) can become part of the error term in the regression model. If one or more such factors are used to set the independent variable, then the variable is endogenous, which means that the estimated effect of the independent variable will be biased in standard estimation approaches such as OLS.

The bias can be substantial in practical applications, and therefore considering endogeneity issues is important both in marketing academia and marketing practice. However, this is more easily said than done. As a first step, we recommend researchers to expand the regression model with covariates that could capture the unobserved factors. In the ice-cream vendor example, once the temperature variable was added to the regression model, the OLS estimates were very close to the true parameter values and were estimated with high precision (low standard errors).

In a panel data setting, unobserved factors may be concentrated in unobserved cross-sectional differences between firms or consumers. In this case, adding fixed effects (dummies for each cross-sectional unit) as additional covariates adequately addresses the endogeneity problem. If the endogeneity problem arises because of time-related demand shocks, adding fixed time effects (time dummies) as additional covariates solves the problem. Equivalently, we can use the within estimators presented in the previous section.

If we have a single time series or a single cross section of observations or if we have a panel data setting and the fixed effects approach does not fully solve the endogeneity problem (e.g., Germann et al. 2015; Ebbes et al. 2004), then we need to split the exogenous variation in the independent variable from the endogenous variation and only use the exogenous variation to estimate the causal effect. This is the essence of the IV (or 2SLS) approach. It is, however, challenging to find IVs that satisfy two seemingly contradictory conditions: they need to be strong (explain the endogenous independent variable) yet be uncorrelated with the error term of the main regression equation. While the first condition (IV strength) can be investigated empirically, the second condition (IV exogeneity) can only be supported with theoretical arguments.

The difficulty of finding suitable IVs has sparked researchers' interest in finding ways of accounting for endogeneity in observational data without the need to use observed instruments. In the marketing literature, two approaches have been

proposed. Firstly, Ebbes et al. (2005) develop the method of latent instrumental variables (LIV) that provides identification through latent, discrete components in the endogenous regressors. Similar to the observed IV approach, the LIV approach shares the underlying idea that the endogenous regressor is a random variable that can be separated into two components, one which represents the exogenous variation and one which represents the endogenous variation. The endogenous component is correlated with the error term of the main regression equation through a bivariate normal distribution. The LIV model may be estimated using, e.g., a maximum likelihood approach. Secondly, Park and Gupta (2012) introduce a method that directly models the correlation between the endogenous regressor and the model error term using Gaussian copulas. This approach can be implemented through a control function approach. Both the LIV and the Gaussian copulas approaches require non-normality in the endogenous regressor and normality of the error term of the main regression equation. We refer to Papies et al. (2017) for a detailed discussion of these two approaches.

Since exogenous variation in the independent variable of interest is essential for the estimation of causal effects, perhaps the best way to estimate causal effects is through *field experiments* (e.g., Ledolter and Swersey 2007; Ascarza et al. 2017). In a field experiment, we randomly set the values of the independent variable(s). After observing the realizations of the dependent variable, we can estimate the causal effect of the independent variable(s) on the dependent variable with a standard OLS regression approach⁸, as the variation in the independent variable is exogenous because of the randomization. Field experiments are suitable in business applications where there are many potential customers, who can be accurately targeted and randomized into two or more treatment groups and whose outcomes or responses can be measured. But even in those settings, the implementation can be a challenge. For instance, a field experiment that attempts to randomly manipulate prices in an online setting can be detected by consumers, who realize that their price changes from one online session to another or notice that their price is different from the price of another consumer who is purchasing the same product. This may lead to a potential backlash, especially in today's world of online connectivity (see, e.g., Verma 2014).

This chapter is deliberately written rather informally, as a way to introduce the problem of endogeneity in regression models to marketing researchers. For a more formal, econometric treatment of endogeneity problems in regression models, we refer to the excellent textbooks by Wooldridge (2010) and Verbeek (2012) and the article by Basile (2008). We hope that this chapter has clarified the issue of endogeneity and has provided some perspective on approaches to address endogeneity.

⁸Ascarza et al. (2017) leverage a field experiment to address endogeneity concerns in the context of a customer relationship management (CRM) campaign, customer targeting, and social influence.

References

- Albers, S., Mantrala, M. K., & Sridhar, S. (2010). Personal selling elasticities: A meta-analysis. *Journal of Marketing Research*, 47(5), 840–853.
- Andrews, R. L., & Ebbes, P. (2014). Properties of instrumental variables estimation in logit-based demand models: Finite sample results. *Journal of Modelling in Management*, 9(3), 261–289.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics. An empiricist's companion*. Princeton: Princeton University Press.
- Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects in CRM campaigns. forthcoming in the *Journal of Marketing Research*.
- Baltagi, B. H. (2013). *Econometric analysis of panel data* (5th ed.). Chichester: Wiley.
- Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, 6(3), 285.
- Baum, C. F., Schaffer, M. E., & Stillman, S. (2007). Enhanced routines for instrumental variables/GMM estimation and testing. *Stata Journal*, 7(4), 465–506.
- Baum, C. F., Schaffer, M. E., & Stillman, S. (2015). *IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation*. Boston College Department of Economics. Retrieved from <https://ideas.repec.org/c/boc/bocode/s425401.html>
- Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The Rand Journal of Economics*, 25(2), 242–262.
- Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R., & Telang, R. (2007). The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science*, 53(9), 1359–1374.
- Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42(2), 141–156.
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430), 443–450.
- Chintagunta, P. C., & Desiraju, R. (2005). Strategic pricing and detailing behavior in international markets. *Marketing Science*, 24(1), 67–80.
- Dinner, I. M., van Heerde, H. J., & Neslin, S. A. (2014). Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. *Journal of Marketing Research (JMR)*, 51(5), 527–545.
- Ebbes, P., Böckenholt, U., & Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, 58(2), 161–178.
- Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing & Economics*, 3(4), 365–392.
- Ebbes, P., Papies, D., & van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science*, 30(6), 1115–1122.
- Franses, P. H. (2005). On the use of econometric models for policy simulation in marketing. *Journal of Marketing Research*, 42, 4–14.
- Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, 97(3), 713–744.
- Germann, F., Ebbes, P., & Grewal, R. (2015). The chief marketing officer matters! *Journal of Marketing*, 79(3), 1–22.
- Greene, W. H. (2011). *Econometric analysis*. Upper Saddle River: Prentice Hall.
- Homburg, C., Kuester, S., & Krohmer, H. (2009). *Marketing management: A contemporary perspective*. London: McGraw-Hill Higher Education.
- Karaca-Mandic, P., & Train, K. (2003). Standard error correction in two-stage estimation with nested samples. *The Econometrics Journal*, 6(2), 401–407. <https://doi.org/10.1111/1368-423X.t01-1-00115>.
- Ledolter, J., & Swersey, A. J. (2007). *Testing 1-2-3 experimental design with applications in marketing and service operations*. Stanford, California: Stanford Business Books.

- Leeflang, P., Wittink, D. R., Wedel, M., & Naert, P. A. (2000). *Building models for marketing decisions*. Dordrecht: Kluwer.
- Leenheer, J., Van Heerde, H. J., Bijmolt, T. H. A., & Smidts, A. (2007). Do loyalty programs really enhance behavioral loyalty? An empirical analysis accounting for self-selecting members. *International Journal of Research in Marketing*, 24, 31–47.
- Liebowitz, S. J. (2016). How much of the decline in sound recording sales is due to file-sharing? *Journal of Cultural Economics*, 40(1), 13–28.
- Lindgren, B. W. (1993). *Statistical theory*. London: Chapman and Hall.
- Manchanda, P., Rossi, P. E., & Chintagunta, P. K. (2004). Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research (JMR)*, 41(4), 467–478.
- Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2), 307–342.
- Papies, D., Ebbes, P., Van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In P. S. H. Leeflang, J. E. Wieringa, T. H. A. Bijmolt, & K. H. Pauwels (Eds.), *Advanced methods for modelling marketing*. Switzerland: Springer. <http://www.springer.com/de/book/9783319534671>
- Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4), 567–586.
- Petrin, A., & Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research (JMR)*, 47(1), 3–13.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioural research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903.
- Roederkerk, R. P., van Heerde, H. J., & Bijmolt, T. H. A. (2013). Optimizing retail assortments. *Marketing Science*, 32(5), 699–715.
- Rossi, P. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science*, 33(5), 655–672.
- Sanderson, E., & Windmeijer, F. (2016). A weak instrument -test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, 190(2), 212–221.
- Sethuraman, R., Tellis, G. J., & Briesch, R. A. (2011). How well does advertising work? Generalizations from meta-analysis of brand advertising elasticities. *Journal of Marketing Research*, 48(3), 457–471.
- Shaver, J. M. (1998). Accounting for endogeneity when assessing strategy performance: Does entry mode choice affect. *Management Science*, 44(4), 571–585.
- Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business Statistics and Economic Statistics*, 20, 518–529.
- Verbeek, M. (2012). *A guide to modern econometrics* (4th ed.). Hoboken: Wiley.
- Verma, I. M. (2014). Editorial expression of concern: Experimental evidence of massive scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(29), 10779–10779.
- Villas-Boas, J. M., & Winer, R. S. (1999). Endogeneity in brand choice models. *Management Science*, 45(10), 1324–1338.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2), 420–445.

Part II
Methods



Cluster Analysis in Marketing Research

Thomas Reutterer and Daniel Dan

Contents

Introduction	222
An Overview of Clustering Methods	223
Data Quality and Proximity Measures	224
Distance-Based Cluster Algorithms	227
Cluster Analysis of Market Basket Data	229
Data Characteristics and Choice of Distance Measure	229
Hierarchical Clustering	232
K-Medoid Clustering	236
K-Centroids Cluster Analysis	240
Conclusions	244
Cross-References	246
References	246

Abstract

Cluster analysis is an exploratory tool for compressing data into a smaller number of groups or representing points. The latter aims at sufficiently summarizing the underlying data structure and as such can serve the analyst for further consideration instead of dealing with the complete data set. Because of this data compression property, cluster analysis remains to be an essential part of the marketing analyst's toolbox in today's data rich business environment. This chapter gives an overview of the various approaches and methods for cluster analysis and links them with the most relevant marketing research contexts. We also provide

T. Reutterer (✉)

Department of Marketing, WU Vienna University of Economics and Business, Vienna, Austria
e-mail: thomas.reutterer@wu.ac.at

D. Dan

Department of New Media, Modul University Vienna, Vienna, Austria
e-mail: daniel.dan@modul.ac.at

pointers to the specific packages and functions for performing cluster analysis using the R ecosystem for statistical computing. A substantial part of this chapter is devoted to the illustration of applying different clustering procedures to a reference data set of shopping basket data. We briefly outline the general approach of the considered techniques, provide a walk-through for the corresponding R code required to perform the analyses, and offer some interpretation of the results.

Keywords

Cluster analysis · Hierarchical clustering · k -centroid clustering · k -medoid clustering · Marketing analysis · Marketing research

Introduction

Cluster analysis is a generic term for exploratory statistical techniques and methods aiming at detecting groupings in data sets that are internally more homogeneous than the entities across the categorized groups. One of the primary goals of clustering is data compression, i.e., to summarize the original entities by a smaller number of groups or representing points instead of considering the complete data set. Cluster analysis has a long history and emerged as a major topic in the 1960s and 1970s under the label “numerical taxonomy” (cf., Sokal and Sneath 1963; Bock 1974). The origins of cluster analysis appeared in disciplines such as biology for deriving taxonomies of species or psychology to study personality traits (Cattell 1943). Over the years, a large variety of clustering techniques has been proposed for numerous types of applications in diverse fields of research. From a historical perspective, excellent books on cluster analysis have been written by Anderberg (1973), Hartigan (1975), Späth (1977), Aldenderfer and Blashfield (1984), Jain and Dubes (1988) or Kaufman and Rousseeuw (1990). Additionally, Arabie and Lawrence (1996) provide an extensive compilation of contributions on various aspects of cluster analysis; for more development updates in the field, see Everitt et al. (2011) or Hennig et al. (2015). From a marketing researcher’s perspective, Punj and Stewart (1983) or Arabie and Lawrence (1994) provide comprehensive reviews of cluster analysis.

The “classical” marketing problems involving the application of clustering methods are market segmentation (Wedel and Kamakura 2000; Dolnicar et al. 2018) and competitive market structure (CMS) analysis (DeSarbo et al. 1993). The former entails deriving segments of customers who either react homogeneously to various marketing mix variables (response-based segmentation) or are more homogeneous with respect to some psychometric constructs such as product attitudes or product images, perceived value, or preferences (construct-based segmentation); see Mazanec and Strasser (2000) or Reutterer (2003) for this distinction and an overview of corresponding clustering methods.

The task of CMS analysis is to derive a configuration of brands in a specific product class which adequately reflects inter-brand competitive relationships as perceived by consumers (DeSarbo et al. 1993). This is typically accomplished via an arrangement of the rivaling brands in ultrametric trees, overlapping or fuzzy cluster structures (Rao and Sabavala 1981; Srivastava et al. 1981, 1984). Because they utilize identical data structures but just differ in the mode of data compression, segmentation (compression of the consumer mode) and CMS (compression of the brand mode) turn out to be “reverse sides of the same analysis” (Grover and Srinivasan 1987; Reutterer 1998). Yet another very similar data structure arises when companies keep record of their customer transactions (e.g., by tracking them over time in customer relationship management systems). Such data sets tend to be huge and accrue as clickstreams of visitation and corresponding purchasing patterns on a website or as sequences of shopping baskets comprising jointly purchased items or product categories. The data compression tasks involved in the so-called exploratory market basket analysis (Mild and Reutterer 2003; Boztuğ and Reutterer 2008; Reutterer et al. 2017) are analogous to those in market segmentation vs. CMS analysis and also entail some suitable clustering method. The marketing literature refers to the task of discovering subgroups of distinguished cross-category interrelationship patterns among jointly purchased items or product categories also as “affinity analysis” (Russell et al. 1999; Manchanda et al. 1999; Russell and Petersen 2000).

The remainder of this chapter is organized as follows: In the next section, we provide a brief overview of the various clustering methods and focus on how to proceed when conducting distance-based clustering in more detail. We also present the most popular distance/proximity measures and algorithms as well as the most commonly used software implementations in the computational environment R available to analysts. To demonstrate the application and the results obtained from using various clustering methods, we then provide a couple of hands-on examples on how to put cluster analysis into action. Using one and the same data set, we demonstrate the specific quality of data compression achieved when utilizing a specific type of cluster analysis. While most textbooks use market segmentation as the standard case for illustrating cluster analysis in marketing (see, e.g., Chapman and McDonnell Feit (2019) and Dolnicar et al. (2018) for excellent examples), we focus in our demonstration on the exploratory analysis of shopping basket data representing customers’ joint purchase decisions across a wide range of product categories.

An Overview of Clustering Methods

We can distinguish between two major groups of clustering methods: model-based clustering and distance-based clustering. While model-based methods explicitly assume some statistical probability model as an underlying data generating process, the latter are more exploratory by nature. The idea behind model-based methods is that the observations arise from a probability distribution which is a mixture of two

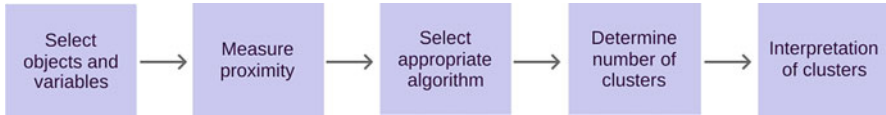


Fig. 1 Steps to conduct cluster analysis

or more components (i.e., clusters). Each of these components is a density function with an associated weight in the mixture (e.g., a mixture of multivariate normal distributions) and the task is to determine the mixture distribution which fits the data best. This is usually done by varying the number of components and optimizing some fit or information criterion. A specific variant of model-based clustering is discussed in the chapter ▶ “[Finite Mixture Models](#)” of this handbook. For more technical details on model-based clustering, see Titterton et al. (1985), McLachlan and Basford (1988), Fraley and Raftery (2002), and Frühwirth-Schnatter (2006).

In the following, we will focus on distance-based clustering. The aim of distance-based clustering is to find groupings in the data such that the distance between entities within a group is minimized, while it is maximized for entities between groups for some predefined distance measure. The steps required to employ a distance-based clustering procedure are shown in Fig. 1. After the objects and variables of interest are selected from an available data base, the second step involves the choice of an appropriate proximity measure to quantify the (dis-)similarity between the objects to be clustered. In the next step, a specific cluster algorithm is selected, and once the results are obtained, the number of clusters is determined and the cluster solution interpreted accordingly. Because of their crucial impact on the resulting cluster solutions, we next discuss some of the most popular proximity measures and clustering algorithms used in research practice.

Data Quality and Proximity Measures

The choice of a proximity measure depends on the nature of the data to be clustered, more specifically, the scaling properties of the variables at hand. Generally speaking, we can distinguish between numerical (quantitative, metric) and categorical (qualitative, nonmetric) data. Metric data is characterized by a scale with numerically equal distances representing values of the underlying characteristic being measured, such as age, income, or the number of units sold over a month. With this kind of data, any mathematical operation can be performed, it can be displayed from the greatest to the least and vice versa. In contrast, categorical data include binary and nominal data with no natural order, for example: product choice, gender, ethnicity, etc. If categorical data imply an order relationship (such as preference rankings) of the measured objects, the scale is denoted as being ordinal. The latter include rating scales (e.g., brand attitudes measured using itemized scales), which occasionally are

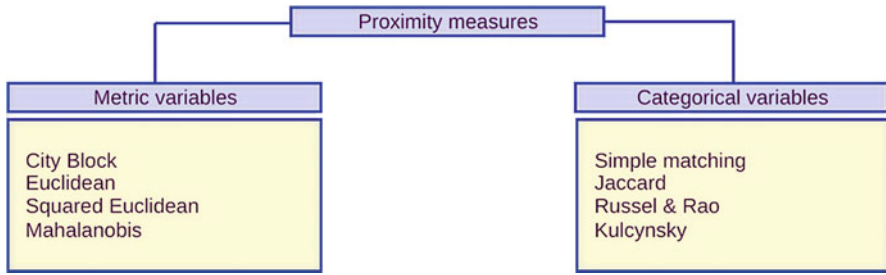


Fig. 2 Common examples of proximity measures

called pseudometric as they are treated as numerical data despite their ordinal properties. For a detailed discussion of measurement issues and data quality see, e.g., Mooi et al. (2018). Given the scaling properties of the data and the type of (dis-)similarity desired, the choice of a proximity measure determines how close/similar or how far/dissimilar objects in a data set are situated. A major distinction of alternative measures arises when we distinguish between metric and nonmetric data, see Fig. 2. In the case of metric data, most of the proximity measures are based on the summed distances of the objects with respect to all variables or dimensions of the data.

In an n -dimensional space, the most well-known and widely used distance measure between two data points $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ arises as a family of metrics denoted as Minkowski distance or L_p norm (Adams and Fournier 2003), i.e., a metric where the distance between two vectors is given by the norm of their difference. The outcome of this metric is given by Eq. 1:

$$d_p(X, Y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \tag{1}$$

The distance is a metric if $p \geq 1$. The most commonly used norms are L_1 known as the Manhattan norm and L_2 known as the Euclidean norm. The distances derived from these norms are called Manhattan distance

$$d_1(X, Y) = \sum_{i=1}^n |x_i - y_i|, \tag{2}$$

and Euclidean distance

$$d_2(X, Y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}, \tag{3}$$

respectively. In cases where more weight should be put on the measurement of very distant data objects, the squared Euclidean distance can be used:

$$d_2^2(X, Y) = \sum_{i=1}^n |x_i - y_i|^2 \tag{4}$$

If the data is nonmetric (i.e., nominal, binary, or ordinal scales), the most common way of quantifying the (dis)similarity between data points is based on a two-way cross-classification of objects which counts for having a binary attribute: present or absent (note that ordinal data can be transformed into a series of binary variables accordingly). The corresponding similarity coefficients mentioned in Fig. 2 mainly differ in their assumptions on whether the common absence of a characteristic reflects similarity (such as the simple matching coefficient) or not and how much weight they put on the matched presence of an attribute.

Since we will use the Jaccard (dis)similarity coefficient in the following applications of cluster analysis using shopping basket data, we briefly illustrate the construction of the Jaccard index. The latter is used to assess the similarity s between two sets A, B or categories c_A, c_B . Formally, it measures the size ratio of their intersection $c_A \cap c_B$ divided by their union $c_A \cup c_B$ and can be written as follows:

$$s_{c_{AB}} = \frac{c_A \cap c_B}{c_A \cup c_B} = \frac{c_A \cap c_B}{c_A + c_B - c_A \cap c_B} \tag{5}$$

As discussed below in more detail, in the context of market-basket analysis the analyst’s interest is in quantifying the (dis)similarity of products or product categories depending on whether they are jointly purchased in a set of transactions or not. In doing so, the product purchases are represented as binary elements with (1) denoting presence and (0) denoting absence of the specific product in a shopping basket. By cross-classifying a pair of products, we can calculate the Jaccard coefficient with the help of the following contingency table (cf. Sneath 1957; Kaufman and Rousseeuw 1990; Leisch 2006).

		Product 1		
		1	0	sum
Product 2	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	p

For a set of p shopping baskets, the Jaccard similarity coefficient (also often referred to as the Tanimoto similarity coefficient (Anderberg 1973)) for products 1 and 2 can be calculated as:

$$s_{\text{prod1,prod2}} = \frac{a}{a + b + c}, \tag{6}$$

and the corresponding dissimilarity coefficient is:

$$d_{\text{prod1,prod2}} = \frac{b + c}{a + b + c}, \tag{7}$$

with the elements in the contingency table representing:

- a , the number of transactions with purchases of both product 1 and 2
- b , the number of incidences of product 2 but no product 1 purchases
- c , the number of incidences of product 1 but not product 2 purchases
- d , the number of transactions with neither product 1 nor product 2 purchases

Note that $s_{prod1, prod2} = 1 - d_{prod1, prod2}$ and in practice d in the above contingency table is usually the cell with the (by far) highest counts. This particularly applies to the context of shopping basket analysis but is not limited to this case. In such situations, any proximity measure that treats co-incidences of common zeros (in our case: nonpurchases of two specific products or categories) the same way as common ones would be biased towards the absence of two characteristics. For example, this is the case for the simple matching coefficient (which is $s_{prod1, prod2} = \frac{(a+d)}{p}$) or the Hamming distance (i.e., the number of different bits $d_{prod1, prod2} = b + c$). Thus, in many scenarios it makes sense to use asymmetric proximity measures like the abovementioned Jaccard coefficient which gives more weight to common ones than to common zeros.

Distance-Based Cluster Algorithms

Regarding the choice of a cluster algorithm (step three in the above by Fig. 1), one popular way to distinguish variations of distance-based clustering methods is to

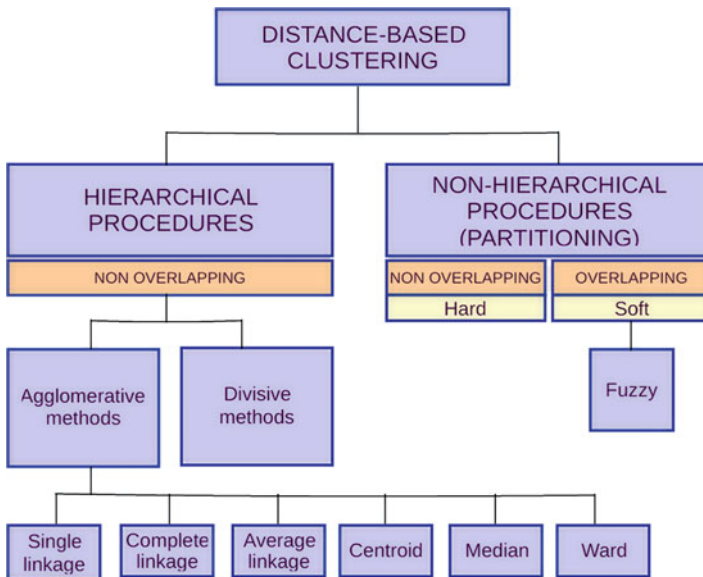


Fig. 3 Overview of distance-based clustering algorithms

divide them into hierarchical and nonhierarchical procedures. As illustrated by Fig. 3, the former can be split into agglomerative and divisive, while the later can be branched into nonoverlapping, overlapping, or fuzzy clustering methods (Hruschka 1986; Wedel and Kamakura 2000).

While in the case of nonoverlapping clustering each entity is assigned to one single group, overlapping clustering techniques allow for the simultaneous membership of objects to multiple groups. For example, depending on the consumption, brands and/or consumers might belong to more than one cluster of products or to several segments, respectively (see, e.g., Arabie et al. 1981). Fuzzy clustering abandons the idea of a “hard” partitioning of the data and replaces fixed cluster assignments by a degree of membership assigned to each entity and cluster (Hruschka 1986). Note that despite similar in idea but different in conception and interpretation, the notion of a “soft” group membership assignment becomes also apparent in model-based clustering methods when probabilities of cluster memberships are estimated and thus each data point can be assigned to more than one cluster. In this case, the inclusion of a data point in multiple clusters is due to a probabilistic approach, not of a distance.

Nonoverlapping clustering approaches can be further classified into hierarchical and nonhierarchical methods (Punj and Stewart 1983). Techniques for hierarchical clustering either start out with all entities in a single cluster (divisive algorithms or top down) or with each entity in its own cluster (agglomerative algorithms or bottom up). The latter approach is more popular among marketing researchers and successively links pairs of clusters (or still isolated entities) from a previous stage based on their shortest mutual distance. The agglomeration schedule stops when all entities are combined into one single cluster. We illustrate the application of some common hierarchical clustering procedures below in section “[Hierarchical Clustering](#).”

Nonhierarchical clustering starts with a (typically randomly initialized) grouping of the data for a prespecified number k of clusters and aims to gradually improve the partition by optimizing a “minimum variance criterion,” i.e., by minimizing the inner (within-group) dispersion of the k -partition (cf. Bock 1974; Strasser 2000). The k -means algorithm was first proposed by MacQueen (1967) and its many variations (see, e.g., Jain and Dubes 1988; Kaufman and Rousseeuw 1990) are popular examples for such nonhierarchical distance-based clustering procedures. There is a huge variety of clustering procedures available in packages and functions provided by the R (R Core Team 2019) ecosystem for statistical computing. The most commonly used packages are the following:

- `stats`: The base R package provides a number of implementations for both partitioning and hierarchical clustering techniques. Function `kmeans()` comprises several algorithms for computing Euclidean distance-based partitions, while `hclust()` provides agglomerative hierarchical clustering algorithms. The `stats` package also provides various auxiliary functions like `dendrogram()` for visualizing cluster hierarchical solutions.
- `cluster`: This package provides R implementations of methods introduced in Kaufman and Rousseeuw (1990) and comprises a number or both partitioning (`pam()`, `clara()`, and `fanny()`) and hierarchical cluster algorithms (`agnes()`,

`diana()`, and `mona()`). The package also contains many extensions of these base methods and visualization functions (Struyf et al. 1996).

- `mclust`: A set of model-based clustering methods for fitting Gaussian finite mixture models using an expectation maximization (EM) algorithm is provided by the `mclust` package. It also provides numerous functions to assist cluster validation and evaluating the number of mixture components using the Bayesian Information Criterion (BIC) (Fraley and Raftery 2003).
- `flexclust`: This package provides an environment for partitioning cluster analysis with non-Euclidean distance measures using k -centroids cluster algorithms (KCCA) (Leisch 2006). There are also functions for deriving neighborhood graphs and image plots for visualization of partitions.

A comprehensive list of R packages for performing model-based or distance-based clustering is maintained by Friedrich Leisch and Bettina Grün and made available via the following CRAN task view: <https://CRAN.R-project.org/view=Cluster>.

Cluster Analysis of Market Basket Data

We next illustrate three different clustering procedures applied to a reference data set of shopping basket data. In doing so, we briefly outline the data used and the general approach of the selected procedure, then provide a walk-through for the corresponding R code required to perform the analyses and give some interpretation of the results. The three clustering procedures under consideration are: hierarchical clustering using the function `hclust()` and two prototype generating clustering methods: function `pam()` from the `cluster` package and function `kcca()` from the `flexclust` package.

Data Characteristics and Choice of Distance Measure

To illustrate the clustering methods, we use one month (30-days) of real-world point-of-sale transaction data from a local grocery outlet. The data set is included in the widely used R package `arules` and consists of an easy-to-handle set of 9835 retail transactions representing purchases in 169 different categories. The data come as a sparse matrix with each observation (row) representing a retail transaction and each column a binary variable with 1 denoting that a specific grocery category is present in the transaction and 0 otherwise. Thus, the row-wise sums indicate the number of categories purchased together in each transaction. A typical transaction is expressed as a list of categories such as: {tropical fruit, yogurt, coffee}, {citrus fruit, semifinished bread, margarine, ready soups}, and so on.

To obtain the data, in the R console, we add the `arules` package and activate the `Groceries` data set included in the library. To get a first visual impression of the proposed data set, we plot a histogram for the basket sizes, see Fig. 4, which

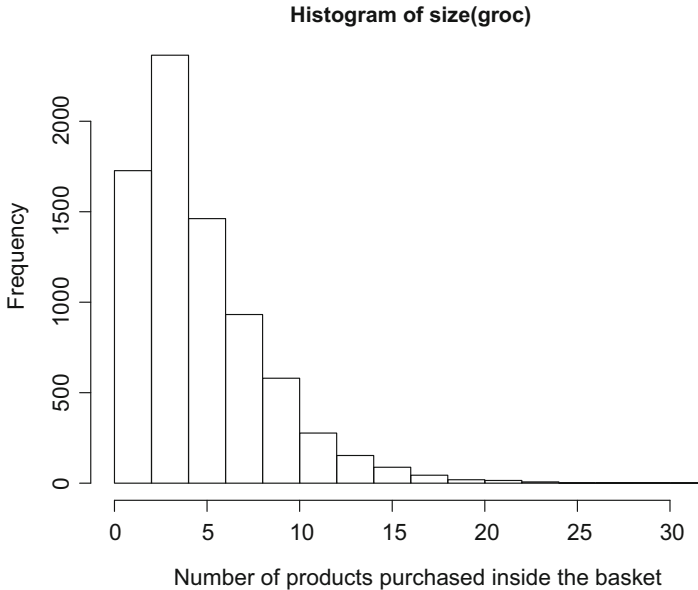


Fig. 4 Histogram of basket sizes

shows a right-skewed pattern we typically observe in supermarket transaction data: the majority of baskets are composed of only a few categories. Thus, the binary basket data are extremely sparse, with a mean basket size `mean(size(Groceries))` of only 4.41 categories per shopping trip. Note that throughout the chapter, we will omit the category “shopping bags” because the latter does not reflect any specific consumption preference but in a grocery shopping context merely serves to carry the bought items around. The `arules` library, the `Groceries` data and the `groc` variable indicated below will be the same throughout all the further examples and will be reported only once in the R code examples.

To get a better understanding about which categories are purchased most frequently, we can plot the frequency distribution of categories exceeding a threshold (support) of 5%. As we can see from Fig. 5, the most frequently purchased categories are typical grocery products such as whole milk, other vegetables, rolls/buns, soda, yogurt, etc.

```
library("arules")
data("Groceries")
groc <- Groceries[size(Groceries)>1,
which(itemLabels(Groceries) != "shopping bags")]
hist(size(groc), xlab =
"Number of products purchased inside the basket")
```

With the `itemFrequencyPlot` function, we plot the most frequent items.

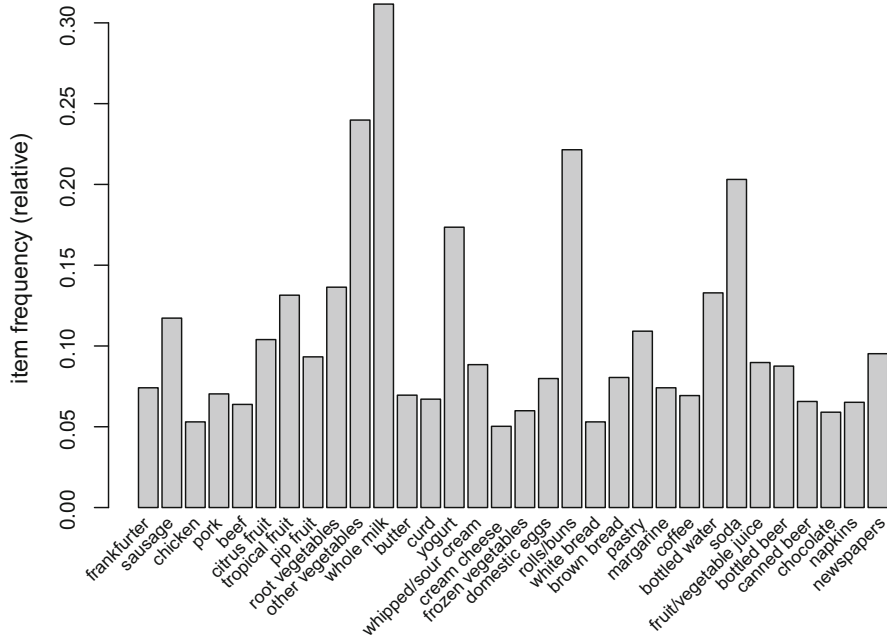


Fig. 5 Histogram of the most frequent items purchased

```
itemFrequencyPlot(groc, support = 0.05)
```

As we discussed in the previous section, in order to perform distance-based clustering, we need to specify a proximity measure which quantifies the distance between the objects to be clustered. Two aspects need to be considered in our present application. Firstly, an appropriate distance measure needs to adequately account for the data sparsity we observe for our market basket data at hand. Secondly, from a more substantive perspective, we are typically interested in finding groupings in the data which reflect jointly purchased categories, i.e., we aim at detecting complementary cross-purchase incidences. Thus, an asymmetric distance measure giving more weight to joint purchases than to common zeros (i.e., nonpurchases) is preferred in such situations. The previously discussed Jaccard coefficient (cf. Kaufman and Rousseeuw 1990) has such properties and is used in the present application.

For a given data set of market baskets $X^T = [x_n]$, $n = 1, \dots, N$ containing binary purchase incidences $x_n \in \{1, 0\}^J$ we can compute a frequency matrix $X^T \times X = C = [c_{ij}]$ of pairwise co-purchases of categories $i, j = 1, \dots, J$ and derive the Jaccard distance as follows (cf. Sneath 1957):

$$d_{ij} = 1 - \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}}, \forall i, j = 1, \dots, J \quad (8)$$

Note that in the present context, the corresponding Jaccard similarity $s_{ij} = 1 - dij$ measures the percentage of joint purchases in all baskets which contained at least one of the two categories.

Hierarchical Clustering

In our first example, we employ hierarchical clustering to the described set of shopping basket data. In such a setting, the task is to explore subgroups of jointly purchased product categories based on pairwise co-purchase “affinities” across the categories included in the data set. Thus, we aim at detecting clusters of product categories which tend to be purchased together more often by the customers of the local supermarket.

As already mentioned in section “[Distance-Based Cluster Algorithms](#),” the most popular hierarchical clustering method is agglomerative clustering, which can be performed by using the function `hclust()`. In this family of clustering methods, the agglomeration procedure initially considers each singleton object (here: product category) as a cluster and then, step by step, merges the objects iteratively into groups of clusters until one final cluster is generated. The merging mechanism is directed by (i) the proximity or (dis)similarity measure and (ii) a linking criterion.

In our example, we use the Jaccard distance as defined above. As a starting point, we thus compute a dissimilarity matrix $D = [d_{ij}]$ according to Eq. 8:

```
diss <- dissimilarity(groc[, itemFrequency(groc) > 0.02], method =
"jaccard", which = "items")
```

Note that for simplicity reasons and to keep the resulting tree structure easy to inspect, we only select categories which are contained in more than 2% of the retail transactions at hand. This is done by specifying `itemFrequency(Groceries) > 0.02`.

The choice of a linking criterion determines how the distance between two groups of observations is calculated during the agglomeration procedure. It uses the previously computed (dis)similarity measure and one of the many available methods for measuring the distance between two clusters or between a cluster and a singleton object. The most popular linkage methods are represented in Fig. 6. While single linkage uses the distance between the two closest elements of two clusters, complete linkage measures the distance between the two farthest or most distant elements in two sets. The average linkage criterion compromises between the two previously mentioned ones and takes the mean distance between the elements of each cluster. Ward’s method aims at minimizing the within-cluster variance and at each step merges the pair of clusters that leads to a minimum increase in total within-cluster variance after merging (cf. Kaufman and Rousseeuw 1990).

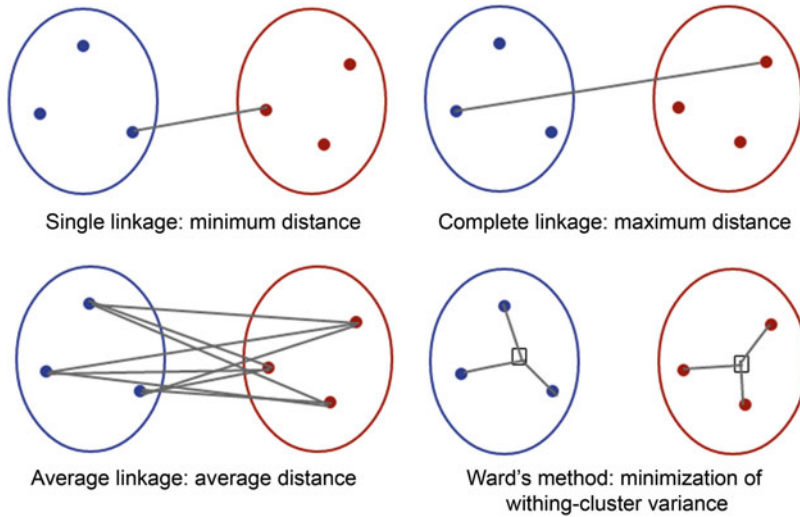


Fig. 6 A selection of popular linkage criteria in hierarchical clustering

Note that the choice of a cluster linkage method has a decisive impact on the resulting cluster solution. To illustrate this, consider the following two cases using single linkage clustering and the minimum variance method proposed by Ward (1963). Using the function `hclust()`, single linkage clustering can be performed as follows:

```
# Single linkage method
hc.single <- hclust(diss, method = "single")
plot(hc.single, cex=0.7)
abline(h = c(0.75, 0.80, 0.85, 0.90, 0.95),
col = "gray", lty = 3)
```

One common way to visualize the outcome of hierarchical clustering is by using a so-called dendrogram. The word dendrogram comes from the combination of two ancient Greek words: *déndron* (“tree”) and *grámma* (“written character, letter, that which is drawn.”) The analogy with a reversed tree is obvious, each leaf represents one object (in our case a product category), each branch represents one cluster at a certain point of the agglomeration process, and the root encompasses all the clusters. Notice that at a certain “height” of the cluster dendrogram, the branches are merged together. The height of the fusion, also known as the *cophenetic distance* (Farris 1969), is inversely proportional to the similarity of the objects. A common way to assess how the generated dendrogram reflects the data is to compare the cophenetic distances with the original distances by correlating them.

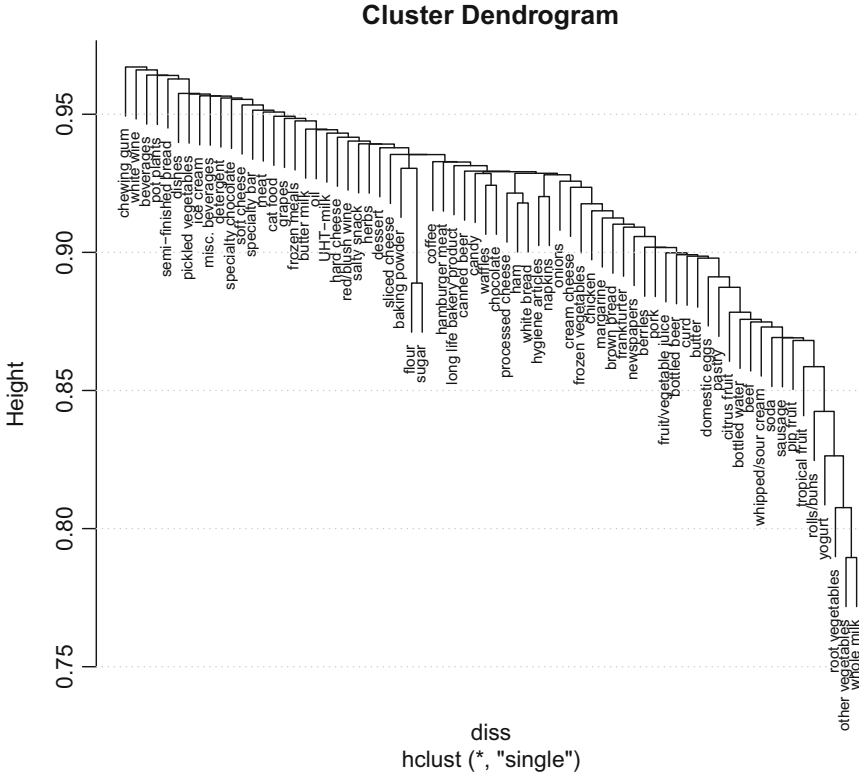


Fig. 7 Hierarchical agglomerative clustering applying single linkage method

A strong correlation indicates a good linking of the objects in the dendrogram (Saraçlı et al. 2013).

The dendrogram derived for our market basket data is obtained by the `plot()` command for object `hc.simple`. It is given in Fig. 7 and depicts a typical property inherent to single linkage clustering, namely that very “similar” categories (i.e., those which are purchased together very often, here whole milk, other vegetables, followed by root vegetables, yogurt, etc.) are merged at a very early stage of the agglomeration process and those categories which rarely appear in the same shopping baskets (chewing gum, beverages, etc.) at a later stage.

The single linkage criterion employs a “nearest neighbor” rule to merge sets and thus is able to reveal rather complex, elongated, or snake-shaped data structures (Kaufman and Rousseeuw 1990; Dolnicar et al. 2018). On the other hand, single linkage typically induces a chaining effect in the hierarchical agglomeration procedure, which can be clearly seen in Fig. 7 in the creation of long straggling “clusters.” This is due to the fact that objects are added sequentially to clusters, and at each stage, the “closest distant” (or most similar) object is merged with the already

existing configuration. Because of this property, single linkage cluster is also sometimes used for outlier detection (i.e., those entities merged with the configuration towards the end of the agglomeration process).

In contrast, Ward’s linking method aims at forming minimum inner variance partitions. To achieve this, at each step, the pair of clusters which minimizes the incremental increase in within-cluster variance is merged. Ward’s method can be called in the R code by specifying the `ward.D` option. As we see from the dendrogram in Fig. 8, using this method, the categories are merged more evenly from the start, and thus we have a more balanced distribution of clusters.

```
# Ward linkage method
hc.wardd <- hclust(diss, method = "ward.D")
plot(hc.wardd, cex=0.7)
rect.hclust(hc.wardd, k = 5, border = 2:7)
```

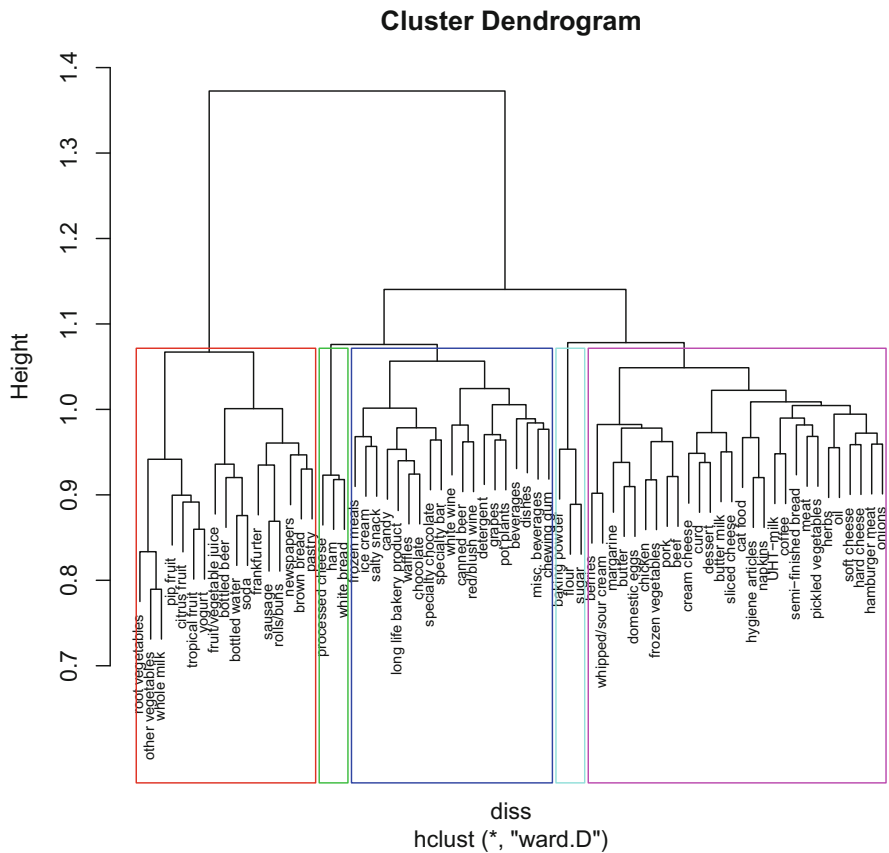


Fig. 8 Hierarchical agglomerative clustering applying Ward’s method

As we move up from the leaves to the root, we notice that the branches get linked together at a variable height. The height of the linkage, indicated on the vertical axis, measures the (dis)similarity between two objects or clusters. The more we move towards the root of the dendrogram, the more dissimilar the merged objects are. In our illustrative example, we can see that several branches merge frequently co-purchased categories from the assortment of fresh products (e.g., vegetables, whole milk, fruits, yogurt) at the left-hand side of the tree, which are later linked with drinks (e.g., bottled beer, juices, water, soda, etc.) and a combination of categories associated with snacks (e.g., frankfurter, sausage, rolls, etc.) into one cluster marked by a red box. In the other (right-hand side) branch from the root, other distinct category combinations (such as salty and sweet snacks, candies and chocolate categories, etc.) are represented.

Notice that at any horizontal “cut” of the tree structure a specific cluster solution consisting of groups of product categories with internally more intense cross-category purchase relationships emerge. For example, in Fig. 8, we marked a solution with $k = 5$ clusters: three larger clusters and two smaller, each representing only three categories which are purchased together very frequently (i.e., {processed cheese, ham, white bread} and {baking powder, flour, sugar}). Such formations of “supra-categories” can be helpful for store managers to design shelf placements of categories within the store but also to consider the representation of categories in leaflets or promotional activities.

K-Medoid Clustering

Hierarchical clustering is a useful tool for data compression and visualization if the number of objects used for clustering is reasonably small. For example, this is the case if we are interested in analyzing the joint purchase affinities among the product categories J represented in a matrix X of shopping baskets as just illustrated above. However, as J increases, the dimensionality of the to-be-derived distance-matrix reaches computational limits and/or visualization of the dendrogram for the linkage procedure becomes intractable and interpretation cumbersome. The latter also applies to the task of deriving a segmentation of the shopping baskets or the households behind the observed retail transactions. From a substantive perspective, such a focus also implies moving away from studying category purchase interdependencies for a pooled set of transaction data. When we compress the shopping baskets into a smaller number of representing basket classes, we aim at finding a partition of the data set at hand with outstanding or more distinguished complementary cross-category purchase incidences within the detected classes (for more details, see Boztuğ and Reutterer 2008; Reutterer et al. 2006, 2017).

For such tasks, nonhierarchical or partitioning clustering is a feasible alternative. Formally, the task is to find a partition $P = \{P_1, \dots, P_K\}$ of the data set into a fixed number of K basket classes which fulfills the following objective function:

$$\sum_k \sum_{n \in P_k} d(x_n, c(x_n)) \rightarrow \min_{P, C} \quad (9)$$

where $C = (c_1, \dots, c_K)$ is a set of centroids or prototypes and $d(\cdot)$ a distance measure, such as the Jaccard distance in Eq. 8, we are using in the present application. In the clustering and classification literature, the “minimum dispersion criterion” in Eq. 9 is also known as the principal point or k -centroids problem (Jain and Dubes 1988; Leisch 2006). One important property of resolving Eq. 9 is that for any optimum configuration (P^*, C^*) , the condition $c^*(x_n) = \arg \min \{d(x_n, c_k), \forall k\}$ holds, which warrants that each basket x_n is mapped onto its minimum distant or closest centroid. With the notable exception of Ward’s method (which follows a similar objective function), this is in sharp contrast to the way most linkage procedures proceed in forming clusters. Instead of minimizing a global objective function, agglomerative hierarchical clustering aims at minimizing a distance function at each step of the cluster fusion but can result in a potentially suboptimal global solution.

Before we illustrate using a generic method for solving the k -centroids problem in the next section, we first employ an iterative, easy-to-implement, relocation-based heuristic proposed by Kaufman and Rousseeuw (1990) under the name Partitioning Around Medoids (PAM). Combined with clustering objective function (Eq. 9), this algorithm requires from the centroid to have the property $c_k \in \{x_n\}_{n \in P_k} \forall k$, i.e., the “medoid” is defined as the shopping basket which minimizes the mean distance with all other transactions in the same cluster P_k . This medoid property guarantees that the centroids are real shopping baskets, which tend to result in more robust cluster solutions in the presence of outliers and facilitates interpretation. On the other hand, PAM is suitable for relatively small- to medium-sized data sets, but this problem can be overcome by selecting randomly from the available data or following other resampling methodologies.

For clustering larger data sets using the medoid approach, one may use CLARA (Clustering LARGE Applications; see Kaufman and Rousseeuw 1990) or CLARANS (Clustering Large Applications based upon RANdomized Search; see Ng and Han 2002). The former does not use the entire data, but it randomly chooses multiple samples with fixed size and repeatedly applies PAM to each of these samples and selects the representative k -medoids. Afterwards, the objects in the data set are assigned to the closest medoid. CLARA finds the best clustering if the sampled medoid is among the best k -medoids by calculating the mean of the dissimilarities of the data to their closest medoid. CLARANS interprets the search space as a hypergraph, where each node represents a set of k -medoids. The algorithm randomly chooses a set of neighbor nodes as new medoids in an iterative manner. If the neighbor discovered is better than the previous one, a local optimum is discovered. The whole process is repeated until the whole graph is sufficiently explored and an optimal solution is found.

We apply the k -medoid partitioning to the Groceries data set by taking into account only transactions that contain at least two different product categories.

After this preselection, we are left with 7,676 transactions and 168 categories. To retain a dissimilarity matrix of moderate size, we randomly select 2,000 transactions and use the Jaccard coefficient as distance measure. The cluster solutions for a sequence of $K = 1, \dots, 8$ clusters are generated using the function `pam()`. Setting a seed value to secure reproducibility of the obtained results completes the following code for performing k -medoid clustering of the available shopping basket data:

```
library("cluster")
set.seed(42)
samp <- sample(groc, 2000)
diss <- dissimilarity(samp, method = "Jaccard")
clust <- lapply(1:8, function (x) pam(diss, k= x))
```

Determining a suitable number of clusters based on distance-based clustering methods is an open issue and the relevant literature offers a huge variety of “validity” metrics to assist the analyst with this task. Popular metrics include the cluster separation measure proposed by Davies and Bouldin (1979) or indices based on the agreement of repeated cluster solutions like the measures proposed by Rand (1971). An overview and detailed performance comparisons of alternative metrics for determining the number of clusters is provided by Milligan and Cooper (1985) or Dimitriadou et al. (2002). Among these heuristics is also the easy-to-use silhouette coefficient proposed by Rousseeuw (1987) which takes into consideration the discrepancies of the average within-cluster dissimilarities and the nearest data points of each neighboring cluster. Based on this heuristic, we opt for a solution of $K=5$ clusters (see also the discussion in Reutterer et al. 2007 or Reutterer et al. 2017).

As discussed before, the derived clusters should reflect classes of shopping baskets with more distinguished complementary cross-category purchase incidences within the detected basket classes. To explore these particular patterns, we select two exemplary clusters, namely cluster number 2 and 5, and characterize their specific properties using the function `itemFrequencyPlot()`. The resulting plots exhibited in Figs. 9 and 10 represent the relative purchase frequencies across categories in the complete data set as continuous lines and contrast them with the respective cluster-specific distributions. Note that for space and illustrative reasons, we include only categories which are present in at least 5% of transactions.

```
itemFrequencyPlot(samp[clust[[5]]$clustering == 2],
population = groc, support = 0.05)
```

```
set.seed(42)
inspect(samp[clust[[5]]$medoids[2]])
##      items
##      [1] {citrus fruit,
##          tropical fruit,
##          root vegetables,
```

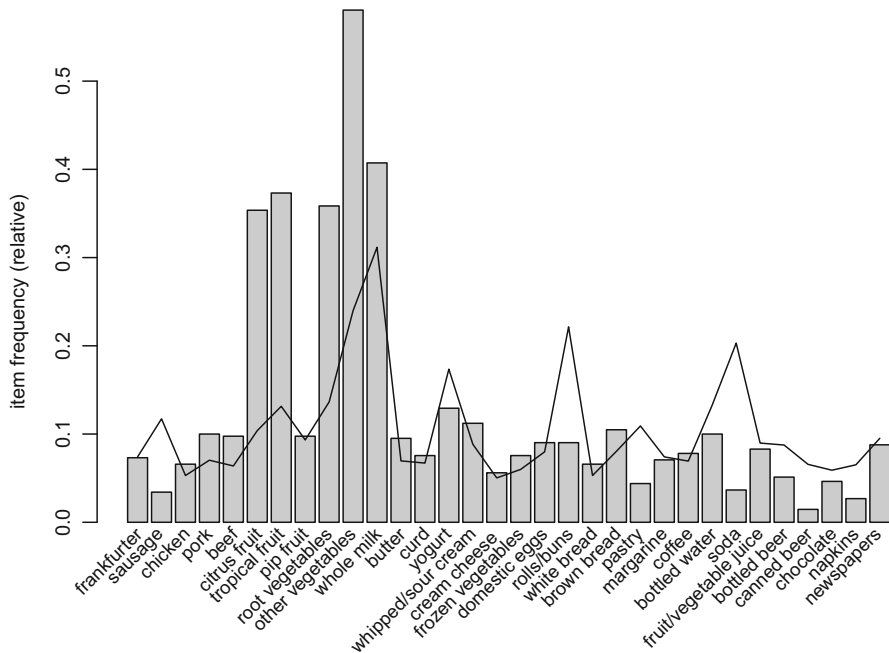


Fig. 9 Profile of relative purchase frequencies across categories in cluster 2

```
##          other vegetables,
##          whole milk}
```

Comparing the two clusters, they clearly point to considerable differences between the shopping baskets summarized by them. The transactions represented by cluster 2 are characterized by a shopping pattern with elevated purchase likelihood in fruits (citrus fruit, tropical fruit) and vegetables (root vegetables, other vegetables) categories as well as whole milk. In contrast, the purchase behavior behind cluster 5 transactions is clearly dominated by remarkably high purchase incidences in certain beverage categories (bottled water, soda, bottled beer) and only moderate class-conditional choice probabilities in the remaining categories.

The `inspect()` function returns us the respective medoid shopping baskets for these two clusters which confirm the above interpretation (i.e., {citrus fruit, tropical fruit, root vegetables, other vegetables, whole milk} for cluster 2 and {bottled water, soda, bottled beer} for cluster 5).

```
itemFrequencyPlot(samp[clust[[5]]$clustering == 5],
population = groc, support = 0.05)
```

```
set.seed(42)
```

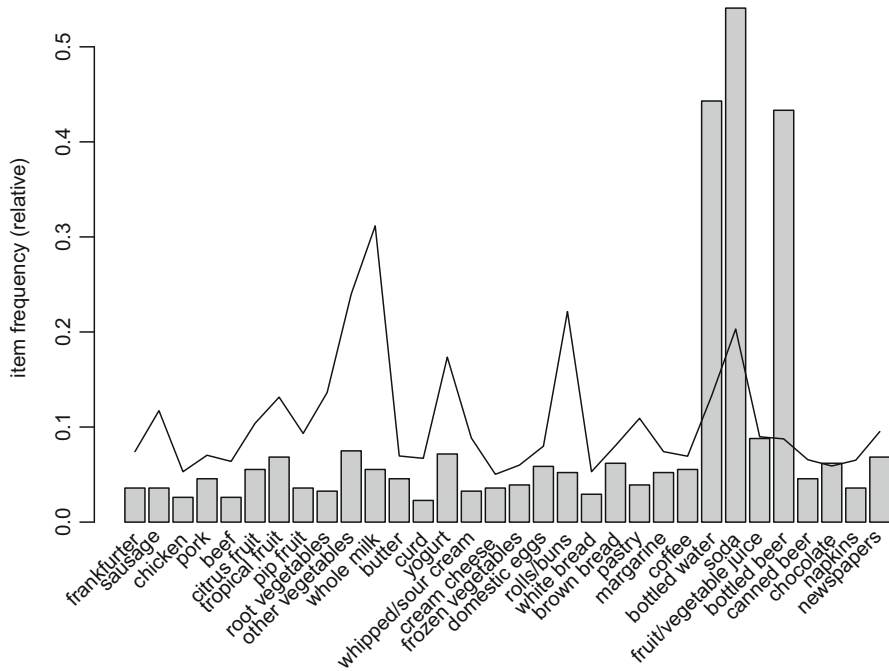


Fig. 10 Profile of relative purchase frequencies across categories in cluster 5

```
inspect(samp[clust[[5]]$medoids[5]])
##      items
##      [1] {bottled water,soda,bottled beer}
```

K-Centroids Cluster Analysis

A more flexible approach to solve the k -centroids problem introduced in the previous section is offered by the function `kcca()` in package `flexclust` (Leisch 2006). The R package `flexclust` includes a multitude of functions for various cluster algorithms. The main function in this package is `kcca()` which implements generalizations of k -means clustering for arbitrary distance measures. Thus, it can be used as a unifying partitioning framework for finding canonical centroids in both metric and nonmetric spaces including binary data (Hartigan and Wong 1979). Like the previous PAM application, once applied to the data structure at hand and using the same distance measure (i.e., Jaccard), this method also derives clusters of retail transactions with internally more homogeneous and pronounced cross-category dependencies.

To illustrate this approach, we use again the `Groceries` data set from the `arules` library and apply the same preselection for categories (threshold of being present in more than 2% of all transactions) as we did in the above application of

hierarchical clustering. In order to be able to perform the necessary computations, we transform the remaining data set into a matrix `grc` with logical values structure and then just add a zero to transform the result into a usable numerical values container.

The `flexclust` family provides, among others, the Jaccard coefficient option as a distance measure for binary data and the `kcca()` function returns a set of real-valued centroids representing class conditional expectations which are directly accessible for the interpretation of the derived cluster solution. The issue of choosing a suitable number of clusters follows the same line of arguments as discussed in the previous subsection and is usually addressed by generating cluster solutions for a sequence of increasing K numbers of clusters and applying some internal “validity” metrics or by systematically studying the stability of alternative solutions using ensemble clustering techniques (Hornik 2005). Here, to exemplify and simplify our illustration, we use five clusters like in the above illustration of the PAM method (`num_clusters = 5`).

One way to inspect the separation of the derived clusters is by visualizing a lower-dimensional representation of the cluster solution. This can be accomplished by applying a data projection method, such as principal component analysis to the data set at hand. We note that principal components or factor analysis is problematic for non-Gaussian (here binary) data, but following Leisch (2006) we consider using it as appropriate for the mere purpose as a simple and easy-to-use data projection device used to visualize a cluster solution (we are not interested in the underlying interpretation of the derived dimensions). Other appropriate methods would be, for example, correspondence or homogeneity analysis.

Combining the results of `kcca()` and `prcomp()`, the projection of the data points together with indicators of their cluster membership on the first two dimensions can be done by using the `plot()` function. In Fig. 11, the centroids of the five clusters solution are plotted as numbers and connected by a neighborhood graph, which thickness represents the degree of connectedness. Even though the projected clusters apparently overlap, the scatter plot suggests an underlying structure of five diagonally separated groups of data points.

```
library("ggplot2")
library("flexclust")
Gr <- groc[, itemFrequency(groc) > 0.02]
grc <- as(Gr, "matrix")
grc <- grc + 0

# flexclustControl object holds the "hyperparameters"

fxc <- new("flexclustControl")
lc <- list(iter.max=500, tol=0.001, verbose=0)
fxc <- as(lc, "flexclustControl")
fc_seed <- 100
num_clusters <- 5
```

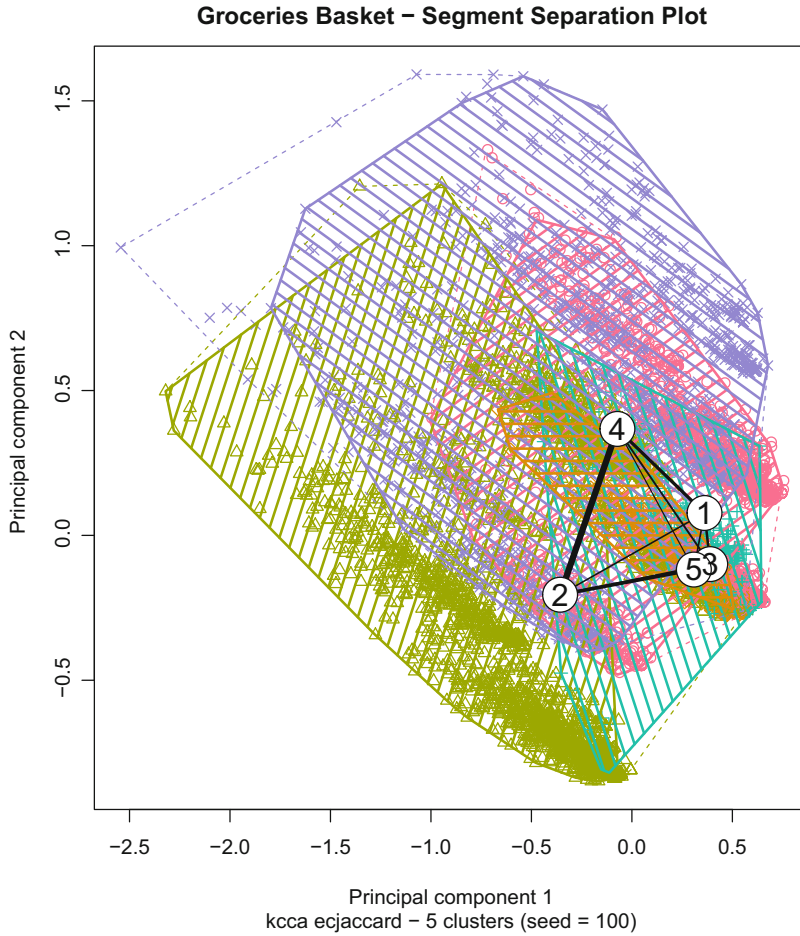


Fig. 11 Neighborhood plot of the Groceries data using the Jaccard distance on the two principal components

```
set.seed(fc_seed)

# verbose > 0 will show iterations
vol.cl <- kcca(grc, k = num_clusters, save.data = TRUE,
control = fxc, family = kccaFamily("ejaccard"))
main_text1 <- "Groceries Basket"
sub_text <- "kcca ejaccard - 5 clusters (seed = 100)"

# plot on first two principal components
vol.pca <- prcomp(grc)
plot(vol.cl, data = grc, project = vol.pca, which = 1:2,
```

```

main = paste0(main_text1, " - Segment Separation Plot"),
xlab = "Principal component 1",
ylab = "Principal component 2",
points = TRUE, hull.args = list(density=10),
sub = sub_text)

barchart(vol.cl, strip.prefix = "# ",
shade = TRUE, layout = c (vol.cl@k, 1),
main = paste0(main_text1, " - Cluster Profile Plot"),
which = hc.wardd$order)

```

Similar to the cluster-specific barplots for the relative category purchase frequencies of the k -medoid cluster solution, the `barchart()` function allows for a graphical representation of the cluster solution contained in a `kcca()` object and thus helps interpreting the findings (see Dolnicar et al. 2014). The argument `shade = TRUE` detects and displays the marker variables in colors. The argument `which` defines the order of the variables. In our case, we chose the order of the product categories such that it corresponds to the order in the dendrogram represented in Fig. 8.

Figure 12 shows in the header the absolute number (i.e., the number of shopping baskets assigned to the respective cluster) and the percentage size of each cluster. The individual barplots help to compare the overall against the cluster-specific centres or mean values per category, which in the present context can be interpreted as the respective percentage of transactions containing a specific category. The line with the full dot represents the relative purchase frequencies over the complete sample. Thus, as we already know from above, the categories with highest shares in the shopping baskets are whole milk, other vegetables, rolls/buns, etc. The bars represent the respective within-cluster purchase shares for each category. They are colored if the difference to the overall mean exceeds a certain threshold value, the bar has a gray contour if this difference is not relevant for interpretation but might be a relevant characteristic of the cluster.

From the visual inspection of Fig. 12, it becomes obvious that the five derived basket classes differ in their basket composition from the overall “average” shopping basket by only a few categories, which makes them distinctive from each other. For example, the shopping baskets represented by cluster 4 are characterized by an outstanding share of rolls/buns and clearly above average purchase incidences in the sausage and frankfurter categories (the three categories together representing typical items demanded for making snacks). A slightly different variation of this cluster is represented by segment 3, in which drinks are represented by the bottled beer and food by sausages and bread. In contrast, the 36% of shopping baskets represented by cluster 2 contain typical grocery shopping categories, such as whole milk, other vegetables, root vegetables, butter, and domestic eggs above average.

From a managerial perspective, knowledge of such behavioral segments is an important prerequisite for designing customized target marketing actions (see Reutterer et al. 2006). For example, categories with distinguished purchase propensities within a specific segment (such as beer in segment 3 or water and soda

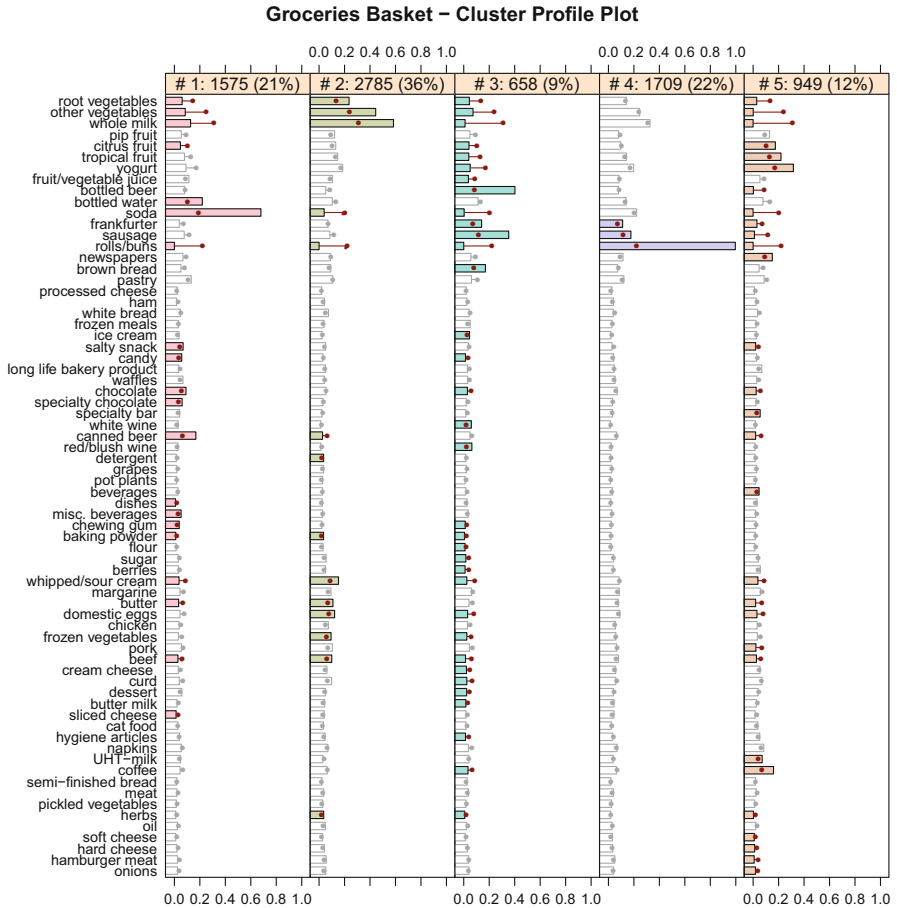


Fig. 12 Segment profile plot for the five-segment solution based on the *K*-centroids algorithm

in segment 1) are promising candidates for deriving targeted promotions to increase store traffic. On the other hand, direct marketers could also aim at promoting categories which are systematically underrepresented in certain segments by cross-promoting them in combination with certain “draw” categories; for example, in segment 3, customers could be stimulated to purchase more vegetables or milk by cross-promoting bottled beer in a way similar to “get a bottle of beer free if you purchase one liter of milk” (Breugelmans et al. 2010; Dréze and Hoch 1998).

Conclusions

As demonstrated in the previous section, the cluster solutions and corresponding interpretations vary considerably depending on the applied clustering procedure. This clearly reflects the exploratory nature of cluster analysis which implies that

there is no single “correct” or natural solution for a specific data set at hand. The achieved data compression effect and thus the specific data structure uncovered by a particular approach to cluster analysis rather depends on a number of factors under the analyst’s control. The main factors are the choice of the cluster analysis procedure, the distance measure, the number of clusters, and the data mode to be clustered.

While there is some more or less sophisticated methodological guidance available to decide on some of these factors, it merely depends on the specific research objective for others. For example, there is extensive research on the determination of the number of clusters (Milligan and Cooper 1985; Dimitriadou et al. 2002) or on the stability properties of cluster solutions (Hornik 2004; Hornik 2005). However, the choice of an appropriate clustering procedure or the specific focus on a particular data mode to be compressed is a decision that is determined by the substantive research question or the analysts subjective judgement. Generally speaking, hierarchical linkage methods have their merits when the task is to explore differences and commonalities among objects on a more fine-granular level (e.g., by “zooming in” the representing dendrogram) and the number of clusters is not fixed a priori. On the other hand, partitioning methods like k -centroid clustering tend to be the preferred method when the analyst aims at compressing larger data sets into a smaller number of representing points (centroids or prototypes), each characterizing a subset of the data as accurately as possible and simultaneously are distinctive from the other cluster centroids.

Using a widely used benchmark set of market basket data, we demonstrated in this chapter that the analytical task of exploring the specific structures of cross-category purchase relationships can be achieved by reducing the dimensionality of the data set using hierarchical clustering (i.e., by analyzing the (dis)similarity structure among the variables of the data matrix). We have also shown that the derived structural patterns strongly depend on the specific method applied and they would also vary if we chose a different distance metric. On the other hand, compressing the number of baskets using nonhierarchical partitioning methods results in a set of specific classes of shopping baskets with distinguished complementary cross-category purchase incidences within the classes. The latter effect is obtained by choosing an appropriate distance measure (in our case Jaccard distances) and the partitioning as well as the interpretation of the classes which would be different for other distance metrics. All these examples demonstrate the generic idea behind cluster analysis as an exploratory data compression tool. This “idea” is to uncover structure in the data, which in the case of distance-based clustering is based on a specific conceptual understanding of quantifying proximity between data points.

The field of cluster analysis is a very dynamic one and the analysts’ toolbox is constantly growing. New methods which aim to cope with the specific challenges of today’s data-rich environments are emerging. Such challenges are not limited to but include real-time (online) updating of cluster solutions for data streams (Ghesmoune et al. 2016), clustering of very high dimensional data sets (Strehl and Ghosh 2003), bootstrap aggregated clustering (Dolnicar and Leisch 2003), and other ensemble methods to improve the quality and robustness of cluster solutions (Hornik 2004;

Hornik 2005); see also the extension package `clue` for R (R Core Team 2019) which provides a computational environment for cluster ensembles.

Modern clustering methods also comprise a variety of unsupervised machine learning methods (for an overview, see Hastie et al. 2009). Marketing applications of such machine learning methodologies include the employment of vector quantization techniques (e.g., Decker 2005; Reutterer et al. 2006), neural networks (e.g., Hruschka and Natter 1986; Mazanec 1999; Reutterer and Natter 2000), topic models for “soft-clustering” unstructured texts (e.g., Tirunillai and Tellis 2014; Büschken and Allenby 2016), or graph partitioning methods (Netzer et al. 2012).

Cross-References

- ▶ [Finite Mixture Models](#)
- ▶ [Market Segmentation](#)

References

- Adams, R. A., & Fournier, J. J. (2003). *Sobolev spaces* (Pure and applied mathematics) (Vol. 140). Amsterdam: Elsevier.
- Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills: Sage.
- Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic.
- Arabie, P., & Lawrence, J. H. (1994). Cluster analysis in marketing research. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Cambridge, MA: Blackwell.
- Arabie, P., & Lawrence, J. H. (1996). *An overview of combinatorial data analysis. Clustering and classification* (pp. 5–63). Singapore: World Scientific.
- Arabie, P., Carroll, J. D., DeSarbo, W., & Wind, J. (1981). Overlapping clustering: A new method for product positioning. *Journal of Marketing Research*, 28(3), 310–317.
- Bock, H. H. (1974). *Automatische Klassifikation*. Göttingen: Vandenhoeck & Ruprecht.
- Boztuğ, Y., & Reutterer, T. (2008). A combined approach for segment-specific market basket analysis. *European Journal of Operational Research*, 187(1), 294–312.
- Breugelmans, E., Boztuğ, Y., & Reutterer, T. (2010). A multistep approach to derive targeted category promotions. Working paper series of the Marketing Science Institute, MSI report no. 10-118, Cambridge, MA.
- Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6), 953–975.
- Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, 38(4), 476–506.
- Chapman, C., & McDonnell Feit, E. (2019). *Segmentation: Clustering and classification. R for marketing research and analytics* (pp. 299–338). New York: Springer.
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2), 224–227.
- Decker, R. (2005). Market basket analysis by means of a growing neural network. *The International Review of Retail, Distribution and Consumer Research*, 15(2), 151–169.
- DeSarbo, W. S., Ajay, K. M., & Lalita, A. M. (1993). Non-spatial tree models for the assessment of competitive market structure: An integrated review of the marketing and psychometric literature. In J. Eliashberg & G. L. Lilien (Eds.), *Handbooks in operations research and management science* (Vol. 5, pp. 193–257). Amsterdam: Elsevier.

- Dimitriadou, E., Dolničar, S., & Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1), 137–159.
- Dolnicar, S., & Leisch, F. (2003). Winter tourist segments in Austria: Identifying stable vacation styles using bagged clustering techniques. *Journal of Travel Research*, 41(3), 281–292.
- Dolnicar, S., Grün, B., Leisch, F., & Schmidt, K. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research*, 53(3), 296–306.
- Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market segmentation analysis. Understanding it, doing it, and making it useful*. Singapore: Springer.
- Drèze, X., & Hoch, S. J. (1998). Exploiting the installed base using cross-merchandising and category destination programs. *International Journal of Research in Marketing*, 15(5), 459–471.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis: Wiley series in probability and statistics*. New York: Wiley.
- Farris, J. S. (1969). On the cophenetic correlation coefficient. *Systematic Zoology*, 18(3), 279–285.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611–631.
- Fraley, C., & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *Journal of Classification*, 20(2), 263–286.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer Science & Business Media.
- Ghesmoune, M., Lebbah, M., & Azzag, H. (2016). State-of-the-art on clustering data streams. *Big Data Analytics*, 1(13), 1–27.
- Grover, R., & Srinivasan, V. (1987). A simultaneous approach to market segmentation and market structuring. *Journal of Marketing Research*, 24, 139–153.
- Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 28(1), 100–108.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485–585). New York: Springer.
- Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of cluster analysis*. Boca Raton/London/New York: CRC Press.
- Hornik, K. (2004). Cluster ensembles. In C. Weihs, W. Gaul (Eds.), *Classification – The ubiquitous challenge. Proceedings of the 28th annual conference of the Gesellschaft für Klassifikation E.V* (pp. 65–72). Heidelberg: University of Dortmund/Springer.
- Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software*, 14(12), 1–25.
- Hruschka, H. (1986). Market definition and segmentation using fuzzy clustering methods. *International Journal of Research in Marketing*, 3(2), 117–134.
- Hruschka, H., & Natter, M. (1986). Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *European Journal of Operational Research*, 114(2), 346–353.
- Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Upper Saddle River: Prentice-Hall.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Hoboken: Wiley.
- Leisch, F. (2006). A toolbox for k-centroids cluster analysis. *Computational Statistics & Data Analysis*, 51(2), 526–544.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281–297.
- Manchanda, P., Ansari, A., & Gupta, S. (1999). The “shopping basket”: A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2), 95–114.
- Mazanec, J. A. (1999). Simultaneous positioning and segmentation analysis with topologically ordered feature maps: A tour operator example. *Journal of Retailing and Customer Services*, 6(4), 219–235.

- Mazanec, J. A., & Strasser, H. (2000). *A nonparametric approach to perceptions-based market segmentation: Foundations* (Vol. 1). Wien: Springer.
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.
- Mild, A., & Reutterer, T. (2003). An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. *Journal of Retailing and Consumer Services*, 10(3), 123–133.
- Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159–179.
- Mooi, E., Sarstedt, M., & Mooi-Reci, I. (2018). Data. In *Market research* (pp. 27–50). Singapore: Springer.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543.
- Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5), 1003–1016.
- Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research*, 20(2), 134–148.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna: R Development Core Team.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.
- Rao, V. R., & Sabavala, D. J. (1981). Inference of hierarchical choice processes from panel data. *Journal of Consumer Research*, 8(1), 85–96.
- Reutterer, T. (1998). Competitive market structure and segmentation analysis with self-organizing feature maps. In P. Anderson (Ed.), *Proceedings of the 27th EMAC conference. Track 5: Marketing research* (pp. 85–105). Stockholm: EMAC.
- Reutterer, T. (2003). Bestandsaufnahme und aktuelle Entwicklungen bei der Segmentierungsanalyse von Produktmärkten. *Journal für Betriebswirtschaft*, 53(2), 52–74.
- Reutterer, T., & Natter, M. (2000). Segmentation-based competitive analysis with MULTICLUS and topology representing networks. *Computers & Operations Research*, 27(11–12), 1227–1247.
- Reutterer, T., Mild, A., Natter, M., & Taudes, A. (2006). A dynamic segmentation approach for targeting and customizing direct marketing campaigns. *Journal of Interactive Marketing*, 20(3–4), 43–57.
- Reutterer, T., Hahsler, M., & Hornik, K. (2007). Data mining und marketing am beispiel der explorativen warenkorbanalyse. *Marketing ZFP*, 29(3), 163–180.
- Reutterer, T., Hornik, K., March, N., & Gruber, K. (2017). A data mining framework for targeted category promotions. *Journal of Business Economics*, 87(3), 337–358.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76(3), 367–392.
- Russell, G. J., Ratneshwar, S., Schocker, A. D., Bell, D., Bodapat, A., Degeratu, A., Hildebrandt, L., Kim, N., Ramaswami, S., & Shankar, V. H. (1999). Multiple-category decision-making: Review and synthesis. *Marketing Letters*, 10(3), 319–332.
- Saraçlı, S., Doğan, N., & Doğan, I. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013(1), 203.
- Sneath, P. H. (1957). Some thoughts on bacterial classification. *Journal of General Microbiology*, 17, 184–200.
- Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy* (A series of books in biology). San Francisco: W.H. Freeman.
- Späth, H. (1977). *Cluster-analyse – Algorithmen zur Objektklassifizierung und Datenreduktion* (2nd ed.). München/Wien: Oldenbourg Wissenschaftsverlag.

- Srivastava, R. K., Leone, R. P., & Shocker, A. D. (1981). Market structure analysis: Hierarchical clustering of products based on substitution-in-use. *Journal of Marketing*, 45(3), 38–48.
- Srivastava, R. K., Alpert, M. I., & Shocker, A. D. (1984). A customer-oriented approach for determining market structures. *Journal of Marketing*, 48(2), 32–45.
- Strasser, H. (2000). Reduction of complexity. In J. Mazanec & H. Strasser (Eds.), *A nonparametric approach to perceptions-based market segmentation: Foundations* (pp. 99–140). Wien/New York: Springer.
- Strehl, A., & Ghosh, J. (2003). Relationship-based clustering and visualization for high-dimensional data mining. *INFORMS Journal on Computing*, 15(2), 208–230.
- Struyf, A., Hubert, M., & Rousseeuw, P. (1996). Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4), 1.
- Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using Latent Dirichlet allocation. *Journal of Marketing Research*, 51(4), 463–479.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley.
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation – Conceptual and methodological foundations*. New York: Springer.



Finite Mixture Models

Sonja Gensler

Contents

Introduction	252
Basic Idea of Finite Mixture Models	253
Illustrative Example	253
Finite Mixture Model and Likelihood Function	254
Probability to Observe a Specific Value of the Segmentation Variable and Mixed Density Function	256
Assignment of Consumers/Objects to Segments Within Finite Mixture Models	257
Determining the Number of Segments	258
Popular Applications of Finite Mixture Models in Multivariate Methods of Analysis	260
Conclusion	261
References	263

Abstract

Finite Mixture models are a state-of-the-art technique of segmentation. Next to segmenting consumers or objects based on multiple different variables, Finite Mixture models can be used in conjunction with multivariate methods of analysis. Unlike approaches combining multivariate methods of analysis and cluster analysis, which require a two-step approach, the parameters are then directly estimated at the segment level. This also allows for inferential statistical analysis. This book chapter explains the basic idea of Finite Mixture models and describes some popular applications of Finite Mixture models in market research.

S. Gensler (✉)
Marketing Center Münster – Institute for Value-based Marketing, University of Münster, Münster,
Germany
e-mail: s.gensler@uni-muenster.de

Keywords

Finite Mixture models · Latent class analysis · Segmentation · Maximum likelihood estimation · Multivariate methods

Introduction

Finite Mixture models are segmentation approaches (in this article, we make no distinction between Finite Mixture and Latent Class models). Segmentation considers heterogeneity among consumers/objects and is crucial for developing marketing strategies. For example, marketing managers may want to know whether there are groups of consumers who exhibit similar shopping behaviors or share particular preferences for product features. Such knowledge offers opportunities to target specific groups of consumers and to develop targeted products and services.

Cluster analysis has traditionally been used to identify groups of consumers who are similar with respect to some specified variables (e.g., shopping behavior, attitude, preferences) – either on its own or in combination with multivariate methods of analysis in a two-step procedure. An example for the latter is the use of a conjoint study to elicit consumers' preferences and the subsequent implementation of a cluster analysis with the estimated preferences as segmentation variables (Green and Krieger 1991).

In recent years, Finite Mixture models have gained popularity as alternative approaches for segmentation. What are advantages of Finite Mixture models compared to traditional clustering approaches? A Finite Mixture model is a *model-based approach*. This means that a statistical model is assumed for the population from which the data stems from. Specifically, it is postulated that a *mixture* of underlying probability distributions generates the data. The assumption of an underlying statistical model has important consequences. Finite Mixture models aim to recover the actual observations in the dataset, while traditional cluster approaches just intent to find homogenous groups of consumers/objects that are distinct from each other (heterogeneity across groups). Thus, *goodness-of-fit measures* for Finite Mixture models are available and support the confidence in the obtained solution. Moreover, *rigorous statistical criteria* help the researcher to identify the most appropriate segment structure in the market. In contrast, researchers use rather arbitrary criteria (e.g., dendrogram) to decide on the number of segments when using traditional cluster analyses (Magidson and Vermunt 2002). Another advantage of Finite Mixture models is that they reduce the experiment-wise error. If traditional cluster approaches are combined with multivariate methods of analysis, a two-step approach is implemented (see example above). Such two-step processes inflate experiment-wise error since there are two different objective functions that are optimized. Finite Mixture models allow for formulating a model that incorporates the identification of segments in the original analysis. That means, instead of separately conducting two different types of analysis and optimizing two different objective functions, one

objective function is formulated (*one-step approach*). Finally, Finite Mixture models are flexible in the sense that variables measured at different scales can be considered. These advantages have contributed to the increasing popularity of Finite Mixture models in market research.

It is the aim of this book chapter to illustrate the basic idea of Finite Mixture models and to discuss how Finite Mixture models can be combined with different multivariate methods of analysis. Finally, the book chapter refers to some specific applications in academic literature.

Basic Idea of Finite Mixture Models

Illustrative Example

A simple example should help to illustrate the basic idea of Finite Mixture models. We observe the purchase frequency of chocolate bars for 450 consumers (see Fig. 1; example adapted from Dillon and Kumar 1994).

Before considering the Finite Mixture model approach, we have a look at the solution derived from traditional cluster analysis.

In this example, the observed purchase frequency serves as the segmentation variable. Traditionally, the researcher would start with running a hierarchical cluster analysis to determine the number of segments. When we use Ward's algorithm, we conclude that there are three segments in the data. We use this information to conduct a K-means clustering, and we find that the three segments have purchase frequencies of 0.74, 4.64, and 12.50, respectively. With K-means clustering each consumer is assigned to one specific segment based on his/her purchase frequency. In this example, the relative segment sizes are 45%, 41%, and 14%. Figure 2 illustrates the result graphically. Obviously, the K-means solution does not represent the observed purchase frequencies well. One reason is the deterministic assignment of consumers to segments.

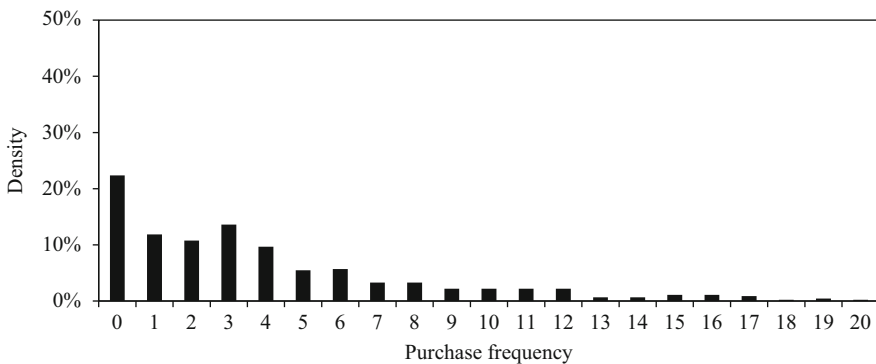


Fig. 1 Density function of the observed purchase frequencies of chocolate bars

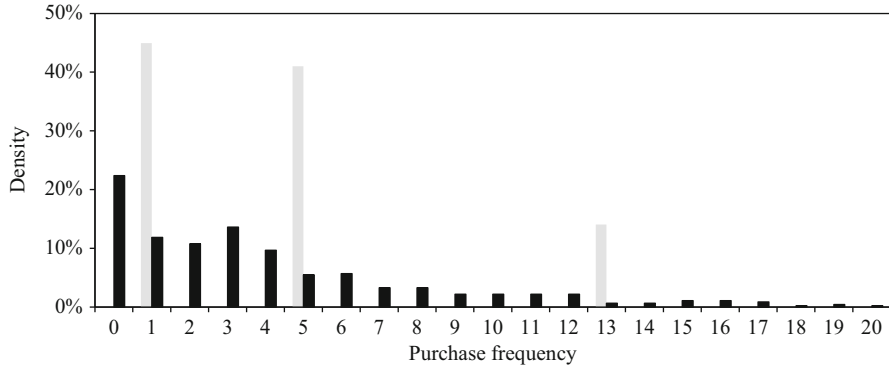


Fig. 2 Density function of the observed purchase frequencies and K-means result for three segments (chocolate bar example)

Finite Mixture Model and Likelihood Function

In the following, the basic idea of the Finite Mixture approach is illustrated. The Finite Mixture approach is a model-based approach and assumes that any observation of one or more variables of interest (i.e., segmentation variable(s)) stems from an underlying density function. In this example, we use again the purchase frequency for the segmentation. Purchase frequency is a count variable, and, thus, the Poisson distribution serves as the underlying density function. The observed density function (see Fig. 1) results from two or more segment-specific density functions that are *mixed*. The segment-specific density functions stem from the same distribution and only differ in their characteristic parameters. That means, we observe multiple Poisson distributions that differ in their means. The objective is now to *unmix* (separate) the density functions of the observations to identify the segment-specific density functions (Green et al. 1976).

The density function of the Poisson distribution is:

$$g(y|\mu) = \frac{\mu^y}{y!} \cdot \exp(-\mu) \tag{1}$$

where

$g()$: density function

y : value of the variable of interest (here: observed purchase frequency)

μ : mean value of the Poisson distribution

In order to identify the segment-specific density functions of the Finite Mixture model, we formulate a likelihood function:

$$L = \prod_{h \in H} \prod_{s \in S} \hat{\eta}_s^{\lambda_{h,s}} \hat{g}_{h|s}(y_h | \hat{\mu}_s)^{\lambda_{h,s}} \tag{2}$$

where

$\hat{\eta}_s$: estimated relative size of segment s

$\lambda_{h,s}$: indicator variable for segment membership of consumer h to segment s

$$\lambda_{h,s} = \begin{cases} 1, & \text{if consumer } h \text{ belongs to segment } s, \\ 0 & \text{otherwise.} \end{cases}$$

$\hat{g}_{h|s}(\cdot)$: estimated conditional density function of consumer h if this consumer is a member of segment s

y_h : value of the variable of interest for consumer h (here: purchase frequency of consumer h)

$\hat{\mu}_s$: estimated mean of Poisson distribution for segment s

H : index set of consumers

S : index set of segments

The likelihood function represents the mixture of distributions, and the relative segment size serves as a weighting variable. Moreover, the likelihood function considers an indicator variable that indicates to which segment a consumer belongs. The likelihood function is maximized using iterative optimization algorithms such as the Newton-Raphson algorithm or expectation-maximization (EM) algorithm (Wedel and Kamakura 2000; Wedel and DeSarbo 1994).

The different algorithms require starting values. In this example, we need to set starting values for the segment-specific means of the Poisson distributions and the relative segment sizes. This implies that we also have to specify the number of segments we want to consider. Since we do not know how many segments represent our data, we estimate models with different numbers of segments. To define the starting values for the means of the Poisson distribution and the relative segment sizes, we can use the result of the K-means clustering. However, since the likelihood function is multimodal in nature, we might find a local and not the global maximum of the likelihood function. To circumvent this issue, one should use different starting values.

Table 1 shows the result for a three-segment solution from the maximization of the likelihood function (2), and Fig. 3 shows the segment-specific density functions. (An Excel spreadsheet for the “chocolate bars” example using the Newton-Raphson algorithm can be request from the author).

The estimated means of the segment-specific Poisson distributions and the relative segment sizes differ from the K-means results. A comparison of the

Table 1 Mean values of the segment-specific Poisson distributions and their relative sizes in the “chocolate bars” example

	Mean value of Poisson distribution	Relative size (%)
Segment 1	0.3	27.7
Segment 2	3.5	54.3
Segment 3	11.2	18.0

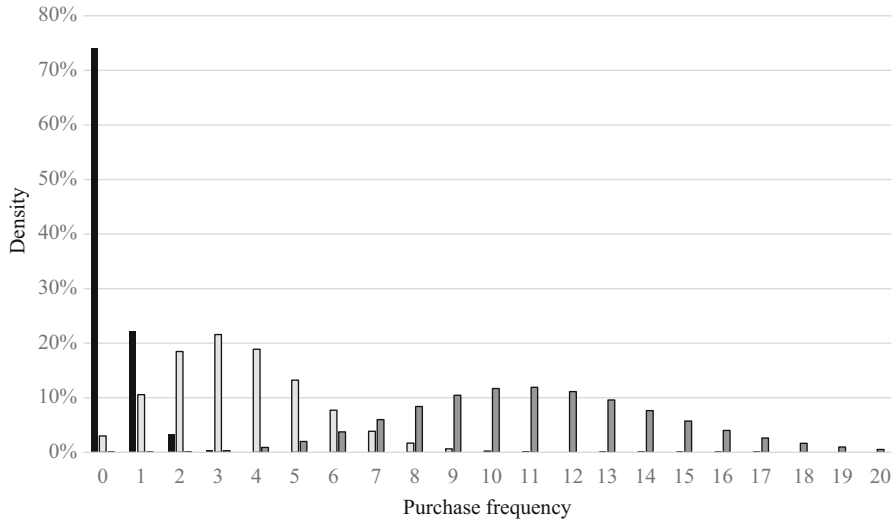


Fig. 3 Segment-specific density functions for the three-segment solution (chocolate bar example)

In-likelihood values (natural logarithm of the value of Eq. 2) for each solution allows assessing which set of estimated parameters reflects the observed purchase frequencies better. The K-means solution leads to an In-likelihood of $-1,147.70$ for Eq. 2, while the Finite Mixture solution results in an In-likelihood of $-1,132.04$. Thus, the Finite Mixture solution fits the observed data better.

Probability to Observe a Specific Value of the Segmentation Variable and Mixed Density Function

The segment-specific (conditional) density functions weighted by the estimated relative segment sizes allow deriving the probability to observe a certain value of the purchase frequency. For any individual consumer in the above example, the *unconditional* individual probability to observe his/her purchase frequency is as follows:

$$\begin{aligned}
 \hat{g}_h(y_h|\hat{\mu}) &= \sum_{s \in S} \hat{\eta}_s \cdot \hat{g}_{h|s}(y_h|\hat{\mu}_s) \\
 &= 0.277 \cdot \frac{0.3^{y_h}}{y_h!} \cdot \exp(-0.3) + 0.543 \cdot \frac{3.5^{y_h}}{y_h!} \cdot \exp(-3.5) \\
 &\quad + 0.18 \cdot \frac{11.2^{y_h}}{y_h!} \cdot \exp(-11.2) \quad \forall h \in H
 \end{aligned}
 \tag{3}$$

For an individual consumer who buys two chocolate bars, this yields, for example, a probability of 0.11. That means, the probability to observe a purchase frequency of

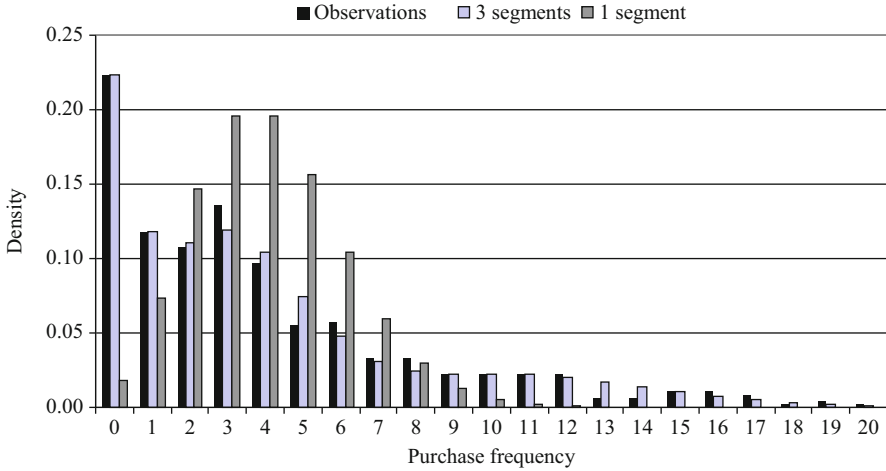


Fig. 4 Density function of the observed purchase frequencies and mixed density functions with one and three segments (chocolate bar example)

two chocolate bars equals 11%. This way, we can compute the probability for every observed value of the purchase frequency. We can use these probabilities to construct the mixed density function (Fig. 4). Figure 4 illustrates that the three-segment solution provides a very good fit with the observed purchase frequencies. Figure 3 also shows the density function for the one-segment solution. In this case, the estimated mean of the Poisson distribution equals 3.99. The one-segment solution does not capture the heterogeneity in purchase frequencies adequately.

Assignment of Consumers/Objects to Segments Within Finite Mixture Models

In contrast to traditional clustering approaches, Finite Mixture models assign consumers to a segment with a certain probability (probabilistic assignment). That means each consumer has a certain probability to belong to a specific segment. This probability is determined based on the estimated relative segment sizes and means of the Poisson distribution. Specifically, the a posteriori probability of segment membership equals

$$\omega_{h,s} = \frac{\hat{\eta}_s \hat{g}_{h|s}(y_h | \hat{\mu}_s)}{\sum_{s \in S} \hat{\eta}_s \hat{g}_{h|s}(y_h | \hat{\mu}_s)} \quad \forall h \in H, s \in S \tag{4}$$

where

$\omega_{h,s}$: probability that consumer h belongs to segment s

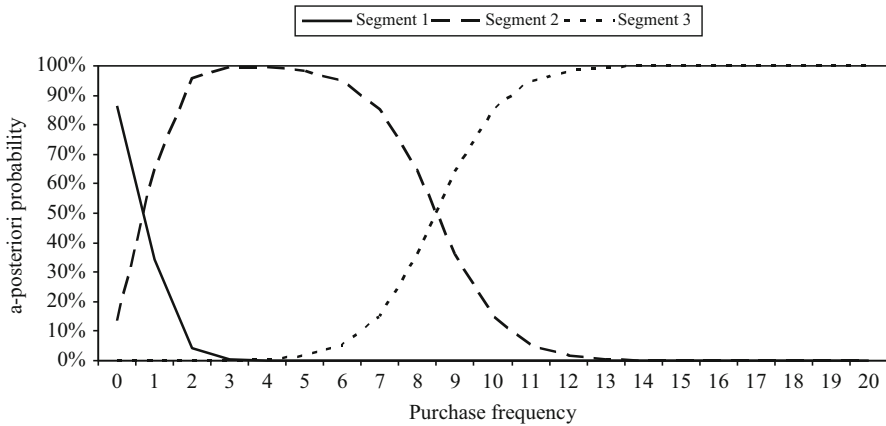


Fig. 5 Relation between purchase frequency and a posteriori probability of segment membership

The following conditions need to hold:

$$0 \leq \omega_{h,s} \leq 1 \quad \forall h \in H, s \in S \tag{5}$$

$$\sum_{s \in S} \omega_{h,s} = 1 \quad \forall h \in H \tag{6}$$

The a posteriori probability considers the probability to observe the actual purchase frequency given that a consumer belongs to a specific segments and weighs this probability with the relative size of the segment. For managerial purposes, the consumer is ‘assigned’ to the segment for which he/she has the highest a posteriori probability of segment membership.

Figure 5 shows the relationship between the observed purchase frequency of a consumer and his/her a posteriori probability of segment membership. For example, Fig. 5 illustrates that a consumer who buys four chocolate bars has a posteriori probability of segment membership that is close to one for segment 2. Thus, this consumer would be assigned to segment 2, which is characterized by a mean purchase frequency of 3.5 (“medium buyers”). In contrast, a consumer who buys 14 chocolate bars belongs to segment 3 with probability almost equal to one. Segment 3 represents the “heavy buyers” of chocolate bars (mean = 11.2).

Determining the Number of Segments

The most critical decision when conducting a segmentation analysis is to determine the number of segments. Traditional clustering approaches use rather arbitrary criteria to make this decision. In Finite Mixture models, statistical decision criteria are available. These criteria use the model fit and a posteriori segment membership probability to determine the number of segments.

The most prominent criteria are the so-called *information criteria*. The information criteria use the deviance (i.e., $-2\ln L$ of the estimated model) which reflects the model fit and a *penalty factor*. The penalty factors take the number of estimated parameters and/or observations into account. The lower the value for the information criterion the better. Thus, when comparing two models that differ with respect to the number of segments, the researcher selects the model with the lower value for the information criterion. The idea behind is that the deviance decreases when more segments are considered since an increasing number of segments improves the flexibility, that is, it becomes easier to capture the nature of the observed data. However, increasing the number of segments results in an increasing number of parameters and less degrees of freedom. The decreasing degrees of freedom should be taken into account when making a decision of which model is most appropriate (similar to adjusted R^2 in OLS regression analysis). In the following, different popular information criteria are described.

The Akaike's information criterion (AIC) suggests using two times the number of estimated parameters as a penalty factor (Bozdogan 1987; McLachlan and Peel 2001, p. 203):

$$\text{AIC}(S) = -2 \ln L + 2|K| \quad (7)$$

where

$|K|$: number of elements in the index set of estimated parameters

The modified Akaike information criterion (also called AIC3) uses a penalty factor of three. The number of estimated parameters has a stronger negative effect (Andrews and Currim 2003a):

$$\text{AIC3}(S) = -2 \ln L + 3|K| \quad (8)$$

Consistent Akaike information criterion (CAIC) and the Bayesian information criterion (BIC) consider the number of observations in addition to the number of estimated parameters (Wedel and Kamakura 2000, p. 92):

$$\text{CAIC}(S) = -2 \ln L + (\ln(|H||I|) + 1)|K| \quad (9)$$

$$\text{BIC}(S) = -2 \ln L + (\ln(|H||I|))|K| \quad (10)$$

where

$|H|$: number of elements in the index set of consumers

$|I|$: number of elements in the index set of observations for each consumer

There is no single best information criterion. Several simulation studies indicate the weaknesses of the different information criteria in certain settings. The studies suggest that the AIC tends to overestimate the number of segments (Ramaswamy et al. 1993). Andrews and Currim (2003a, b) suggested that AIC3 is an appropriate criterion in many settings – especially with smaller sample sizes. With large samples ($n > 300$), BIC and CAIC perform well (Andrews and Currim 2003a). Sarstedt et al. (2011) find that AIC4 (penalty factor of 4) performs generally better than BIC and

CAIC. Given these ambiguous results, a researcher may want to discuss multiple information criteria and argue for a certain solution using also alternative criteria.

An alternative criterion is the entropy. The entropy metric considers the a posteriori segment membership probability:

$$\text{Entropy} = 1 + \frac{\sum_{h \in H} \sum_{s \in S} \omega_{h,s} \ln \omega_{h,s}}{|H| \ln |S|} \quad (11)$$

where

$|S|$: number of segments

The entropy measure ranges between zero and one. If the a posteriori segment membership probabilities are very similar across all segments, the solution is fuzzy, and the entropy measure would be close to zero. Imagine a three-segment solution and the a posteriori segment membership probability is one third for all consumers for all three segments. In this specific case, the resulting value for the entropy measure equals zero. If the a posteriori segment membership probability is exactly one for one specific segment for all consumers, entropy equals one (Ramaswamy et al. 1993). Thus, the closer the value of the entropy to one, the better solution in the sense that the segments are better separated. A good separation is critical for deriving managerial implications later on.

In addition to these model-based criteria, one should evaluate whether the identified segments are actionable, differentiable, and substantial (e.g., Kotler and Keller 2012). Hence, the finally chosen segment solution might not always be the “best” one in statistical terms.

Popular Applications of Finite Mixture Models in Multivariate Methods of Analysis

A main reason for the popularity of Finite Mixture models is that they can easily be implemented within other multivariate methods of analysis. While segment-level solutions are attractive from a managerial standpoint, multivariate methods of analysis inherently assume either an individual- (e.g., conjoint analysis, multi-dimensional scaling) or aggregate-level of analysis (e.g., logit models, structural equation models).

For multivariate methods that originally perform an individual estimation of the parameters, employing a Finite Mixture model will reduce the variance of the estimated parameters through segment-based estimation. For multivariate methods that originally estimate the parameters at aggregate level (i.e., assuming homogeneity), incorporating the inherent heterogeneity of the consumers into the model can reduce systematic biases in the estimated parameters. Note that ultimately there will always be a trade-off between variance and systematic bias: segment-based estimation rather than individual estimation leads to some systematic bias, as the heterogeneity of the consumers is less accurately captured. On the other hand, performing

segment-based estimation rather than aggregate estimation has a negative effect on the variance of the estimated parameters. Nevertheless, a segment-level analysis is attractive from a managerial perspective and builds the basis of many marketing strategies.

Before the advent of Finite Mixture models, researchers used two-step procedures to derive results at the segment level. In case of an original analysis at the individual level, the researcher used the individual-level parameters as segmentation variables in traditional clustering approaches (see, e.g., Green and Krieger 1991). However, this approach ignores that the segmentation variables are estimates in themselves. Moreover, two objective functions are optimized independently, for example, minimizing the squared errors in a regression and minimizing the within-group variance in K-means clustering. In case of an original analysis at the aggregate level, the researcher used a priori segmentation and then estimated the parameters for the predefined segments. This approach requires a thorough knowledge of the source of heterogeneity when defining the segmentation variables.

Implementing the Finite Mixture approach in multivariate methods of analysis leads to the specification of one likelihood function. Thus, there is only one single optimization step and no need of optimizing multiple functions with different and maybe conflicting objectives.

The most popular applications of Finite Mixture models in combination with multivariate methods of analysis are regression analysis (special case: conjoint analysis), logit models (special case: choice-based conjoint analysis), multi-dimensional scaling, and structural equation models. Table 2 lists the advantage of using the Finite Mixture model in combination with the multivariate method of analysis and refers the interested reader to articles that describe the approach in more detail or represent some recent applications of the approach.

Conclusion

Segment-level analyses are particularly useful to managers, as it enables them to target consumers effectively. Finite Mixture models are therefore highly relevant in practice, and software developments, such as LatentGold[®], Sawtooth[®], or SmartPLS[®] have supported the increasingly widespread adoption.

Finite Mixture models provide a flexible framework for performing model-based estimations of segment-specific parameters. Combining Finite Mixture models with multivariate methods of analysis makes it possible to estimate segment-specific parameters, segment sizes, and the a posteriori probability of segment membership simultaneously for each consumer. Thus, for multivariate methods of analysis that traditionally operate at an aggregate level (i.e., assuming homogeneity across consumers), a reduction in systematic bias in the estimated parameters can be achieved by considering consumer heterogeneity. For multivariate methods of analysis that traditionally operate at an individual level, estimating segment-specific parameters instead can yield more stable estimates.

Table 2 Overview of main applications of Finite Mixture models in multivariate methods

Multivariate method of analysis	Original level of aggregation	Advantage of using a Finite Mixture model	Description of approach	Exemplary applications
Regression analysis	Aggregate	Reduced systematic bias in the estimated utility parameters	Wedel and DeSarbo (1994)	Decker and Trusov (2010) Petersen and Kumar (2015) Srinivasan (2006)
Conjoint analysis	Individual	Reduced variance of the estimated utility parameters	DeSarbo et al. (1992) Kamakura et al. (1994)	DeSarbo et al. (1992)
Logit models (choice-based conjoint analysis)	Aggregate	Reduced systematic bias in the estimated utility parameters	DeSarbo et al. (1995) Natter and Feurstein (2002) Kamakura et al. (1994)	Papies et al. (2011) Steiner et al. (2016) Ailawadi et al. (2014)
Multidimensional scaling	Individual	Reduced variance of the estimated utility parameters	DeSarbo et al. (1994) DeSarbo et al. (1991) DeSarbo and Wu (2001) Wedel and DeSarbo (1996)	Natter et al. (2008)
Structural equation modelling	Aggregate	Reduced systematic bias in the estimated utility parameters	Jedidi et al. (1997) Sarstedt and Ringle (2010)	DeSarbo et al. (2006) Haapanen et al. (2016) Wilden and Gudergan (2015)

Finite Mixture models are based on a fuzzy partition of the consumers into segments, and they assume that there exists a finite number of segments that are homogenous in themselves. Yet, this assumption is a weakness of the Finite Mixture model approach and has been addressed in research (e.g., Lenk and DeSarbo 2000).

References

- Ailawadi, K. L., Gedenk, K., Langer, T., Ma, Y., & Neslin, S. A. (2014). Consumer response to uncertain promotions: An empirical analysis of conditional rebates. *International Journal of Research in Marketing*, 31(1), 94–106.
- Andrews, R., & Currim, I. (2003a). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, 40(2), 235–243.
- Andrews, R., & Currim, I. (2003b). Retention of latent segments in regression-based marketing models. *International Journal of Research in Marketing*, 20(4), 315–321.
- Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC). The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370.
- Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, 27(4), 293–307.
- DeSarbo, W., Howard, D., & Jedidi, K. (1991). MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. *Psychometrika*, 56(1), 121–136.
- DeSarbo, W., Wedel, M., Vriens, M., & Ramaswamy, V. (1992). Latent class metric conjoint analysis. *Marketing Letters*, 3(3), 273–288.
- DeSarbo, W., Manrai, A., & Manrai, L. (1994). Latent class multidimensional scaling. A review of recent developments in the marketing and psychometric literature. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 190–222). Cambridge, MA: Blackwell Publishers.
- DeSarbo, W., Ramaswamy, V., & Cohen, S. (1995). Market segmentation with choice-based conjoint analysis. *Marketing Letters*, 6(2), 137–147.
- DeSarbo, W. S., & Wu, J. (2001). The joint spatial representation of multiple variable batteries collected in marketing research. *Journal of Marketing Research*, 38(2), 244–253.
- DeSarbo, W., Di Benedetto, C., Jedidi, K., & Song, M. (2006). Identifying sources of heterogeneity for empirically deriving strategic types: A constrained finite-mixture structural-equation methodology. *Management Science*, 52(6), 909–924.
- Dillon, W. R., & Kumar, A. (1994). Latent structure and other mixture models in marketing: An integrative survey and overview. In R. P. Bagozzi (Ed.), *Advanced methods for marketing research* (pp. 295–351). Cambridge, MA: Blackwell Publishers.
- Green, P., & Krieger, A. (1991). Segmenting markets with conjoint analysis. *Journal of Marketing*, 55(4), 20–31.
- Green, P., Carmone, F., & Wachspress, D. (1976). Consumer segmentation via latent class analysis. *Journal of Consumer Research*, 3(3), 170–174.
- Haapanen, L., Juntunen, M., & Juntunen, J. (2016). Firms' capability portfolios throughout international expansion: A latent class approach. *Journal of Business Research*, 69(12), 5578–5586.
- Jedidi, K., Jagpal, H., & DeSarbo, W. (1997). Finite mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1), 39–59.
- Kamakura, W. A., Wedel, M., & Agrawal, J. (1994). Concomitant variable latent class models for conjoint analysis. *International Journal of Research in Marketing*, 11(5), 451–464.
- Kotler, P. T., & Keller, K. L. (2012). *Marketing Management*. Pearson.
- Lenk, P., & DeSarbo, W. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1), 93–119.
- Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, 20, 36–43.
- McLachlan, G., & Peel, D. (2001). *Finite mixture models*. New York: Wiley.
- Natter, M., & Feurstein, M. (2002). Real world performance of choice-based conjoint models. *European Journal of Operational Research*, 137(2), 448–458.
- Natter, M., Mild, A., Wagner, U., & Taudes, A. (2008). Planning new tariffs at tele.ring: The application and impact of an integrated segmentation, targeting, and positioning tool. *Marketing Science*, 27(4), 600–609.

- Papies, D., Eggers, F., & Wlömert, N. (2011). Music for free? How free ad-funded downloads affect consumer choice. *Journal of the Academy of Marketing Science*, 39(5), 777–794.
- Petersen, A., & Kumar, V. (2015). Perceived risk, product returns, and optimal resource allocation: Evidence from a field experiment. *Journal of Marketing Research*, 52(2), 268–285.
- Ramaswamy, V., DeSarbo, W., Reibstein, D., & Robinson, W. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, 12(1), 103–124.
- Sarstedt, M., & Ringle, C. M. (2010). Treating unobserved heterogeneity in PLS path modeling: A comparison of FIMIX-PLS with different data analysis strategies. *Journal of Applied Statistics*, 37(8), 1299–1318.
- Sarstedt, M., Becker, J.-M., Ringle, C., & Schwaiger, M. (2011). Uncovering and treating unobserved heterogeneity with FIMIX-PLS: Which model selection criterion provides an appropriate number of segments? *Schmalenbach Business Review*, 63, 34–62.
- Srinivasan, R. (2006). Dual distribution and intangible firm value: Franchising in restaurant chains. *Journal of Marketing*, 70(3), 120–135.
- Steiner, M., Wiegand, N., Eggert, A., & Backhaus, K. (2016). Platform adoption in system markets: The roles of preference heterogeneity and consumer expectations. *International Journal of Research in Marketing*, 33(2), 276–296.
- Wedel, M., & DeSarbo, W. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, MA: Blackwell Publishers.
- Wedel, M., & DeSarbo, W. (1996). An exponential-family multidimensional scaling mixture methodology. *Journal of Business & Economic Statistics*, 14(4), 447–459.
- Wedel, M., & Kamakura, W. (2000). *Market segmentation. Conceptual and methodological foundations*. Norwell: Kluwer.
- Wilden, R., & Gudergan, S. (2015). The impact of dynamic capabilities on operational marketing and technological capabilities: Investigating the role of environmental turbulence. *Journal of the Academy of Marketing Science*, 43(2), 181–199.



Analysis of Variance

Jan R. Landwehr

Contents

Introduction	266
Between-Subjects: One Observation per Person	268
Two Means: One-Factorial ANOVA or Independent-Samples <i>t</i> -Test	268
More Than Two Means: One-Factorial ANOVA	274
Multiplicative Effects: Factorial ANOVA	277
Within-Subjects: Two or More Observations per Person	285
Two Means: One-Factorial RM-ANOVA or Paired-Samples <i>t</i> -Test	286
More Than Two Means: One-Factorial RM-ANOVA	288
Multiplicative Effects: Factorial RM-ANOVA/Mixed-ANOVA	289
Extensions	291
Analysis of Covariance (ANCOVA)	291
Multivariate Analysis of Variance (MANOVA)	292
Conclusion	293
References	296

Abstract

Experiments are becoming increasingly important in marketing research. Suppose a company has to decide which of three potential new brand logos should be used in the future. An experiment in which three groups of participants rate their liking of one of the logos would provide the necessary information to make this decision. The statistical challenge is to determine which (if any) of the three logos is liked significantly more than the others. The adequate statistical technique to

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-3-319-57413-4_16) contains supplementary material, which is available to authorized users.

J. R. Landwehr (✉)
Marketing Department, Goethe University Frankfurt, Frankfurt, Germany
e-mail: landwehr@wiwi.uni-frankfurt.de

assess the statistical significance of such mean differences between groups of participants is called analysis of variance (ANOVA). The present chapter provides an introduction to the key statistical principles of ANOVA and compares this method to the closely related *t*-test, which can alternatively be used if exactly two means need to be compared. Moreover, it provides introductions to the key variants of ANOVA that have been developed for use when participants are exposed to more than one experimental condition (repeated-measures ANOVA), when more than one dependent variable is measured (multivariate ANOVA), or when a continuous control variable is considered (analysis of covariance). This chapter is intended to provide an applied introduction to ANOVA and its variants. Therefore, it is accompanied by an exemplary dataset and self-explanatory command scripts for the statistical software packages R and SPSS, which can be found in the Web-Appendix.

Keywords

ANOVA · ANCOVA · RM-ANOVA · MANOVA · Mixed-ANOVA · Split-plot ANOVA · *t*-test · GLM · Experimental design · *F*-distribution · Between-subjects · Within-subjects · Mean comparison · Sum of squares · Effect size · Confidence intervals · Effect coding · Simple effects · Disordinal interaction · Crossover interaction · R · SPSS

Introduction

The term analysis of variance (ANOVA) refers to a family of statistical methods that are closely linked to the analysis of experimental data where a continuous outcome variable (i.e., dependent variable, DV) is explained by one or more experimental factors with discrete levels (i.e., independent variable(s), IVs). For instance, if a company conducts an experiment in which consumers have to rate their liking of an emotional advertisement or a reason-based advertisement for the same product, ANOVA can be used to determine whether the mean liking ratings (the DV) of the two types of advertisements (the IV) differ significantly. To determine whether a significant mean difference is present, ANOVA follows an indirect testing strategy rather than directly comparing the means. In particular, it compares the explained part of the variance in the data (i.e., the systematic variance) to the unexplained part of the variance (i.e., the error variance) and determines whether the explained part of the variance is significantly larger. ANOVA thus compares the relative size of the variances, which gives the approach its name.

The general structure of the present chapter follows a crucial distinction of experimental design: between-subjects and within-subjects. In a between-subjects



design, the experimental variable is manipulated such that each participant is only exposed to one level of the independent variable. In the earlier advertisement example, this would mean that a participant is randomly assigned to either the emotional or the reason-based advertisement, and the means of these two groups of participants are compared. In contrast, in a within-subjects design, the independent variable is manipulated within each participant such that a participant is exposed to both versions of the advertisement and must provide two judgments. These judgments are compared across all participants using an extension of ANOVA called repeated-measures analysis of variance (RM-ANOVA). The differences between these two approaches and their respective statistical advantages and disadvantages are discussed in the second part of this chapter.

For both between- and within-subjects designs, three versions of ANOVA models will be discussed: a simple comparison of two means, a slightly more complicated version with three means, and the simultaneous examination of two IVs and their interactive effect on a DV (each IV with two levels, resulting in $2 \times 2 = 4$ means to compare). It is important to note that the naming of an ANOVA model depends on the number of factors considered (in the terminology of ANOVA, the terms factor and IV can be used interchangeably). A model with just one factor is called a one-way ANOVA, a model with two factors is called a two-way ANOVA, and so forth (there are few ANOVA models with more than three factors because such models are very hard to interpret). After an extensive discussion of these ANOVA models, this chapter ends with a brief introduction to two variants of ANOVA: analysis of covariance (ANCOVA), where in addition to IVs with discrete levels, continuous IVs are included in the ANOVA model, and multivariate analysis of variance (MANOVA), which is used to analyze several DVs at the same time.

To provide an easy access to these ANOVA models, the presentation of the theory behind these models will be accompanied by an exemplary dataset. The dataset is simulated and is based on a series of hypothetical market research experiments. The parameters used to simulate the data ensure that we obtain statistically significant results throughout the chapter. It is important to note that real studies do usually not produce such a perfect pattern of results. Moreover, for ease of presentation, the results of all market research experiments are saved within only one dataset, as if each participant took part in all experiments. In reality, one would rather use different samples of participants for different experiments and would hence save the data to different data files. Throughout the chapter, names printed in *italics* refer to the variable names of this exemplary dataset.

All analyses described in this chapter were conducted using the statistical software R. This powerful statistical software is free of charge and is continuously improved and extended by world-leading statisticians on an open-access basis. The downside is that it is not as easy to use as menu-based statistical software, such as SPSS. To provide a versatile applied introduction to ANOVA, the Web-Appendix contains the simulated dataset and all R scripts used for the analyses described in this chapter in addition to a corresponding SPSS syntax file with explanatory comments.

Table 1 Experimental conditions of Study 1

Condition (1) Simple font	Condition (2) Complex font
	

Note: All imaginary bicycle brand logos used in this text are designed by Veronika König (www.nachtundtag.com)

For ease of recognition, the R commands used throughout this chapter are printed in Courier New.

Between-Subjects: One Observation per Person

Two Means: One-Factorial ANOVA or Independent-Samples t-Test

Let us start with a simple comparison between two means. Suppose a company would like to launch a new brand of exclusive bicycles targeted at successful businesspeople. The brand is to be called “BUYCYCLE.” The brand manager considers a simple reduced font and a more complex font as candidates for printing the brand name (see Table 1). To explore which of these two versions is preferred by their target group of consumers, he decides to conduct a market research experiment to compare the liking of the two fonts. He randomly splits a sample of 120 potential consumers into two groups. Each group is shown one of the font versions of the brand name and is asked to rate their liking of the brand name. The aim of this research is to determine whether there is a difference in the mean liking of the simple-font group compared to the complex-font group.

A typical data matrix for this type of study would look like the matrix shown in Table 2, where the first six cases and the last case of a sample of 120 participants are shown (the full dataset can be found in the Web-Appendix). The experimental factor is named *iv_2*, and the dependent variable liking is named *dv_2*.¹ Before computing inferential test statistics, the first reasonable step is to visually display the key patterns in the data. In the current example, we are interested in the statistical significance of the mean difference between the two experimental groups. Hence, we are interested in displaying the two means, which is usually done by a barplot for discrete IVs² (see Fig. 1). Moreover, to gain an impression of the random noise in the data, we would like to include an indicator of random variation. Usually, either the standard error of the mean or, more often and hence recommended, a 95% confidence interval is used (which is, for sufficiently large samples of $N > 30$, equal to approximately 1.96 times the standard error, Field et al. 2012, pp. 45–46).

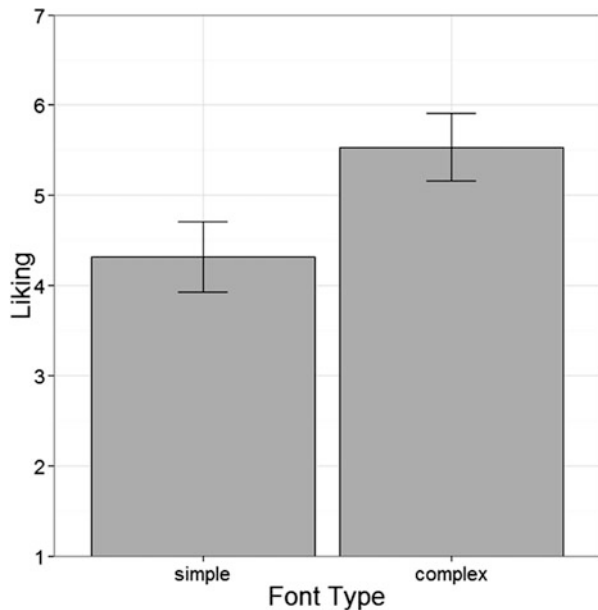
¹The naming of the variables throughout the chapter follows the key characteristic of the respective experimental scenario. “_2” refers to the two factor levels employed in the present experiment. All variable names are constructed following the same logic.

²All barplots in this chapter were produced using the `ggplot2`-library in R (Wickham 2009).

Table 2 Data matrix for the first few participants of Study 1

id	gender	age	iv_2	dv_2
1	male	28	complex	7
2	male	23	simple	1
3	male	20	simple	3
4	male	24	complex	6
5	female	25	simple	4
6	female	25	simple	7
⋮	⋮	⋮	⋮	⋮
120	female	23	complex	3

Fig. 1 Mean liking ratings of Study 1 with 95% confidence intervals



Interestingly, in a between-subjects design, the statistical significance of a mean difference can easily be inferred from a mere visual inspection of the means and their 95% confidence intervals (“inference by eye”; Cumming and Finch 2005). When the 95% confidence intervals do not overlap (as in Fig. 1), the mean difference is significant at the $p < 0.01$ level. When they overlap by less than half the length of the intervals’ whiskers (i.e., half the length between the upper/lower bound of the interval and the mean), the mean difference is significant at the $p < 0.05$ level. This technique is an efficient way to obtain a quick but solid impression of the data pattern. A more precise way to assess the statistical significance in terms of an exact probability estimate (i.e., p -value) is to conduct either a t -test or an ANOVA. In practice, only the

t -test would be conducted and reported for the given research question because it is the more parsimonious technique for a comparison of just two means. However, I will also introduce the key ideas of ANOVA using this simple dataset.

An independent-samples t -test can compare a maximum of two means. It evaluates the ratio of the mean difference between the two experimental groups and the standard error (SE) of the mean difference:

$$t_{\text{emp}} = (M_1 - M_2) / \text{SE}(M_1 - M_2) \quad (1)$$

The higher the empirical t -value, the less likely is a purely random mean difference. The theoretical t -distribution can be used to compute the exact likelihood of observing an empirical mean difference given the null hypothesis that both means are the same, which is denoted as the p -value. For the present example, the mean is 4.32 for the simple font condition and 5.53 for the complex font condition. The empirical t -value for the difference between these two means is -4.53 (the negative sign indicates that the first mean is smaller than the second). To derive the corresponding p -value from the theoretical t -distribution, the degrees of freedom for estimating the difference between two means are calculated. The total sample size is 120. Since two means need to be estimated, two degrees of freedom are “consumed” and 118 degrees of freedom remain for the analysis. The p -value corresponding to a t -value of -4.53 with 118 degrees of freedom can be looked up in a t -table and is smaller than 0.001. In practice, the statistical software package automatically provides the p -value, and there is no need to find a classical statistics book that still contains tables with exact values for the t -distribution. When reporting the result of a t -test in a scientific manuscript, one would write that the difference in means between the simple font condition ($M = 4.32$) and the complex font condition ($M = 5.53$) is statistically significant ($t(118) = -4.53; p < 0.001$).

Alternatively, we could also run an ANOVA model to estimate whether there is a significant mean difference between the two experimental groups. Since ANOVA is the key topic of the present chapter, I will present the theoretical background of this method in greater detail. When considering ANOVA in general, it is important to note that it is not a unique statistical method but is closely linked to many other statistical techniques. In particular, as is almost every method of inferential statistics, ANOVA is a special case of the general linear model (GLM) underlying regression analysis (Rutherford 2001). Hence, the model formula of a one-way ANOVA is similar to the usual regression model, with y_{ig} indicating the DV of participant i in condition g , \bar{y} indicating the grand mean across all observations, α_g indicating the impact of the experimental manipulation with g levels, and e_{ig} indicating the residual:

$$y_{ig} = \bar{y} + \alpha_g + e_{ig} \quad (2)$$

As the name analysis of variance suggests, the aim of ANOVA is to examine the sources of variation in the data. In particular, the total amount of variance in the data

is partitioned into an explained and an unexplained part of the total variance. A reasonable ANOVA model should explain more variance than is left unexplained by the model, that is, the ratio of explained to unexplained variance should be greater than 1. How much greater it must be to reach statistical significance is determined based on the F -distribution, which is a probability density function of the ratios of variances.

To provide a better understanding of these different parts of variance and the necessary steps to determine F - and p -values, we consider the first six cases of the dataset featured in the Web-Appendix. As hopefully will become clear, ANOVA is a mathematically simple technique that can be easily computed by hand (although it would be quite annoying for larger datasets). Figure 2 shows the key elements necessary to understand the mechanics of ANOVA: the computation of the total sum of squares (SS_T) in Fig. 2a, the computation of the model sum of squares (SS_M) in Fig. 2b, and the computation of the residual sum of squares (SS_R) in Fig. 2c. All three figures depict the same three key elements: first, the observed liking evaluations of the first six participants (participant ID is shown on the x-axis). The dark gray circles indicate that the person is in the simple font condition, and the light gray diamonds indicate that the person is in the complex font condition; second, the mean evaluation across all 120 participants (independent of the experimental group), as indicated by the horizontal black line, which is also called the “grand mean”; and third, the mean evaluation of the 60 observations of the simple font (horizontal dark gray line at 4.32) and the mean evaluation of the 60 observations of the complex font condition (horizontal light gray line at 5.53).

The dashed vertical lines differ between the three figures and indicate three different sources of variation in the data. In Fig. 2a, the dashed lines indicate the deviation of each individual observation from the grand mean. As shown in formula (3a), the sum of these deviations squared is defined as SS_T . Because the magnitude of this measure directly depends on the number of observations (i.e., every additional participant will add a squared deviation from the mean), it has no meaningful interpretation. To obtain meaning, we can calculate the total mean squares (MS_T), as shown in formula (4a), by dividing SS_T by the corresponding degrees of freedom. Since the grand mean needs to be estimated from the data, one degree of freedom is “consumed.” Hence, the degrees of freedom are computed as the sample size minus one (df: $N - 1$; here: $120 - 1 = 119$). Consequently, MS_T is the average squared deviation from the grand mean, which is better known by the term “variance” (i.e., $MS_T =$ the total variance in the data).

The SS_T can be decomposed into two components: the part of the variation in the data that is explained by the statistical model SS_M and the part that is unexplained SS_R (i.e., $SS_T = SS_M + SS_R$). Figure 2b shows how SS_M is computed. The ANOVA model defines the two experimental groups as sources of systematic variation. Hence, the mean of the simple and complex group, respectively, is the part of the variation in the data that is explained by the statistical model. SS_M is defined as the sum of the squared deviations of the group means from the grand mean (see formula (3b)). As indicated by the vertical dashed lines in Fig. 2b, there are as many squared deviations entering SS_M as there are participants. We can compute the average SS_M by dividing SS_M by the

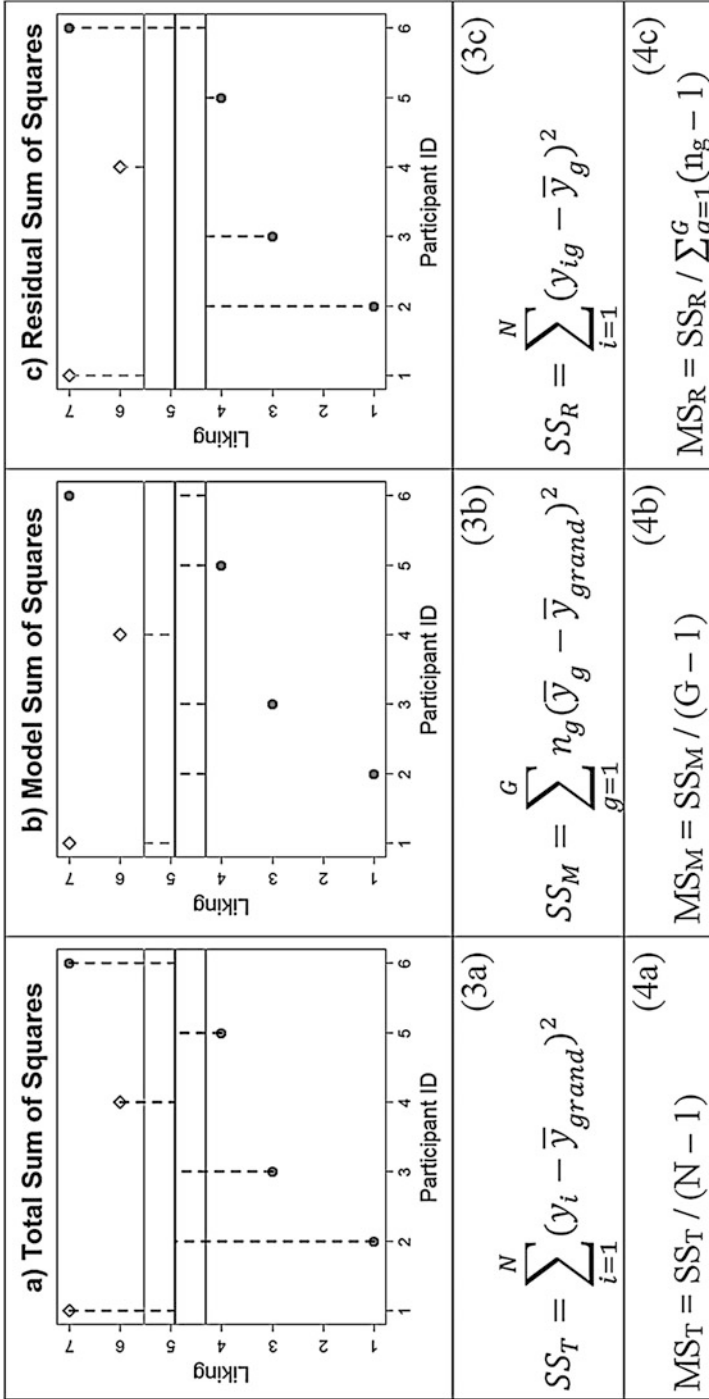


Fig. 2 Visualization of the different sources of variation in a one-factorial ANOVA with two groups. *Note:* The figures show the first six cases of the exemplary dataset. The *dark gray circles (light gray diamonds)* indicate individual observations from the simple (complex) font condition. The *horizontal black line* indicates the grand mean across all observations, and the *dark (light) gray line* indicates the mean of the simple (complex) font condition. The sum of the squares of the *vertical dashed lines* is computed in the formulas below the figures

corresponding df (i.e., the number of factor levels minus 1; here: $2 - 1 = 1$; see formula (4b)). The resulting MS_M is the systematic variance in the data.

Finally, Fig. 2c shows how the unsystematic variance in the data is computed. The vertical dashed lines indicate the deviation of individual observations from the mean of the group the observation belongs to. As defined in formula (3c), the sum of these squared deviations across all observations is called SS_R . To compute the residual variance, we again divide SS_R by the corresponding df (i.e., the sum of the $n - 1$ participants per experimental group; here: $59 + 59 = 118$; see formula (4c)).

Given the variances computed by formulas (4b) and (4c), we are ready to perform the test of statistical significance. To this end, we compute the empirical F -value:

$$F_{\text{emp}} = MS_M / MS_R \quad (5)$$

Hence, F_{emp} represents the ratio of the systematic and the unsystematic variance in the data. Clearly, if F_{emp} is smaller than 1, there cannot be a significant effect in the data because the unsystematic variance is larger than the systematic variance. When F_{emp} is larger than 1, the question arises of how much larger than 1 F_{emp} must be to call the effect statistically significant. To compute the precise level of statistical significance (i.e., the p -value), the empirical F -value F_{emp} is compared to the theoretical probability density function of F -values. This distribution describes how ratios of variance are distributed and provides the critical thresholds that have to be exceeded to infer that the systematic variance is so much larger than the unsystematic variance that a mere random difference between the variances is unlikely (i.e., less than 5%). In contrast to the previously discussed t -distribution with only one parameter, the F -distribution has two parameters that determine the shape of the distribution. The first parameter is the degrees of freedom of the systematic variance (i.e., MS_M); the second parameter is the degrees of freedom of the unsystematic variance (i.e., MS_R). As in the previous t -test example, classic textbooks on ANOVA contain tables of critical F -values ordered by the model degrees of freedom and the residual degrees of freedom. Nowadays, statistical software packages do the tedious job of computing the exact probability for a given F_{emp} with its two corresponding degrees of freedom. For the given dataset, the ANOVA output of R, including all discussed elements, is shown in Fig. 3.

In the figure, MS_M is 44.41, MS_R is 2.17, and F_{emp} is $44.41/2.17 = 20.48$. The model degrees of freedom is 1. The residual degrees of freedom is 118. R computed that an F_{emp} of 20.48 given 1 and 118 degrees of freedom is very unlikely ($\text{Pr}(>F) = p = 0.0000145$).³ In a scientific text, the results of the present ANOVA would commonly be described as follows: An ANOVA showed that the factor font type has a significant influence on participants' liking evaluations ($F(1, 118) = 20.48$; $p < 0.001$). In particular, the complex font ($M = 5.53$) is systematically liked better than the simple font ($M = 4.32$).

³R denotes the p -value by " $\text{Pr}(>F)$," which refers to the probability of observing the empirical F -value given the null hypothesis. R uses exponential notation to show small numbers. Hence, the value $1.45\text{e-}05$ in Fig. 3 is equivalent to 0.0000145.

```

> summary(aov(dv_2 ~ iv_2, data=Data.Anova))
              Df Sum Sq Mean Sq F value    Pr(>F)
iv_2           1  44.41   44.41   20.48 1.45e-05 ***
Residuals     118 255.92    2.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig. 3 R output of a one-way ANOVA with two factor levels (Study 1)

Table 3 The third condition of the second study that is added to the two conditions shown in Table 1

Condition (3) Super-complex font



More Than Two Means: One-Factorial ANOVA

The example described in the previous section did not require an ANOVA model but could have been analyzed by a t -test because only two means were compared. We turn now to the key strength of ANOVA: situations where more than two means are involved. Since the t -test is limited to a comparison of two means, it cannot be applied to such situations. In what follows, we extend the two group example of the previous section by a third experimental group.

When the brand manager of our imaginary company sees the results of the first study, he gets excited about complex fonts. He asks his team of designers to find an even more complex font, expecting an even higher liking due to the increased complexity. However, the head of market research is skeptical and proposes a second study where all three fonts are compared. She sets up a study with one factor (font type) that has three levels: the two levels of the first study plus a third level, the “super-complex font” condition (see Table 3).

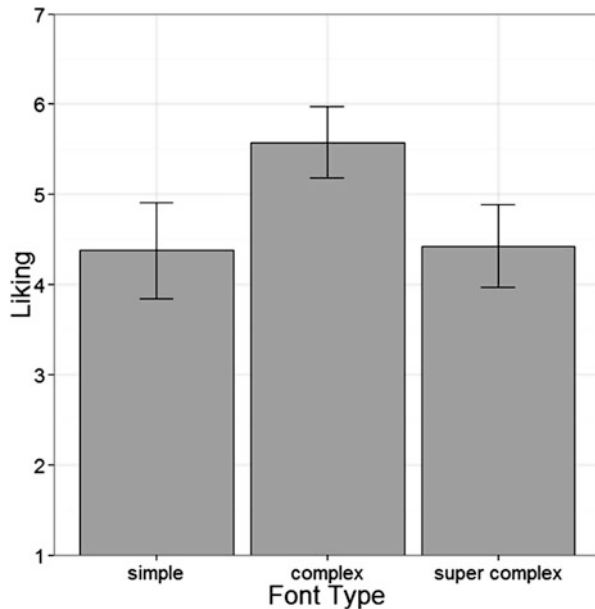
For ease of data handling, the dataset in the Web-Appendix contains the data of this second experiment, too.⁴ The experimental factor of this second study is named iv_3 , and the dependent variable is named dv_3 . A total of 120 participants

⁴In real data collections, we would collect a second independent dataset from new participants. Please assume that although the data for the second experiment (and all further studies) are stored in the same dataset, these datasets are independent and come from different participants.

completed this study, with 40 participants per experimental condition. Before conducting statistical tests, we inspect the descriptive pattern of the results by plotting the means and their 95% confidence intervals (see Fig. 4). The pattern of means suggests that the medium complex font still performs best.

To formally confirm this “inference by eye,” we conduct a one-factorial ANOVA. The model formula is identical to formula (2), and the approach to compute the variances and the empirical F -value is identical to Fig. 2. The only difference is that the group index g now has three levels instead of two. Before running the actual analysis, it is important to note that ANOVA compares the variance explained by one factor (in the present case with three levels) to the unexplained variance and computes one F -value and one p -value per factor, that is, the ANOVA shows whether the factor as a whole explains a significant amount of variance. However, one does not learn which levels of the factor are responsible for the effect. Consider the ANOVA output produced by R for the given dataset (Fig. 5).

Fig. 4 Mean liking ratings of Study 2 with 95% confidence intervals



```
> summary(aov(dv_3 ~ iv_3, data=Data.Anova))
              Df Sum Sq Mean Sq F value Pr(>F)
iv_3           2  36.87  18.433   8.664 0.00031 ***
Residuals    117 248.92   2.128
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 5 R output of a one-way ANOVA with three factor levels (Study 2)

The ANOVA shows that the factor “font type” has a significant influence on participants’ liking evaluations ($F(2, 117) = 8.66; p < 0.001$).⁵ This result tells us that, according to the presented rationale, at least one of the three means is significantly different from at least one other mean. However, this result alone does not inform the researcher which of the three potential pairwise mean comparisons is significant. This is a very important characteristic of ANOVA that, when ignored, leads to unjustified claims about statistically significant mean differences. To assess the significance of the pairwise mean differences, it is necessary to compare each pair of means separately using a priori contrasts, post hoc tests, or inference by eye.

Which of these techniques should a researcher use? A priori contrasts require that one knows prior to the analysis the exact means or group of means one would like to compare. Given that market researchers are usually interested in novel phenomena, it is unlikely that exact hypotheses about specific mean differences can be derived a priori. Moreover, it is usually difficult to convince a critical reader that a contrast was actually proposed prior to the analysis and is not simply declared as being a priori (Rodger and Roberts 2013). Hence, convincing theorizing is required to justify a priori contrasts.⁶ Post hoc contrasts perform an exploratory comparison of all pairwise group means. One problem associated with such post hoc tests is that the same data are used to perform multiple statistical tests, which increases the likelihood of falsely rejecting a null hypothesis (also known as Type I error inflation or alpha inflation). Over the years, countless post hoc procedures have been proposed to counter alpha inflation, and it has become difficult for a market researcher to make an informed decision about which of these methods to select. They differ mainly with respect to the severity of the alpha correction, which has the downside of reducing statistical power (overlooking a significant effect that is actually present, i.e., Type II error). A test is called liberal if it adjusts the alpha level only slightly, and it is called conservative if it adjusts the alpha level considerably.

The most prominent liberal test is called Fisher’s least significant difference test (LSD, Fisher 1935). This test consists of a sequential two-step test procedure. First, the global significance of the ANOVA is assessed, which is called the omnibus test. If this omnibus test is not significant, the test procedure stops. If it is significant, uncorrected t -tests are performed for all pairs of means. An even simpler version of this test procedure is to inspect the means and their confidence intervals based on the inference by eye technique if, and only if, the global F -test of the ANOVA is significant. A conservative procedure was proposed by a statistician named Carlo Bonferroni (although there is no traceable publication), which adjusts the alpha level by the number of conducted pairwise tests m ($\alpha_{\text{Bonferroni}} = \alpha/m$). A less-conservative version of the Bonferroni correction was proposed by Holm (1979) that tests the mean differences in the order of their magnitude and uses increasingly more relaxed

⁵Please note how the df of the ANOVA changed compared to Fig. 3 due to three rather than two factor levels.

⁶The interested reader can find more information about a priori contrasts (also called planned contrasts) in the textbooks of Field (2013), Field et al. (2012), and of Klockars and Sax (1986).

alpha values for smaller mean differences. Which of these tests to select mainly depends on the consequences of falsely accepting/rejecting a null hypothesis. In most market research applications, the research examines potential business opportunities, where it is preferable to give an alternative a try rather than miss a potentially valuable business opportunity. Hence, in most market research applications, Fisher’s LSD (or even simpler, the inference by eye technique as a follow-up on a globally significant ANOVA omnibus test) is a reasonable choice. However, in other disciplines (such as medicine), falsely rejecting a null hypothesis may have very negative consequences; hence, conservative tests are better suited in such instances. The implementation of the three post hoc tests described in the present section can be found in the Web-Appendix.

Multiplicative Effects: Factorial ANOVA

After the brand manager learns that his idea of a super-complex font does not work as intended, he comes up with a new idea: He wants to examine whether the influence of font type on liking depends on the positioning of the brand. In addition to the brand name “BUYCYCLE,” which is targeted at business people, he creates the brand name “EASYCYCLE,” which is targeted at leisure-oriented consumers. He asks his market researchers to test whether the optimal font depends on the brand positioning. An experimental design with two factors with two levels each is implemented, as shown in Table 4, and data are collected from 30 people per cell ($N_{total} = 120$). In the dataset of the Web-Appendix, the factor “font type” is named *iv_fac_a*, the factor “brand name” is named *iv_fac_b*, and the corresponding liking judgments are stored in a variable named *dv_fac*.

The key idea of the two-way experimental design is to go beyond simple main effects (i.e., A is better than B) and examine conditional effects (i.e., A is only better than B when C). Statistically speaking, the basic one-way ANOVA formula (2) is extended by the second experimental factor and by the interaction (i.e., multiplication) of both factors:

$$y_{igh} = \bar{y} + \alpha_g + \beta_h + (\alpha_g * \beta_h) + e_{igh} \tag{6}$$

Table 4 Experimental conditions of Study 3





		Factor (1) Font type	
		Simple font	Complex font
Factor (2) Brand name	Business	 BUYCYCLE	 BUYCYCLE
	Leisure oriented	 EASYCYCLE	 EASYCYCLE

Table 5 Comparison of a balanced and an unbalanced experimental design

(a) Balanced experimental design: factors are uncorrelated				(b) Unbalanced experimental design: factors are correlated			
		Factor A				Factor A	
		-1	1			-1	1
Factor B	-1	$n = 30$	$n = 30$	Factor B	-1	$n = 45$	$n = 15$
	1	$n = 30$	$n = 30$		1	$n = 15$	$n = 45$

Accordingly, the two-way ANOVA considers four sources of variation: the variance produced by factor α , the variance produced by factor β , the variance produced by the interaction of α and β , and finally the residual variance e , against which the significance of all three effects is evaluated.

Before conducting the two-way ANOVA, we need to consider three important aspects that are relevant for ANOVA with more than one factor: (1) coding of the factors, (2) different ways of computing the model sum of squares, and (3) the interpretation of main effects when an interaction is present.

The coding of the factors refers to the numeric coding that is used to represent the categorical factor levels. For the example provided in Table 4, the factor levels are denoted by verbal labels (i.e., “simple font” versus “complex font” and “business” versus “leisure oriented”). However, a statistical procedure, such as ANOVA, cannot handle verbal information but requires numeric information. Although most statistical software packages automatically transform verbal codes into numeric codes internally, it is important to keep in mind what is going on under the surface of the statistical software package. Readers familiar with regression analysis are used to so-called “dummy coding,” which refers to a coding scheme in which a base category is denoted by 0 and the other category by 1 (for k factor levels, $k-1$ dummy-coded variables are needed). This is the default internal coding R uses when a verbally labeled factor is processed. However, in the context of ANOVA with more than one factor, dummy coding leads to an incorrect evaluation of the main effects because this type of coding dismisses one experimental cell from the mean comparison.⁷ Therefore, effect coding must be used, where one factor level is coded as -1 and the other factor level as 1 (see Table 5a).

Effect coding ensures that for a given factor, both cells corresponding to the value -1 are compared to both cells corresponding to the value 1. This is the default coding scheme SPSS uses in its ANOVA procedure. Hence, when using SPSS, the defaults of the software take care of the coding issue. However, when using R, it is important to change the coding scheme from dummy to effect coding to obtain

⁷For the example with 2×2 experimental cells provided in Table 4, dummy-coding Factor 1 (simple = 0; complex = 1) and Factor 2 (business = 0; leisure = 1) would mean that the effect of Factor 1 compares the cell denoted by $\{0,0\}$ (i.e., “simple and business”) to the two cells for which Factor 1 has the value 1 (i.e., “complex and business” and “complex and leisure”). The cell “simple and leisure” would be omitted from the test of the main effect, which is an undesirable feature of dummy coding when applied to ANOVA models.

meaningful results (see lines 24–41 of the Web-Appendix R script on how to specify effect coding in R).

The second important issue concerning ANOVA with more than one factor is how the model sum of squares is computed. This issue is only relevant when the experimental design is unbalanced (i.e., an unequal number of observations per experimental cell; see Table 5b), which leads to a correlation between factors. To understand this point, let us consider the key characteristic of a bivariate correlation in the present context: knowing the value of one factor provides information about the other factor. In the balanced design in Table 5a, this is not the case. If I know that a person is in condition “1” of Factor A, the likelihood of being in condition “–1” or “1” of Factor B is exactly the same (50% or $n = 30$ in each). In contrast, when considering the unbalanced design in Table 5b, things are different: If I know that a person is in condition “1” of Factor A, the likelihood of being in condition “1” of Factor B is 75% ($n = 45$) and the likelihood of being in condition “–1” is only 25% ($n = 15$). Hence, in unbalanced experimental designs, the factors are correlated, meaning that they share common variance. This issue is known as multicollinearity between predictor variables in regression analysis and poses the problem of assigning explained variance in the DV to the IVs. Before we dig deeper into this problem, it is important to note that in practice, unbalanced designs are much more common than perfectly balanced designs due to dropouts. The exemplary dataset in the Web-Appendix, however, contains a perfectly balanced design, as shown in Table 5a. As shown in the lower figure of Table 6, this situation is unproblematic in terms of assigning explained variance to the factors.

When confronted with an unbalanced design, however, it is important to distinguish between three ways of computing the model sum of squares. In Type I,⁸ the explained variance is assigned to the experimental factors in the order of their specification in the model. Hence, the first factor in the model formula can potentially explain more variance than factors occurring later in the formula. Table 6 visualizes this situation using Venn diagrams. In these Venn diagrams, the dashed black circle symbolizes the total variance of the DV, the black circle the total variance of Factor A, the light gray circle the total variance of Factor B, and the dark gray circle the total variance of the interaction of Factors A and B. When circles overlap, they share common variance (i.e., they are correlated). In particular, the area of the dashed black circle covered by the other circles is the explained part of the variance of the DV.

Given that the factors are entered into the ANOVA according to the following formula ($DV = A + B + A * B$) using Type I sum of squares, Factor A explains the black area in the DV, Factor B the light gray area, and the interaction the dark gray area. If the order of the factors in the formula is changed, the explained parts of the

⁸It is important to note that the term “Type I” is used to denote more than just one statistical concept, which can be confusing. We already encountered the term in the context of the statistical p -value, where falsely rejecting the null hypothesis is called an alpha or Type I error. In the present context, “Type I” refers to a specific way of computing the sum of squares in an ANOVA model, which is completely unrelated to the “Type I error” in statistical hypothesis testing.

Table 6 Visualization of different ways of computing the model sum of squares

	Type I Sum of squares	Type II Sum of squares	Type III Sum of squares
Correlated factors			
Uncorrelated factors			

variance also change, which is a disadvantageous characteristic of Type I sum of squares because the order of the factors in the model formula is rarely meaningful. The second way of computing the sum of squares is Type II, where the explained variance is assigned to the main effects first and the interaction explains the leftover variance. This way of assigning the explained variance in the DV to the factors can only be applied when no significant interaction is present. In such a situation, Type II sum of squares tests the main effects with high statistical power (i.e., high likelihood of detecting significant effects). However, two-way experimental designs are usually conducted because the researcher is particularly interested in the interaction, which brings us to the third way of computing the sum of squares: Type III, where each factor only explains that part of the variance in the data that is uniquely produced by that factor. The shared explained variance of several factors is not assigned to a particular factor when using Type III sum of squares but is nevertheless counted as explained variance in the computation of the total R^2 . Type III sum of squares allows a meaningful interpretation of main effects and interactions when interactions are present and is usually the best choice in the analysis of experimental designs.

Common statistical packages, such as SPSS, use Type III sum of squares as the default option. R uses Type I as its default in the `aov` function. Because there is rarely a natural order of factors in an ANOVA model, using Type I sum of squares may lead to biased results because the arbitrary order of specifying the model in the statistical software can influence the significance of the factors. Hence, I recommend using Type III sum of squares to avoid such biases. The Web-Appendix shows how to use Type III in R using the `Anova` function of the `car` library (Fox and Weisberg 2011).

The third general issue concerning ANOVA with more than one factor is the question of how the main effects can be interpreted when an interaction is present. This is a particularly important point since the mere significance of a main effect can lead to misleading conclusions when the pattern of an existing interaction is ignored. To clarify this point, Fig. 6 shows four prototypical mean patterns of a two-way experimental design. In Fig. 6a, we see a meaningful main effect of Factor A (“1” is better than “-1”) but no other effects. In Fig. 6b, we see two main effects of Factors *A* and *B* (for both factors “1” is better than “-1”), which are both meaningful, but there is no interaction. In Fig. 6c, we see the same two main effects but in addition also an interaction such that the effect of Factor *B* is stronger for level “1” of Factor *A* than for level “-1” of Factor *A*. This situation is called an ordinal interaction because the order of the means is not changed by the interaction. Therefore, both main effects remain meaningful and can be interpreted, that is, the claim that factor level “1” is better than “-1” is true for both factors, independently of the other factor.

Now let us consider Fig. 6d, which features a disordinal interaction (also called “crossover interaction”). On average, level “1” of Factor *A* is better than level “-1.” Given the exemplary mean pattern, this effect would most likely show up as a significant main effect in an ANOVA model. However, this effect would not be meaningful because the claim that factor level “1” of Factor *A* is better than factor level “-1” is not true unconditionally. This claim only holds in condition “1” of

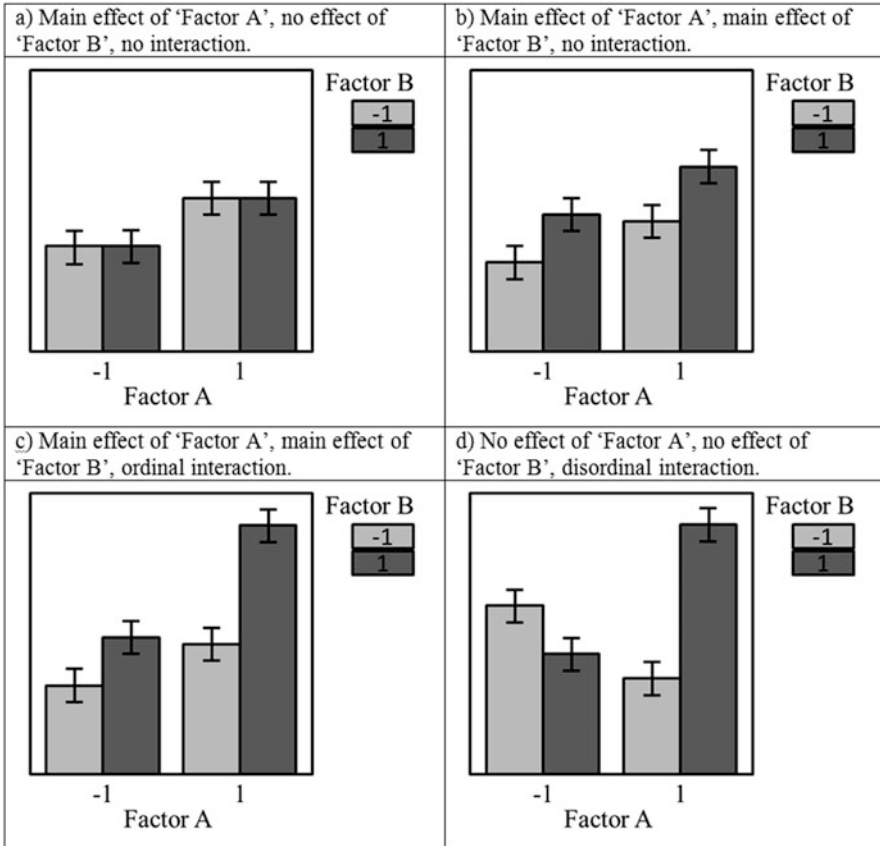


Fig. 6 Prototypical mean patterns of two-factorial experimental designs

Factor B. However, in condition “–1” of Factor B, the effect reverses, that is, the order of the means changes, conditional on the other factor, which makes it a disordinal interaction. Therefore, it is not sufficient to inspect the ANOVA output to judge the significance of main effects; the pattern of means must also be considered because a statistically significant effect can lack substantial meaning when a disordinal interaction is present. When a disordinal interaction is present, the main effects are meaningless and should not be the basis of conclusions.

With the understanding of these three general aspects of ANOVA with more than one factor, we can start to examine the exemplary dataset of the Web-Appendix. When running the model in R (see Fig. 7), we find no effects of font type ($F(1,116) = 0.47$; $p = 0.49$) and brand name ($F(1,116) = 1.44$; $p = 0.23$) but a significant interaction between the factors ($F(1,116) = 32.05$; $p < 0.001$). Thus, in the present example, only the interaction is significant. However, even if one or both of the main effects were statistically significant, they would not be meaningful in the present scenario since we observe a disordinal interaction (see Fig. 8).

```
> Anova(aov(dv_fac ~ iv_fac_a * iv_fac_b, data=Data.Anova), type="III")
Anova Table (Type III tests)

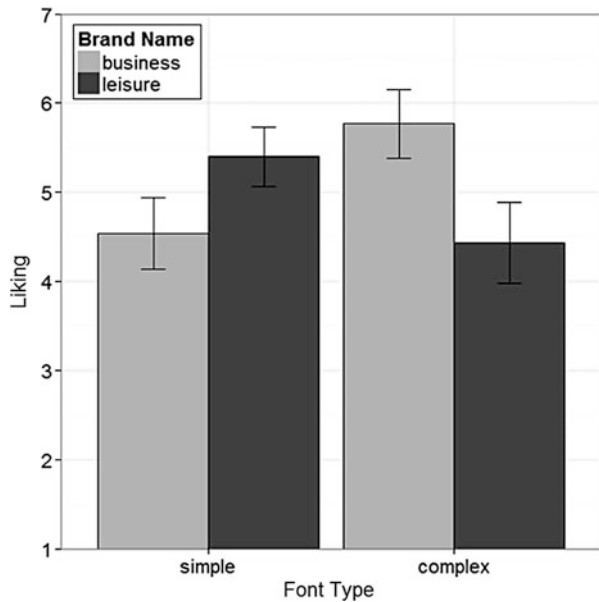
Response: dv_fac

      Sum Sq Df  F value    Pr(>F)
(Intercept) 3040.13  1 2683.8316 < 2.2e-16 ***
iv_fac_a      0.53  1   0.4708   0.4940
iv_fac_b      1.63  1   1.4419   0.2323
iv_fac_a:iv_fac_b 36.30  1  32.0457 1.107e-07 ***
Residuals    131.40 116

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 7 R output of a two-way ANOVA with two factor levels for each factor (Study 3) (R denotes the multiplicative interaction of two factors by “:”)

Fig. 8 Mean liking ratings of Study 3 with 95% confidence intervals



How can we interpret the results of an ANOVA with interaction(s)? A significant interaction indicates that the effect of at least one of the factors is dependent on at least one other factor. A disordinal interaction indicates that the effect of each factor is dependent on the other factor, that is, in the present example, any effect is conditional on the other effect. For a substantial interpretation of the interaction, one needs to know which conditional effects are significant. To this end, we can compute so-called “simple effects.” To introduce the idea of simple effects, let us first recap the types of effects we have encountered thus far using Table 7. The main effect of Factor A compares the mean of $[*{-1,-1}]^*$ and $^{\circ}\{1,-1\}^{\circ}$ with the mean of $[\wedge{-1,1}]^{\wedge}$ and $\#\{1,1\}\#$. Accordingly, the main effect of Factor B compares the mean of $[*{-1,-1}]^*$ and $[\wedge{-1,1}]^{\wedge}$ with the mean of $^{\circ}\{1,-1\}^{\circ}$ and $\#\{1,1\}\#$. The

interactive effect of Factors A and B tests whether the mean of [$\{ -1, -1 \}^*$ and $\{ 1, 1 \}^\#$] differs from the mean of [$\{ 1, -1 \}^\circ$ and $\{ -1, 1 \}^\wedge$].

In contrast, simple effects do not compare combinations of experimental cells but directly compare experimental cells for one factor level of the other factor. For the exemplary design in Table 7, four simple effects can be computed: the simple effect of Factor A conditional on Factor B’s “-1” level (i.e., $\{ -1, -1 \}^*$ vs. $\{ -1, 1 \}^\wedge$), the simple effect of Factor A conditional on Factor B’s “1” level (i.e., $\{ 1, -1 \}^\circ$ vs. $\{ 1, 1 \}^\#$), the simple effect of Factor B conditional on Factor A’s “-1” level (i.e., $\{ -1, -1 \}^*$ vs. $\{ 1, -1 \}^\circ$), and the simple effect of Factor B conditional on Factor A’s “1” level (i.e., $\{ -1, 1 \}^\wedge$ vs. $\{ 1, 1 \}^\#$).

How do we compute these simple effects? The attentive reader would probably suggest that the current chapter started with a situation where two means need to be compared. A *t*-test for independent samples or a one-way ANOVA applied to a subset of the dataset (i.e., just those experimental cells involved in the respective simple effect) appears to be a natural solution. Simple effects are, however, a bit more complicated because they make use of the information contained in all experimental cells (i.e., also those cells that are not involved in the respective simple effect). In particular, simple effects estimate the residual variance based on the full dataset and the model variance based on only the cells involved in the simple effect (Field 2013; Field et al. 2012). This approach increases the residual degrees of freedom for the evaluation of the empirical *F*-value and hence increases the statistical power of the simple effects. Thus, simple effects are a variant of ANOVA that are conducted as a follow-up analysis after a significant interaction has been observed. As with any follow-up technique in the ANOVA world (cf. post hoc contrasts), it is important to only run the follow-up analyses if, and only if, the omnibus test is significant. Therefore, simple effects are only computed if a significant interaction is present in the initial factorial ANOVA. Otherwise, an inflation of the alpha error is likely to occur, and the obtained results would be questionable.

Unfortunately, neither SPSS nor R offers a straightforward, convenient way of testing simple effects (Field 2013; Field et al. 2012). In SPSS, simple effects are not available from the menu but need to be requested by a self-written command in the syntax. In R, simple effects need to be computed by extracting the residual sum of squares and residual degrees of freedom from the initial factorial ANOVA and the model sum of squares and model degrees of freedom from a follow-up ANOVA applied to only the involved experimental cells. Own code must be written to apply the formulas of Fig. 2 to compute the empirical *F*-value and the corresponding

Table 7 Distinguishing main effects, interaction effects, and simple effects

		Factor (A)	
		-1	1
Factor (B)	-1	$\{ -1, -1 \}^*$	$\{ -1, 1 \}^\wedge$
	1	$\{ 1, -1 \}^\circ$	$\{ 1, 1 \}^\#$

Note: The symbols $\{ \cdot \}^* \wedge \circ \#$ are used to facilitate the recognition of the four different cells in the text

Table 8 Comparison of different data formats commonly used for RM-ANOVA

(a) Wide format					(b) Long format				
id	gender	age	dv_within_2_a	dv_within_2_b	id	gender	age	font.type	liking
1	male	28	2	1	1	male	28	simple	2
2	male	23	5	7	1	male	28	complex	1
3	male	20	5	7	2	male	23	simple	5
:	:	:	:	:	2	male	23	complex	7
120	female	23	4	6	3	male	20	simple	5
					3	male	20	complex	7
					:	:	:	:	:
					120	female	23	simple	4
					120	female	23	complex	6

p-value. Examples for how to compute simple effects in SPSS and R can be found in the Web-Appendix.

In the present example, the market researcher is most likely interested in the question of whether the effect of font type is significant in each of the brand name groups, that is, whether the difference between the two light gray bars in Fig. 8 is statistically significant and whether the difference between the two dark gray bars in Fig. 8 is statistically significant. Thus, we will focus on the two managerially most relevant simple effects described above (i.e., the effect of “font type” for the “business brand name” and the effect of “font type” for the “leisure brand name”) and will discard the managerially less relevant effects (i.e., the effect of brand name conditional on font type). The simple effects indicate that a complex font works better for a business brand name ($F(1, 116) = 20.14; p < 0.001$) and a simple font works better for a leisure brand name ($F(1, 116) = 12.37; p = 0.001$).⁹

Within-Subjects: Two or More Observations per Person

The standard ANOVA covered in the first part of this chapter requires that all observations are independent from each other and hence come from different individuals. However, there are situations where it is preferable to collect more than one measurement per person, which requires the application of a different type of ANOVA: repeated-measures ANOVA (RM-ANOVA). If a market researcher would like to compare the effectiveness of two advertisement campaigns, he or she is usually not only interested in the short-term effects but also in the long-term effects. To this end, one could survey the same participants immediately after viewing an advertisement and 1 month later. The time between these two measurements is a within-subjects factor. In the last dataset of this chapter (Study 6), we will consider an example of a longitudinal dataset, where time is the within-subjects factor. Another field of application for RM-ANOVA is the manipulation of a “normal”

⁹Please note that the residual degrees of freedom (i.e., 116) for the simple effects are the same as in the initial factorial ANOVA. This is the reason why simple effects have higher statistical power than other post hoc approaches that would just compare the two means, such as an independent-samples *t*-test.

experimental factor within-subjects instead of between-subjects. If there is no reason to expect that participants respond in a biased way (e.g., demand artifact¹⁰) once they know all factor levels of an experimental factor, it is statistically much more powerful (and hence more efficient) to manipulate an experimental factor within-subjects instead of between-subjects. We will consider examples of such situations in the next two sections (studies 4 and 5).

Two Means: One-Factorial RM-ANOVA or Paired-Samples *t*-Test

In the fourth study of the present chapter, we revisit the first between-subjects study of this chapter (Study 1), where “font type” was manipulated with two levels: simple vs. complex. If we assume that participants can provide an unbiased answer even if they know both levels of the factor “font type,” it would be more efficient to conduct a within-subjects manipulation of this factor because half the number of participants is sufficient to gather the same number of measurements as in the between-subjects scenario. In fact, a key advantage of a within-subjects experiment is that even less than half the number of participants is sufficient to reach a comparable level of power compared to a between-subjects scenario because the random differences between individuals are “pulled out” of the analysis. Let us consider how this is achieved using the dataset provided in the Web-Appendix. In the dataset, a sample of 120 participants rated the simple font (*dv_within_2_a*) and the complex font (*dv_within_2_b*), and each rating is stored in a separate variable (i.e., a separate column in the data matrix). This type of data storage is called “wide format.” We will see later that some types of analyses require a different type of storage called “long format,” where a second row instead of a second column is used to store the second measurement (see Table 8).

To illustrate the mechanics of RM-ANOVA, we again compare the *t*-test approach with the ANOVA approach.¹¹ A paired-samples *t*-test can be used to analyze repeated measures with a maximum of two measurements. The key idea of this type of *t*-test is that it is not the means per se that are analyzed but the differences between pairwise means within an individual. That is, the difference D_i is first computed per participant i by subtracting measurement₂ from measurement₁. Then, the mean of all D_i is computed and divided by its standard error to derive an empirical *t*-test statistic:

$$T_{\text{emp}} = \text{mean}(D_i) / \text{SE}(D_i) \quad (7)$$

¹⁰The term demand artifact indicates that participants guess the hypothesis of an experiment and demonstrate behavior that is consistent with their guess instead of their natural behavior. Therefore, the occurrence of a demand artifact destroys the external validity of the observed effects. Sawyer (1975) provides an excellent discussion of this problem and potential solutions.

¹¹A third possible approach would be an extension of the regression framework called linear mixed models (LMM; for an applied introduction, see West et al. 2015).

Following this approach, the absolute differences between participants are omitted from the analysis; therefore, this source of random noise does not reduce the precision of the estimated mean difference. This key idea of excluding the between-subjects variance when evaluating the significance of the within-subjects effect of interest is part of all repeated-measures techniques and is responsible for their high level of statistical power. The present example finds a significant difference ($D = -0.59$) between the simple ($M = 3.92$) and complex ($M = 4.51$) fonts ($t(119) = -4.34; p < 0.001$).

For illustrative purposes, we evaluate the same mean difference using RM-ANOVA. In practice, however, when only one factor with two levels is present, the paired-samples t -test would be the method of choice. The RM-ANOVA explicitly disentangles the within- and between-subjects variance and uses only the within-subjects portion to evaluate significance. This approach of splitting the total sum of squares into the relevant components is illustrated in Fig. 9.

As can be seen in Fig. 9, RM-ANOVA only considers the within-subjects variance to determine the statistical significance by splitting this source of variation into the two components we previously encountered in the normal between-subjects ANOVA model at the outset of this chapter (see Fig. 2): SS_{MW} and SS_{RW} . The between-subjects variance is, however, isolated and does not influence the estimation of the model’s statistical significance. Different software packages require different data formats to perform RM-ANOVA (see Table 8). SPSS requires the repeated measures to be stored in separate variables called “wide format,” as shown in Table 8a. In R, the required data format depends on the function used: `Anova` requires “wide format,” and `ezAnova` from the `ez` library (Lawrence 2015) requires “long format” (Table 8b) where each participant fills as many rows as there are measurements (in the present example, we have two repeated measures and hence two rows per person). The output shown in Fig. 10 is produced by running the `Anova` procedure on the wide-format data. The results of the RM-ANOVA replicate the findings of the paired-samples t -test by showing a significant effect of “font type” on liking ($F(1,119) = 18.87; p < 0.001$).

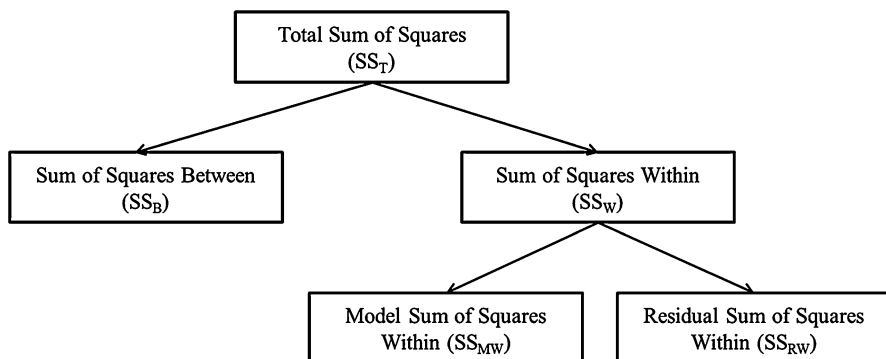


Fig. 9 Sources of variation in a RM-ANOVA

```

> summary(Anova(mod.Rm.2, idata=idata.Rm.2, idesign=~font.type, type="III"), multivariate=F)
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

          SS num Df Error SS den Df      F    Pr(>F)
(Intercept) 4258.8      1  898.66   119 563.951 < 2.2e-16 ***
font.type    21.0      1  132.50   119  18.865 2.968e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig. 10 R output of a one-way RM-ANOVA with two factor levels (Study 4)

More Than Two Means: One-Factorial RM-ANOVA

The key concept of RM-ANOVA introduced in the previous section can be extended to more than two within-subjects factor levels. Suppose we would like to repeat the second study described in the between-subjects part of this chapter using a within-subjects design, where each participant rates all three font types (simple, complex, and super-complex). In the exemplary dataset, 120 participants participated in this study (Study 5). Their answers are stored in the variables *dv_within_3_a* (= simple), *dv_within_3_b* (= complex), and *dv_within_3_c* (= super-complex). This type of dataset with one within-subjects factor with three levels cannot be analyzed by a paired-samples *t*-test; it requires RM-ANOVA.

The existence of more than two factor levels challenges a key assumption of RM-ANOVA called sphericity. Sphericity means that all the differences between pairwise repeated measures have equal variance. This issue arises when more than two repeated measures are involved, such as in Study 5. When the assumption of sphericity is harmed, the *F*-test is too liberal and needs to be corrected. Most statistical software packages (like SPSS and R) automatically check the sphericity assumption using Mauchly's test of sphericity. If the test is significant, the sphericity assumption is violated, and the *F*-test must be corrected using either the Greenhouse-Geisser (Greenhouse and Geisser 1959) or the Huynh-Feldt (Huynh and Feldt 1976) correction, which applies a correction factor to the degrees of freedom of the empirical *F*-value. These two approaches usually differ only slightly, and it is up to the researcher to report either the more conservative Greenhouse-Geisser or the slightly more liberal Huynh-Feldt correction. However, when Mauchly's test of sphericity is significant, one of the two corrections must be applied and reported.

For the present dataset, we observe the highest liking rating for the complex font ($M = 5.10$), followed by the simple ($M = 4.23$) and the super-complex font ($M = 4.03$). The RM-ANOVA produces a significant Mauchly's test ($p < 0.001$). Hence, we report the corrected degrees of freedom for the *F*-test. We use the Huynh-Feldt correction and observe a significant effect of font type on liking ($F(1.78, 211.47) = 25.39; p < 0.001$). The R output of this analysis is shown in Fig. 11. The degrees of freedom in the first output have been multiplied by the "HF eps" (i.e., Huynh-Feldt) correction factor (i.e., $2 \times 0.8885498 = 1.78$; $238 \times 0.8885498 = 211.47$). Post hoc LSD contrasts reveal that the complex font condition is significantly different from the other two ($p < 0.001$), but the simple and super-complex font conditions do not differ from each other ($p = 0.19$). Hence, the complex font is the best-liked option.

```

> summary(Anova(mod.Rm.3, idata=idata.Rm.3, idesign=~font.type, type="III", multivariate=F))

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

          SS num Df Error SS den Df      F  Pr(>F)
(Intercept) 7137.8    1 1181.86   119 718.69 < 2.2e-16 ***
font.type    77.6    2   363.73   238  25.39 1.012e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Mauchly Tests for Sphericity

          Test statistic    p-value
font.type      0.8591 0.00012837

Greenhouse-Geisser and Huynh-Feldt Corrections
for Departure from Sphericity

          GG eps Pr(>F[GG])
font.type 0.8765 1.107e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          HF eps  Pr(>F[HF])
font.type 0.8885498 8.764784e-10

```

Fig. 11 R output of a one-way RM-ANOVA with three factor levels (Study 5)

Multiplicative Effects: Factorial RM-ANOVA/Mixed-ANOVA

The RM-ANOVA approach introduced in the previous two sections is easily extended to two or more within-subjects factors. The key ideas of RM-ANOVA and the handling of interactions in an ANOVA framework have already been described; therefore, we will focus on a slightly more complex situation: the combination of between- and within-subjects factors within one analysis. If these two types of factors are combined within one analysis, the analysis is called either a mixed-ANOVA or a split-plot ANOVA (both terms can be used interchangeably). To illustrate this type of analysis, we revisit the study introduced at the outset of this chapter, in which the simple font was compared to the complex font (Study 1). Market researchers are usually interested not only in the immediate effects but also in the long-term effects of their marketing efforts. Suppose that the participants of the first study were instructed to provide an immediate liking judgment for the brand name printed in one of the fonts and to take a look at the brand name once per day for a 1-month period (Study 6). After the month is over, they are asked to provide a second liking judgment. It then becomes possible to examine the temporal stability of the liking judgment and to determine whether one of the fonts is more prone to habituation.

Technically, such a study has a 2 (between: font type) \times 2 (within: time) mixed factorial design. In the exemplary dataset, the between-subjects factor is named *iv_2*, the immediate liking judgment is *dv_2*, and the follow-up liking judgment is *dv_2_within*. A total of 120 participants were randomly assigned to the two between-subjects conditions (60 per factor level). To understand the mechanics of

a mixed-ANOVA, we first look at the decomposition of the total sum of squares provided in Fig. 12.

The key concept of RM-ANOVA is also applied to the mixed-ANOVA: The residual variance is decomposed into a between- and a within-subjects part, and each effect is only tested against the relevant part of the residual variance. In particular, the main effect of the between-subjects factor (i.e., font type) SS_{MB} is tested against the between-subjects residual variance SS_{RB} . The main effect of the within-subjects factor (i.e., time) SS_{MW} is tested against the within-subjects residual variance SS_{RW} . Finally, the interaction of the between- and the within-subjects factor SS_{MB*w} is only tested against the within-subjects residual variance SS_{RW} . This testing procedure is evident from the output of the described mixed-ANOVA provided in Fig. 13 (note the column denoted by “Error SS”).

The results of the mixed-ANOVA reveal that “font type” ($F(1, 118) = 42.45; p < 0.001$), time ($F(1, 118) = 27.80; p < 0.001$), and the interaction of the factors ($F(1, 118) = 26.04; p < 0.001$) all have a significant influence on liking. The barplot of the means provided in Fig. 14 shows that the interaction is caused by the complex font being equally liked immediately after seeing it the first time and 1 month later,

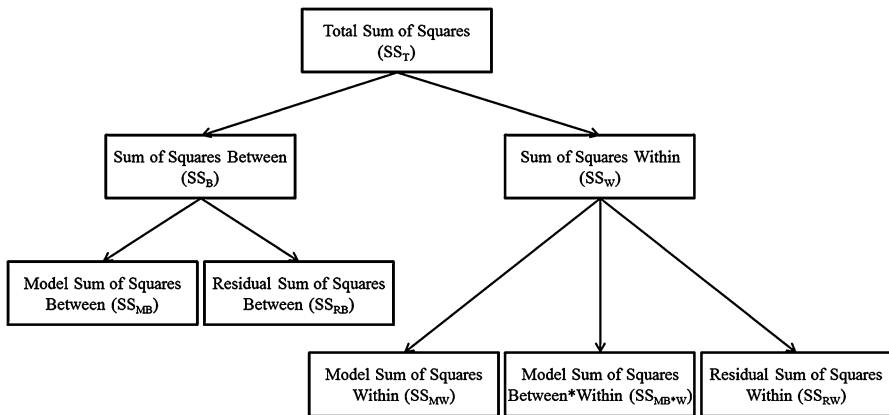


Fig. 12 Sources of variation in a mixed-ANOVA

```

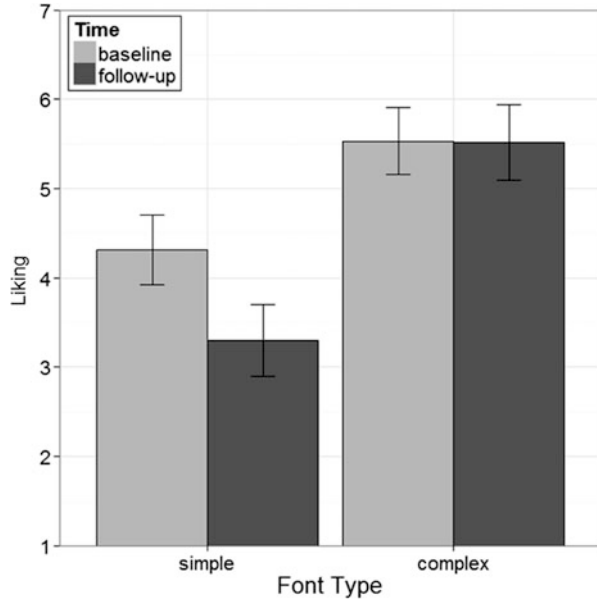
> summary(Anova(mod.Mixed, idata=idata.Mixed, idesign="--Time, type="III"), multivariate=F)

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

              SS num Df Error SS den Df      F      Pr(>F)
(Intercept)   5226.7    1  491.52   118 1254.783 < 2.2e-16 ***
Data.AnovaSiv_2  176.8    1  491.52   118  42.449 1.864e-09 ***
Time           16.0     1   67.98   118  27.800 6.166e-07 ***
Data.AnovaSiv_2:Time  15.0    1   67.98   118  26.036 1.296e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Fig. 13 R output of a mixed-ANOVA with two factors (Study 6)

Fig. 14 Mean liking ratings of Study 6 with 95% confidence intervals



whereas the liking of the simple font decreases over time, which is arguably due to boredom.

Extensions

The standard (RM-)ANOVA approach described thus far is limited to discrete independent variables (i.e., predictors) and to a single dependent variable (i.e., outcome). We will now cover two variants of the standard approach that extend the ANOVA framework to continuous predictors (ANCOVA) and to more than one dependent variable (MANOVA).

Analysis of Covariance (ANCOVA)

Analysis of covariance (ANCOVA) refers to an ANOVA model in which at least one continuous predictor variable is included. The key idea of ANCOVA is to include continuous control variables that are independent of the experimental factors but are related to the dependent variable. Such a control variable reduces the residual variance of the model and hence increases the likelihood of discovering a significant effect of the experimental factors. Before conducting ANCOVA, it is essential to ensure that the covariates are not correlated with the experimental factors (see Miller and Chapman 2001 for more details on this issue). If a correlation is present, the effects of the experimental factors are difficult to interpret since the covariate is

confounded with the experimental effect. To avoid such a correlation, researchers should either use only trait/person variables (e.g., age, education) rather than state variables (e.g., mood, motivation) as covariates or if a state variable is really needed, it must be measured before the experimental manipulation occurs. Moreover, researchers should refrain from including too many covariates in the ANCOVA model. A covariate can only reduce the residual variance if it is highly correlated with the dependent variable and not correlated with the experimental factors. Furthermore, the included covariates should not be correlated with each other. Usually, it is difficult to find a lot of variables that fulfill these criteria. Therefore, ANCOVA often employs a single covariate and rarely more than two to three covariates.

As an example of the application of ANCOVA, consider the one-way ANOVA study at the outset of this chapter. It is reasonable to assume that the general liking of a business bicycle brand differs between people, independent of the employed font. For instance, participants' age could influence the general liking, such that the general preference for a business bicycle brand increases with age. If age is added to the first ANOVA model of this chapter (Study 1) as a covariate, the model becomes an ANCOVA. Thereby, the residual variance can be reduced. The Web-Appendix contains an exemplary application of such an ANCOVA. Please note that in R, the covariate needs to be mean centered (i.e., $x_i - \text{mean}(x)$ for each individual observation i of variable x) before it can be included in the ANCOVA. In SPSS, mean centering is automatically executed by the ANCOVA procedure.

Multivariate Analysis of Variance (MANOVA)

Suppose that in this chapter's first study, participants were not only asked to provide a liking judgment (variable name in the dataset, *dv_2*) but also a willingness-to-buy judgment (*dv_2_manova*). In this situation, there is more than one dependent variable that could be analyzed by ANOVA. One could be tempted to run two univariate ANOVAs sequentially for the two variables. However, it is likely that these two variables are correlated with each other, and this information would be lost when running two separate analyses. Furthermore, as previously discussed in the context of post hoc contrasts, this approach would be an example of multiple statistical testing using the same data, which leads to an inflation of Type I error (i.e., alpha inflation).

There are two potential solutions to this problem. First, if the two (or more) variables are statistically highly correlated and, based on their theoretical meaning, are highly related, it would be advisable to treat them as items of an overarching scale and to aggregate the scores to form a single value (i.e., the mean of the two or more scores). This approach follows the logic of latent constructs that are measured by multiple indicators. Then, one can simply run a univariate ANOVA on the aggregated score and follow the steps described thus far.

Second, if the two variables measure theoretically different concepts that cannot be summarized by one overarching construct, one would perform MANOVA. The

key idea of MANOVA is that the dependent variables are not analyzed in isolation but a linear combination of all dependent variables (called a “variate”) serves as the outcome to be analyzed. Moreover, MANOVA accounts for the correlation between the dependent variables using information present in the error terms. A significant effect in MANOVA means that at least one experimental group differs from one other group on at least one of the considered dependent variables. To determine which group mean/means differs/differ, one usually conducts separate univariate ANOVAs and respective post hoc tests. In this sense, MANOVA serves as a global omnibus test that guards against alpha inflation. The follow-up analyses are only conducted if the MANOVA indicates significant effects; otherwise, the analysis stops. The Web-Appendix contains an exemplary application of MANOVA.

Conclusion

Managers often need to make discrete decisions: Should I run advertisement A or B? Should I launch product A or B? Should I position my brand as A or B? These types of decisions require marketing research to evaluate the effect of a discrete predictor (i.e., experimental factor) on a continuous outcome variable (such as consumers’ liking, willingness to buy, or willingness to pay). ANOVA is the method of choice to analyze such datasets because it determines whether there are differences in the means of different groups of observations. The data entered into an ANOVA are usually produced through experimental research, which has major advantages in terms of causal interpretation of effects. Since ANOVA provides answers to some of the most important types of questions (marketing) managers have in mind, it is not surprising that ANOVA is one of the most important techniques in marketing research. Accordingly, a comprehensive review of articles published in one of the world-leading scientific marketing journals, the *Journal of Marketing Research*, found that ANOVA is the most frequently used statistical technique across all papers published by the journal (Malhotra et al. 1999).

The present book chapter was intended to provide an applied introduction to this important statistical technique. Instead of covering all statistical details and all the different options to run the analysis (see Field 2013 and Field et al. 2012 for a more detailed introduction to the different types of ANOVA), the key intention of this book chapter was to provide clear guidance for how to apply this technique without making mistakes. To this end, the book chapter is accompanied by a comprehensive Web-Appendix covering an exemplary dataset and complete scripts of commands for R and SPSS to conduct all the analyses covered in this chapter. These scripts can be used as blueprints to conduct these analyses on the readers’ own datasets. To close this chapter, I would like to summarize some recommendations for applying ANOVA:

- *Coding of factors*: The main effects of ANOVA are only meaningful when effect coding (i.e., -1 vs. 1) instead of dummy coding (i.e., 0 vs. 1) is implemented for the experimental factors because only effect coding evaluates the main effects

relative to the grand mean (see Table 7 and the corresponding explanations). SPSS uses effect coding by default for ANOVA models, and the user does not need to consider this issue. However, R uses dummy coding by default, and the user needs to actively specify effect coding to make the main effects in ANOVA models meaningful (see lines 24–41 of the Web-Appendix R Script on how to specify effect coding in R). This is a serious issue because the type of coding has a substantial impact on the estimated effects. Incorrect coding can lead to completely wrong inferences.

- *Sum of squares (SS)*: As long as only one experimental factor is involved or if all experimental cells contain exactly the same number of observations, one does not have to consider the different ways of computing the SS. If these conditions do not hold, Type III SS is highly recommended. Type I SS suffers from the fact that the order in which the factors are entered into the model determines how much of the explained variance is ascribed to the factors (the first factor gets more than the later factors). However, there is rarely a meaningful order for the factors. SPSS uses Type III SS by default. However, the `aoV` function in R uses Type I SS, and the `Anova` function is required to estimate Type III SS. It is highly recommended to change the default when using R and to base the analysis on Type III SS (see, e.g., line 156 of the Web-Appendix R Script on how to change to Type III SS).
- *Linear mixed models (LMMs)*: Recent discussions on the sphericity assumption and the treatment of missing values in RM-ANOVA designs have led researchers to conclude that in situations where three or more within-factor levels and/or missing values are present, an alternative analytical technique is superior: LMMs (see West et al. 2015 for an applied introduction). LMMs are an extension of the regression framework and were developed to model any type of multilevel data. Repeated measures on the same participants are one potential field of application for this type of model. The key idea of this class of models when applied to repeated measures is that the between-subjects variance is treated as a random effect. As in RM-ANOVA, this part of the variation is excluded from the evaluation of the experimental effects. However, LMMs can also explicitly model deviations from the sphericity assumption and can easily handle missing values within participants (in contrast to RM-ANOVA). However, LMMs are theoretically and computationally more difficult to understand and more difficult to apply than RM-ANOVA. Nevertheless, LMMs have become increasingly popular in many fields. Recent research (e.g., Westfall et al. 2014) recommends the use of such models for any ANOVA procedure where random stimulus replicates are used within participants.¹²
- *Effect size*: Scientists and practitioners are not only interested in the existence of an effect but in addition also in the magnitude of the effect. For large datasets, very small effects can be statistically significant without having any practical

¹²For example, when the effect of funny vs. rational advertisement is examined, one usually shows several funny and several rational advertisements and compares the aggregated mean evaluations. The random variation between advertisements can be controlled by LMM.

meaning. To quantify the magnitude of an effect, it has become good scientific practice to report the effect size in addition to F - and p -values (Lakens 2013).¹³ In the context of ANOVA, the most common effect size measure is the partial eta squared, reported as η_p^2 . For a Factor A, it is defined as $SS_A / (SS_A + SS_{\text{Residual}})$. Hence, it measures a factor's share of the variance not explained by other factors in the model. According to Cohen (1988, pp. 285–288), a value of 0.01 is regarded as a small effect, a value of 0.06 is regarded as a medium effect, and a value of 0.14 and larger is regarded as a large effect (see also Richardson 2011). The Web-Appendix contains an exemplary computation of this effect size measure using the `BaylorEdPsych` library in R (Beaujean 2012) and the option menu of SPSS.

- *Extensions of ANOVA*: The present chapter exclusively focused on the application of ANOVA to experimental data. We learned that by comparing two portions of variance (systematic vs. unsystematic), ANOVA can determine whether the means of experimental groups significantly differ. The basic principle of ANOVA can, however, be extended to any situation in statistical data analysis where the sizes of variances need to be compared. For instance, many statistical techniques require homogeneity of variances across different groups of observations to fulfill the distributional assumptions of the technique. ANOVA itself makes the assumption that the error variance is homogenous across different experimental groups, which can be tested using Levene's test of variance homogeneity (Levene 1960). Interestingly, Levene's test is just a special case of ANOVA. Another example is the modeling of data using regression analysis. When applying regression, one often has to decide how many predictors should be included in the model. In particular, the question whether adding additional predictors to the model sufficiently increases the fit of the model needs to be answered. For this purpose, ANOVA can be used to compare the residual variances of a more parsimonious model and a model with additional predictors. In this context, ANOVA assesses whether the additional predictors significantly decrease the residual variance and should be included or discarded. Hence, ANOVA can also be used to support model selection in the context of regression analysis. The reader of this chapter should thus not be surprised to see applications of ANOVA to problems beyond the analysis of experimental data.
- *Statistical software*: The present chapter is accompanied by an exemplary dataset and command script files for SPSS and R. I selected these two software packages because SPSS is a widespread and easy-to-use statistical software, and the application of ANOVA models is well implemented in SPSS. By default, SPSS uses the correct coding of factors and computes the correct sum of squares, which prevents potential errors. SPSS also provides many helpful additional outputs by default, for instance, checks of the model assumptions. Hence, I would highly recommend SPSS for beginners in the application of ANOVA and for people who

¹³An excellent introduction to the use of effect size measures and a comparison of different approaches can be found in the referred article by Lakens (2013).

do not want to focus too much on technical issues in their daily research to make life easier and less error prone. The implementation of ANOVA in R is not user-friendly. As discussed above, the defaults in R with respect to factor coding and sum of squares lead to potentially wrong results. Applying RM-ANOVA and computing simple effects are very difficult in R and require computations by hand. I nevertheless chose R because it is a highly powerful statistical software with a steadily increasing number of users. Moreover, because the implementation of ANOVA in R is so difficult and error prone, I found it particularly important to equip the interested reader with the required knowledge to prevent mistakes when using R for ANOVA. I hope that the exemplary R code will help readers to build their own error-free ANOVA models and that it makes applying ANOVA in R less challenging and more fun.

References

- Beaujean, A.A. (2012). *BaylorEdPsych: R package for Baylor University educational psychology quantitative courses*. R package version 0.5.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180.
- Field, A. (2013). *Discovering statistics using R* (4th ed.). Los Angeles: Sage.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles: Sage.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.
- Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Thousand Oaks: Sage.
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.
- Huynh, H., & Feldt, L. S. (1976). Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*(1), 69–82.
- Klockars, A. J., & Sax, G. (1986). *Multiple comparisons*. Newbury Park: Sage.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2013.00863>.
- Lawrence, M.A. (2015). *ez: Easy analysis and visualization of factorial experiments*. R package version 4.3.
- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin et al. (Eds.), *Contributions to probability and statistics* (pp. 278–292). Stanford: University Press.
- Malhotra, N. K., Peterson, M., & Kleiser, S. B. (1999). Marketing research: A state-of-the-art review and directions for the twenty-first century. *Journal of the Academy of Marketing Science*, *27*(2), 160–183.
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48.
- Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135–147.

- Rodger, R. S., & Roberts, M. (2013). Comparison of power for multiple comparison procedures. *Journal of Methods and Measurements in the Social Sciences*, 4(1), 20–47.
- Rutherford, A. (2001). *Introducing ANOVA and MANOVA: A GLM approach*. London: Sage.
- Sawyer, A. G. (1975). Demand artifacts in laboratory experiments in consumer research. *Journal of Consumer Research*, 1(4), 20–30.
- West, B. T., Welch, K. B., & Galecki, A. T. (2015). *Linear mixed models: A practical guide using statistical software* (2nd ed.). Boca Raton: Chapman & Hall.
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020–2045.
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer.



Regression Analysis

Bernd Skiera, Jochen Reiner, and Sönke Albers

Contents

Introduction	300
Statistical Explanation of the Method	300
Problem Statement	300
Objective Function and Estimation of Regression Coefficients	301
Goodness of Fit	304
Significance Testing	305
Standardization of Coefficients	306
Interpretation of Results	307
Results of Numerical Example	307
Assumptions	309
Procedure	310
Efficiency of Estimators	311
Test for Multicollinearity	311
Test for Autocorrelation	313
Test for Heteroscedasticity	315
Identification of Outliers	317
Transformation of Variables	318
Implications of the Analysis	320
Endogeneity	321
Further Topics	324
Software	325
Summary	326
References	326

B. Skiera (✉) · J. Reiner
Goethe University Frankfurt, Frankfurt, Germany
e-mail: skiera@skiera.de; jreiner@wiwi.uni-frankfurt.de

S. Albers
Kuehne Logistics University, Hamburg, Germany
e-mail: soenke.albers@the-klu.org

Abstract

Linear regression analysis is one of the most important statistical methods. It examines the linear relationship between a metric-scaled dependent variable (also called endogenous, explained, response, or predicted variable) and one or more metric-scaled independent variables (also called exogenous, explanatory, control, or predictor variable). We illustrate how regression analysis work and how it supports marketing decisions, e.g., the derivation of an optimal marketing mix. We also outline how to use linear regression analysis to estimate nonlinear functions such as a multiplicative sales response function. Furthermore, we show how to use the results of a regression to calculate elasticities and to identify outliers and discuss in details the problems that occur in case of autocorrelation, multicollinearity and heteroscedasticity. We use a numerical example to illustrate in detail all calculations and use this numerical example to outline the problems that occur in case of endogeneity.

Keywords

Regression analysis · Marketing mix modeling · Elasticities · Multicollinearity · Autocorrelation · Outlier detection · Endogeneity · Sales response function

Introduction

Linear regression analysis is one of the most important statistical methods. It examines the linear relationship between a metric-scaled dependent variable (also called endogenous, explained, response, or predicted variable) and one or more metric-scaled independent variables (also called exogenous, explanatory, control, or predictor variable). Often, the dependent variable describes the success of marketing and the independent variables the factors that explain this success. Variables that describe the success of marketing are either “hard” success factors such as profit, sales volume (here referred to as quantity), and market share or “soft” success factors such as customers’ attitudes, purchase intention, and satisfaction. Variables that influence this success are frequently marketing instruments such as price, product, distribution, and communication. To illustrate how regression analysis works, we focus here on a numerical example with a hard success factor, namely, the estimation of quantity (i.e., sales volume), which is explained by variables such as advertising budget, price, and number of salespersons.

Statistical Explanation of the Method**Problem Statement**

We illustrate the basic idea of linear regression analysis by applying it to data from a (fictitious) company (displayed in Table 1). Their analysis should address the following business problems:

Table 1 Distribution of quantity and marketing instruments across districts

District	Quantity	Salespersons	Price	Advertising	Number of mailings
1	81,996	7	49	228,753	7,106
2	91,735	5	46	370,062	4,733
3	70,830	4	50	297,909	3,734
4	101,192	6	45	271,884	6,152
5	78,319	6	51	299,919	5,734
6	105,369	7	47	367,644	6,640
7	68,564	3	47	241,362	3,115
8	95,523	7	46	244,575	6,859
9	88,834	7	49	296,100	6,905
10	89,511	5	46	372,498	5,142
11	107,836	6	45	359,511	6,196
12	83,310	7	50	324,837	6,801
13	67,817	4	50	288,303	3,965
14	59,207	6	54	289,470	5,830
15	81,410	6	52	363,501	6,124
16	71,431	3	46	361,974	2,509
17	119,000	3	45	250,000	–

- What is the optimal price?
- What is the optimal budget for advertising and the sales force?

The data in Table 1 shows quantity at one point in time for 16 districts that were randomly selected from a much larger number of districts in which the company operates. The districts differ only in terms of price, advertising budget, and salespersons. For the moment, we ignore district 17 and the marketing instrument “number of mailings.”

The company has variable costs per unit sold of \$30 and costs for each salesperson of \$120,000 per year. Moreover, we assume that thus far the company has managed its marketing activities naively and has not systematically selected the value for each of its marketing instruments in each district. With this assumption, we exclude the problem of endogeneity (we do address this topic subsequently).

Objective Function and Estimation of Regression Coefficients

Linear regression analysis investigates the effect of metric-scaled independent variables (here, person, price, and advertising budget) on a metric-scaled dependent variable (here, quantity). We consider each district as one observation. The corresponding regression equation for the linear regression analysis is as follows:

$$y_i = b_0 + \sum_{k \in K} b_k \cdot x_{i,k} + e_i \quad (i \in I), \quad (1)$$

where

y_i = value of the i^{th} observation for the dependent variable,

b_0 = intercept of the regression,

b_k = regression coefficient capturing the influence of the k^{th} independent variable,

$x_{i,k}$ = value of the i^{th} observation for the k^{th} independent variable,

e_i = residual of the i^{th} observation,

K = index set of the independent variables, and

I = index set of the observations.

We can observe values of the dependent variable y_i and independent variables $x_{i,k}$ (see the respective values in Table 1), but we need to estimate the coefficients of the regression equation (Eq. 1), b_0 and b_k ($k \in K$), and the resulting residuals e_i ($i \in I$) (sometimes also referred to as “disturbance” or “error term”). The residual e_i describes the deviation between the observed value of the dependent variable y_i for the i^{th} observation and the predicted value for the dependent variable \hat{y}_i . The predicted value of the dependent variable is based on the estimated regression coefficients b_k and the respective observed independent variables $x_{i,k}$, as illustrated in Eq. 2:

$$\hat{y}_i = b_0 + \sum_{k \in K} b_k \cdot x_{i,k} \quad (i \in I). \quad (2)$$

The upper part of Fig. 1 illustrates the deviations of the observed values from the predicted values (for quantity, depending on advertising budget), with the predicted values lying on the regression line.

The aim of linear regression analysis is to estimate the coefficients of the regression equation b_0 and b_k ($k \in K$) so that the sum of the squared residuals (i.e., the sum over all squared differences between the observed values of the i^{th} observation of y_i and the corresponding predicted values \hat{y}_i) is minimized. The lower part of Fig. 1 illustrates this approach, which is called the “least squares method” (Stock and Watson 2015, p. 162). Simply speaking, the least squares method aims to minimize the sum of all rectangular areas displayed in the lower part of Fig. 1. The approach results in the following objective function:

$$\sum_{i \in I} e_i^2 = \sum_{i \in I} (y_i - \hat{y}_i)^2 = \sum_{i \in I} \left(y_i - b_0 - \sum_{k \in K} b_k \cdot x_{i,k} \right)^2 \rightarrow \min! \quad (3)$$

Using squared residuals is advantageous in that larger residuals receive a greater weight than smaller residuals and the solution for the objective function (Eq. 3) is algorithmically easy to determine (Pindyck and Rubinfeld 1998, p. 5). We do not derive the solution for the objective function (Eq. 3) here because a wide range of software (such as those presented in section “[Software](#)”) can perform this operation.

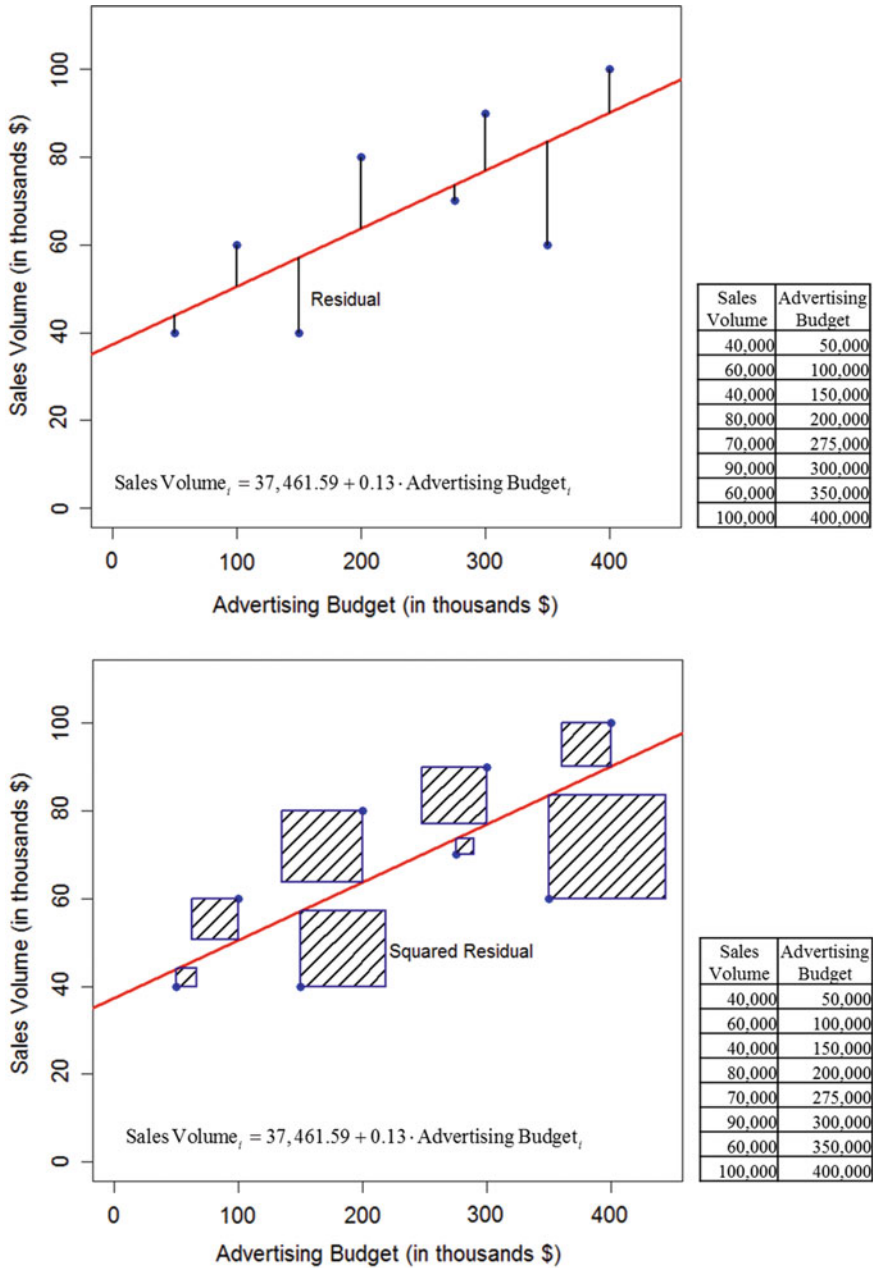


Fig. 1 Regression line with residuals and squared residuals

For more details on deriving the solution, see Wooldridge (2009, p. 27) and Gujarati (2003, p. 58), among many others.

Goodness of Fit

The assessment of goodness of fit for a linear regression analysis builds on the idea that the best estimate of the predicted value of a dependent variable is its mean value if nothing is known about the independent variables. Thus, a linear regression analysis is only meaningful if it can explain deviations from this mean value. More precisely, we determine goodness of fit by evaluating the extent to which considering the independent variables improves the “simple” prediction of just taking the mean value of the dependent variable. The coefficient of determination, called R^2 (see Eq. 4), measures goodness of fit by calculating the proportion of explained variance from the regression equation compared with the variance indicated by simply taking the mean value \bar{y} :

$$R^2 = \frac{\sum_{i \in I} (\hat{y}_i - \bar{y})^2}{\sum_{i \in I} (y_i - \bar{y})^2} \quad (4)$$

Thus, the linear regression analysis always explains at least the variance of the simple approach in that even if all independent variables have no explanatory power (the extreme worst case), the intercept captures the mean value of the dependent variable; in other words, the linear regression analysis reduces to simply predicting the mean value. Fortunately, adding one or more independent variables frequently improves this prediction. At the other extreme, the improvement in prediction might be so large that the predicted values of the dependent variable exactly match their observed values. In this case, the regression equation explains the total variance of the simple approach. R^2 can therefore take values between 0% and 100%. Negative values for R^2 can only arise when the researcher estimates the regression equation without an intercept, which is possible with most statistical programs; however, without an intercept, the linear regression analysis does not necessarily explain the variance of the simple approach.

These considerations should also make clear that including an additional independent variable never leads to a reduction of R^2 , because the explanatory power of the additional variable is at least zero, even if that variable explains nothing. In the extreme case that the number of observations corresponds to the number of estimated coefficients (more precisely, the intercept b_0 plus the number of regression coefficients b_k), we would essentially estimate a linear system of equations. The estimation of such a linear system always results in $R^2 = 100\%$, its largest possible value. To account for the fact that R^2 can only increase and does not penalize the use of variables that have no explanatory power, researchers frequently use the adjusted R^2 (R^2_{adj}), defined as follows (Wooldridge 2009, p. 200):

$$R^2_{adj} = R^2 - \frac{|K| \cdot (1 - R^2)}{|I| - |K| - 1}, \quad (5)$$

where

$|I|$ = number of elements in the index set of observations (equivalent to the number of observations) and

$|K|$ = number of elements in the index set of independent variables (equivalent to the number of regression coefficients).

Note that both numerator and denominator of the fraction are positive; therefore, the adjusted R^2 has at best the same value as the (unadjusted) R^2 . Large differences indicate that the regression equation contains many independent variables that do not explain the dependent variable (i.e., they have no explanatory power).

Significance Testing

Data are usually available only for a sample and not for an entire population, whether in the context of a large amount of surveys carried out by market research companies or data collected via panels and laboratory or field experiments. In the same vein, the presented 16 districts represent a (randomly selected) sample. In such a situation, it is necessary to determine the extent to which the results drawn from the sample also apply for the entire population. To do so, we must make an assumption about the distribution of the residuals.

For regression analysis (and many other methods), it is common to assume a normal distribution for the residuals. This assumption is based on the central limit theorem, which states that the mean from a random sample for any population (with finite variance), when standardized, has an asymptotic standard normal distribution (Wooldridge 2009, p. 758).

Using this assumption, we can conduct significance tests for the regression coefficients by calculating the error probability (often abbreviated as p-value) that the regression coefficients are nonzero. The significance test compares the determined error probability with a predetermined level of significance (often 5%). If the error probability is less than this predetermined level of significance, then the effect is considered significant.

To test the overall significance of a regression – that is, that all regression coefficients of the independent variables are 0 ($b_1 = b_2 = \dots = b_k = 0$) – we conduct the following F-Test with $|K|$ and $(|I| - |K| - 1)$ degrees of freedom:

$$F_{emp} = \frac{\frac{R^2}{|K|}}{\frac{1 - R^2}{|I| - |K| - 1}}. \quad (6)$$

If the F-value (F_{emp}) computed from Eq. 6 exceeds the critical F-value from the F table at a predetermined level of significance (often, as mentioned above, 5%),

the researcher can reject the null hypothesis that all independent variables have no effect on the dependent variable. Alternatively, many statistical software programs immediately report the error probability (p-value) for the F-value. If the error probability is less than the predetermined level of significance, the researcher can also reject the null hypothesis.

Whereas the F-test verifies the existence of a significant relationship between all independent variables and the dependent variable, the t-test determines the significance of each individual regression coefficient separately. Thus, it tests whether the null hypothesis of the coefficient being equal to 0 cannot be rejected with a certain error probability, frequently 5%. The t-test has $(|I| - |K| - 1)$ degrees of freedom and is calculated as follows:

$$t_{k, emp} = \frac{b_k}{s_k} \quad (k \in K), \quad (7)$$

where

$t_{k, emp}$ = empirical t-value for the k^{th} regression coefficient and
 s_k = estimated standard error of the k^{th} regression coefficient.

If the empirical t-value exceeds the critical t-value (from the t-distribution) at the chosen level of significance or, alternatively, the error probability (p-value) is less than the predetermined level of significance, then the coefficient is not equal to 0, and the impact of the k^{th} independent variable is significant.

Standardization of Coefficients

Often, it is necessary to compare the importance of the independent variables against one another. However, directly comparing the regression coefficients is hardly meaningful because the independent variables usually have different orders of magnitude (e.g., in Table 1, the advertising budget is measured in dollar values and person is measured in number of salespersons). The standardized regression coefficients β_{k} allow for a meaningful comparison. They are calculated by multiplying the (unstandardized) regression coefficients b_k with the standard deviation σ_{xk} of the associated independent variable and dividing the result by the standard deviation of the dependent variable σ_y :

$$\beta_k = b_k \cdot \frac{\sigma_{xk}}{\sigma_y} \quad (k \in K). \quad (8)$$

If the dependent and independent variables are standardized before running the regression analysis, then the regression coefficients β_k and b_k will be the same. Thus, comparing the absolute values of all standardized regression coefficient β_k shows the importance of the influence of the individual independent variables: Higher absolute values indicate a stronger influence.

Note, however, that using the standardized regression coefficient is discouraged when the standard deviations of the independent variables can be influenced. For example, if a company varied its product price to a greater extent than its advertising, then the standardized regression coefficients will indicate a stronger influence of the price due to the high standard deviation of the price (see Eq. 8). Therefore, it is preferable to calculate the elasticities of the independent variables, which are dimensionless and measure the responsiveness of the dependent variable to a change in one of the independent variables. Stated differently, they measure the percentage change of the dependent variable that corresponds to a 1% change of the independent variable. In the case of a linear regression, the elasticity is defined as follows:

$$\varepsilon_{y, x_k} = \frac{\frac{\partial y}{y}}{\frac{\partial x_k}{x_k}} = \frac{\partial y}{\partial x_k} \cdot \frac{x_k}{y} = b_k \cdot \frac{x_k}{y} \quad (k \in K). \quad (9)$$

Although the elasticity derived from Eq. 9 varies with the value of the independent variables, the common approach to calculate the elasticity for the independent variable x_k is to use the mean values of the independent variable x_k and the dependent variable y .

Interpretation of Results

Substantive insights should guide the interpretation of the results, rather than simply statistical criteria. In other words, researchers should first consider whether the regression equation captures all relevant variables via a meaningful functional relationship and then examine whether the signs of the regression coefficients are plausible. For the company in our example, the regression coefficients for advertising and person should be positive and negative, respectively, with regard to price. Then, to assess the size of the regression coefficients, calculating elasticities is often helpful. For example, advertising elasticities with values greater than 1 usually make little sense. The same holds for price elasticities with absolute values smaller than 1. Next to evaluating the substantive criteria, researcher should inspect statistical criteria such as the R^2 , the overall significance of the regression equation (F-test), the significance of the regression coefficients (t-test), and the subsequent assumptions for the least squares method.

Results of Numerical Example

In this section, we use the software program *R* to estimate the regression equation (codes for *R* but also other software programs such as STATA and SPSS are available on the authors' website). Figure 2 displays the result for this linear regression analysis with three marketing instruments (person, price, and advertising budget as independent variables and quantity as the dependent variable).

```

Call:
lm(formula = Quantity ~ Person + Price + Advertising, data = regdata)

Residuals:
    Min       1Q   Median       3Q      Max
-7650.1 -3383.9   287.8  3211.0  5047.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210159.444  23729.909   8.856  0.000 ***
Person       6723.478   840.997   7.995  0.000 ***
Price      -3832.503   444.013  -8.632  0.000 ***
Advertising    0.069     0.024   2.903  0.013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4546 on 12 degrees of freedom
Multiple R-squared:  0.9189,    Adjusted R-squared:  0.8986
F-statistic: 45.29 on 3 and 12 DF,  p-value: 8.077e-07

-----
"standardized coefficients - beta"
      Person      Price      Advertising
      0.665      -0.725         0.242
-----

Elasticities, at average values (not provided by R directly)
      Person      Price      Advertising
      0.446      -2.206         0.257

```

Fig. 2 Results of the linear regression analysis

The number of observations (here 16) is equal to the sum of the degrees of freedom (here 12) and the number of estimated coefficients (here 4). The R^2 (“R-squared”) and the adjusted R^2 (“Adjusted R-squared”) have values of 91.9% and 89.9%, respectively. The error probabilities determined for the F-test (“p-value”) and the t-tests (column “Pr (> |t|)”) are lower than the predetermined significance level of 5%; thus, the data indicate that all three marketing instruments have a significant effect. For example, the value 0.013 in column “Pr (> |t|)” means that if we repeatedly draw samples from the entire population, the probability of not observing a relationship between the advertising budget and the quantity is 1.3%.

The column titled “Estimate” provides the values for the (rounded) regression coefficients; their standard errors are in the column “Std. Error.” In contrast to other software programs, R does not automatically calculate the standardized beta values; therefore, separate calculations were necessary (values shown below the regression results in Fig. 2). In line with our expectations, person and the advertising budget have a positive impact, and price has a negative impact on quantity. Yet, we cannot use these regression coefficients to determine the strength of the effects. Instead, we must rely on the standardized beta values, which indicates that price has a slightly higher impact than the salesperson and that both marketing instruments show a considerably higher impact than advertising. Calculating the elasticities using Eq. 9,

we find an elasticity of the salespersons of 0.45, a price elasticity of -2.21 , and an advertising elasticity of 0.26. All elasticities show plausible signs and proportions (see also the meta-analyses conducted by Hanssens et al. 1990 and Albers et al.'s 2010 as well as Assmus et al. 1984; Bijmolt et al. 2005; Lodish et al. 1995; Sethuraman et al. 2011; Tellis 1988).

Assumptions

Thus far, we have focused on the method and the results of the linear regression. Next, we continue by examining whether the underlying statistical assumptions for a linear regression are fulfilled. These assumptions relate to the residuals; the relationship between dependent and independent variables, between the independent variable and the residuals, or between the independent variables; or the number of observations.

Assumptions for the residuals:

- Normal distribution of the residuals e_i
- Expected value of zero for the residuals $E(e_i) = 0$
- No correlation between residuals and independent variables, i.e., $\text{corr}(x_{i,k}, e_i) = 0$
- Constant variance of the residuals (homoscedasticity), i.e., $E(e_i^2) = \sigma^2$
- No correlation between the residuals (missing autocorrelation), i.e., $E(e_i \cdot e_i') = 0$

These assumptions essentially mean that the residuals resulting from the regression equation do not depend on the size of the observed variables (homoscedasticity), any of the independent variables, or the other residuals, particularly not on the residual of the previous period, as often occurs in the presence of time series (autocorrelation). In our example, the residual values should therefore be independent of the values of the three marketing instruments.

Endogeneity refers to the important assumption of the regression analysis that there is no relationship between one of the independent variables and the residuals. This independence should be present in our example because the company was naive in its marketing, which means that the company randomly chose the respective values for the marketing instruments. Therefore, no correlation should exist between the residuals, which reflect high or low level of marketing success, and the marketing instruments.

In reality, this assumption is often not fulfilled, because companies do not set their marketing activities randomly but systematically. These situations suffer from endogeneity. A typical example for such a situation occurs if the company does more marketing in districts where the company expects to be particularly successful. In our regression, “particularly successful” means that the quantity is higher than expected so that residuals are positive. Consequently, the independent variables correlate with the residuals.

Assumptions regarding the relationship between dependent and independent variables

- Consideration of all relevant independent variables
- Linearity of the regression equation (i.e., functional form)

To derive meaningful conclusions, we must ensure that the regression equation includes all relevant independent variables to avoid risking that the regression coefficients reflect the impact of the missing (also called omitted) variables. Furthermore, we assume in our linear regression analysis a linear relationship between the independent and dependent variables.

Assumptions regarding the relationship between the independent variables

- No multicollinearity between the independent variables

The regression analysis assumes that there is no linear relationship between the independent variables, that is, that there is a lack of multicollinearity. If multicollinearity occurs, because, for example, a high correlation exists between two independent variables, then a problem will occur: The effect of neither of the two variables on the dependent variable can be clearly derived.

Number of observations

- Sufficient number of observations

We can only estimate the regression equation in Eq. 1 if a sufficient number of observations are available. More precisely, the number of observations must be at least as large as the number of estimated coefficients (intercept b_0 plus the total number of regression coefficients b_k). Yet detecting a significant influence requires that the number of observations is much larger than the number of estimated coefficients. It is difficult to come up with a general statement about the ratio of the number of observations and the number of estimated coefficients, because this ratio always depends on the characteristics of variables in the data set. However, it is advisable that the number of observations is three times, better five times, as large as the number of estimated coefficients.

Procedure

Next, we describe the procedure for conducting a linear regression. We start by discussing the efficiency of the estimators. Afterwards, we discuss the problems that occur in case of multicollinearity, autocorrelation, heteroscedasticity, and outliers. In addition, we describe the transformation of variables often required to create a linear model.

Efficiency of Estimators

The estimated regression coefficient is efficient if it (1) is unbiased and (2) has the least variance of all unbiased coefficients (Gujarati 2003, p. 79). However, the least squares method only yields efficient estimates if the assumptions presented in section “[Assumptions](#)” hold. If the assumptions do not hold, then we must use a different estimation method or an alternative specification of the regression equation.

When examining the assumptions of the least squares method, we recommend investigating for the presence of multicollinearity, autocorrelation, and heteroscedasticity. If a sufficiently large number of observations are available, then the assumption of a normal distribution of the residuals is of minor importance, as it is usually fulfilled because of the central limit theorem (Stock and Watson 2015, p. 96). Koutsoyiannis (1977, p. 197) notes that this assumption is even fulfilled when samples have only 10–20 observations. In addition, the absence of a normal distribution signifies only that the F- and t-tests (discussed in section “[Significance Testing](#)”) are not meaningfully applicable. The estimated regression coefficients are still unbiased (Koutsoyiannis 1977, p. 197).

Test for Multicollinearity

Multicollinearity occurs when the independent variables are mutually linearly dependent. Multicollinearity usually leads to high standard errors of the estimated coefficients such that it is difficult to interpret them adequately. Inspecting the correlation matrix helps detect multicollinearity in the form of a linear dependency between two independent variables. High correlation values, often greater than -0.5 and $+0.5$, indicate multicollinearity, particularly if there are few observations. In this case, it is necessary to test the degree of the multicollinearity problem. We recommend running additional regressions that leave out some of the highly correlated variables. If the regression coefficients of the kept variables change substantially, multicollinearity represents a serious problem.

Running several linear regressions can help diagnose multicollinearity in the form of a linear dependence between more than two independent variables. In each regression, we use one of the original independent variables as a dependent variable and keep the others as independent variables. The difference between 1 and the R^2 for such a regression is called “tolerance,” and the inverse of this difference, which most statistical software programs report, is the “variance inflation factor” (VIF value). We can only assume a linear independence of the variables if the R^2 values of these regressions are low, which means that tolerance values are high, i.e., close to 1 and VIF values are low. Lower tolerance and higher VIF values, in contrast, indicate problems with multicollinearity. Although VIF values greater than 10 clearly indicate multicollinearity, even values of greater than 3, particularly with smaller data sets, point to problems with multicollinearity (see Hair et al. 2014, p. 200; for additional tests to uncover multicollinearity, see, e.g., Leeflang et al. 2000, pp. 348, 357).

Correlation Matrix			
	Person	Price	Advertising
Person	*****	0.143	-0.080
Price	0.597	*****	-0.158
Advertising	0.767	0.559	*****
upper diagonal part contains correlation coefficient estimates lower diagonal part contains corresponding p-values			

VIF-Values			
	Person	Price	Advertising
	1.024	1.044	1.029

Fig. 3 Testing for multicollinearity

In our example, the correlation matrix (see Fig. 3) indicates consistently low correlations between the independent variables, which shows there are no linear dependencies between two independent variables present. In addition, the VIF values close to 1 indicate that no linear dependencies between several independent variables exist.

With regard to the three marketing instruments, multicollinearity is thus not a problem for our data. Problems begin to emerge, however, if we also consider mailings (see Table 1), because the number of mailings is highly correlated (0.989) with the number of salespersons. Adding mailings to the linear regression analysis changes the results for the salespersons completely (Fig. 4); they now exert a negative influence on the quantity, while the number of mailings is highly positive. The high VIF values for the variables person and mailings also indicate the presence of multicollinearity.

An econometric solution for multicollinearity is difficult, although ridge regression (see Leeflang et al. 2000, p. 360) and partial least squares (Hair et al. 2017) might help. Increasing the number of observations can also help, in particular if the added observations exhibit a lower degree of linear dependence (here, salespersons and mailings). Alternatively, we could combine the linear dependent variables into a single variable (e.g., using factor analysis). We could also eliminate one or several of these variables from the linear regression analysis. For example, eliminating the salespersons from the regression decreases the parameter for mailings to 6.627 and its elasticity to 0.432.

Usually, however, it is difficult to find a truly satisfying solution for the multicollinearity problem because the number of observations is typically fixed and deleting variables is not helpful because the effects of these variables are often of primary interest. Therefore, in our numerical example, we can only assess the joint effect of both variables but not the specific effect of each variable, salespersons and mailings, on quantity. However, this information would certainly be of interest for the marketing manager in this case because this knowledge would enable her to optimize the marketing mix. For better determining the influence, the marketing manager should vary the number of salespersons and mailings so that a high correlation no longer occurs, ideally in the form of a field experiment.

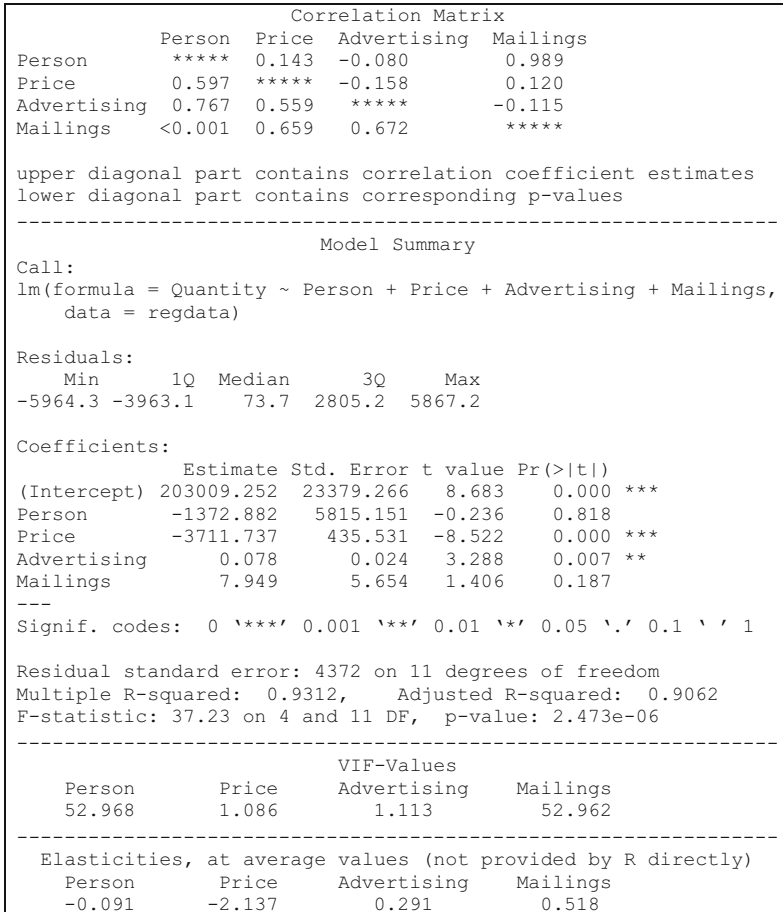


Fig. 4 Result of the regression analysis with multicollinearity problems

Test for Autocorrelation

Autocorrelation means that the residuals correlate with each other. Such correlation is most common in time series analyses, when the independent variables do not adequately cover the cyclical fluctuations in the time series. Autocorrelation usually leads to a situation where the predicted values are too high for some periods and too low for others. Thus, a series of negative residuals alternates with a series of positive residuals.

Autocorrelation causes an underestimation of the standard errors of the regression coefficients and, thus, an overestimation of the significance level of the t-test. (The standard error appears in the numerator of the t-test; see Eq. 7.) The estimated regression coefficients remain undistorted. However, they are no longer efficient because the standard error is not correctly detected (Wooldridge 2009, p. 408).

Most tests for autocorrelation examine the first-order autocorrelation, that is, the correlation between two consecutive residuals. In addition to graphical inspection of the residuals, the Durbin-Watson statistic (Wooldridge 2009, p. 415) is a common method to identify first-order autocorrelation:

$$dw = \frac{\sum_{i \in I'} (e_i - e_{i-1})^2}{\sum_{i \in I} e_i^2} \quad (10)$$

where

dw = value of the Durbin-Watson test and
 I' = index set of observations without the first observation.

If the difference between the two successive residuals is very small (large), positive (negative) autocorrelation is present, and the numerator in Eq. 10 takes small (large) values. The Durbin-Watson value dw approaches 0 (4) in this situation. A value of 2 indicates that no first-order autocorrelation is present. For an accurate representation of the level of significance for the Durbin-Watson test, refer to Wooldridge (2009, p. 415) or Gujarati (2003, p. 970). Also note that tests for autocorrelation often make little sense for cross-sectional data unless the order of observations follows a particular trend.

In time series, the value of the dependent variable of the previous period often can serve as an independent variable, typically called the “lagged variable.” In this case, the Durbin-Watson statistic is not suitable for the detection of autocorrelation, and the Durbin h test should be used (see, e.g., Gujarati 2003, p. 503).

In our numerical example, considering autocorrelation is meaningless because we use cross-sectional data, in which temporal correlation cannot be present. At most, a spatial autocorrelation might be present (see Gujarati 2003, p 442, for an explanation). Autocorrelation, however, can be a substantive problem because it is often an indication of missing (i.e., omitted) independent variables. Therefore, we illustrate the problem of autocorrelation with the numerical example in Table 2.

The columns with the x - and y -values describe the relationship between the x - and y -values in the following form:

$$y = 5 + 2 \times x. \quad (11)$$

The columns “True Error1” and “True Error2” provide the true residuals for two examples that illustrate first-order autocorrelation. These two residuals differ only by their signs. We compute the dependent variables y_1 and y_2 by adding the values of “True Error 1” and “True Error2” to the y -values, that is, $y_1 = y + \text{True Error1}$ and $y_2 = y + \text{True Error2}$.

Table 3 and Fig. 5 present the results of the linear regression analyses for the two numerical examples with the dependent variables y_1 and y_2 and the independent variable x . Note that the estimated residuals are systematically

Table 2 Numerical example to illustrate the problem of autocorrelation

Case	x	y	y1	y2	True Error1	True Error2
1	2	9	11	7	2	-2
2	3	11	15	7	4	-4
3	4	13	17	9	4	-4
4	5	15	17	13	2	-2
5	6	17	15	19	-2	2
6	7	19	15	23	-4	4
7	8	21	17	25	-4	4
8	9	23	21	25	-2	2

Table 3 Results of the linear regression analyses in the case of autocorrelation

	Regression 1 (with y1 as dep. var.)		Regression 2 (with y2 as dep. var.)	
	Intercept_1	Slope_1	Intercept_2	Slope_2
Value	11.286	0.857	-1.286	3.143
Standard error	1.882	0.316	1.882	0.316
Significance level t-test	0.00	0.03	0.52	0.00
R ²	0.55		0.94	
Significance level F-test	0.03		0.00	
Durbin-Watson value	1.27		1.27	

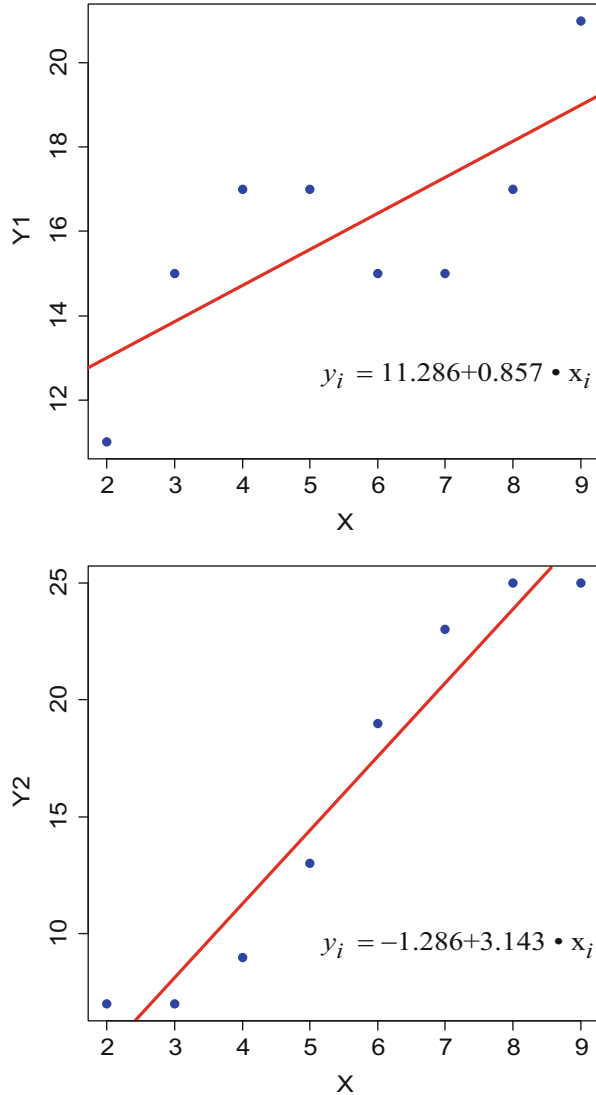
under- or overestimated in both examples and that the estimated regression equation does not reflect the actual functional relationship stated in Eq. 11. However, the estimated regression coefficients are still unbiased because the coefficients of Eq. 11 can be retained if we use a large number of random samples and run linear regression analyses on each of them. This result is also visible in Table 2 because the mean values of the constants and the regression coefficients of the two regression equations correspond to the coefficients in Eq. 11 ($\frac{11.29-1.29}{2} = 5$, respectively $\frac{0.86+3.14}{2} = 2$).

We can solve autocorrelation by detecting the factors responsible for the temporal fluctuations. Otherwise, we should consider econometric solutions such as the Cochrane-Orcutt procedure, the Hildreth-Lu procedure, or the Prais-Winsten estimate (Gujarati 2003, p. 482). These econometric methods, however, can only solve the problem if the regression equation is correctly specified.

Test for Heteroscedasticity

Heteroscedasticity means that the residuals do not have a constant variance. For example, the estimation of market shares for companies with high market shares might have a larger error than companies with small market shares (e.g., Gujarati 2003, p. 392). Heteroscedasticity leads to a situation in which the least squares method does not treat all observations equally and instead puts a greater emphasis on

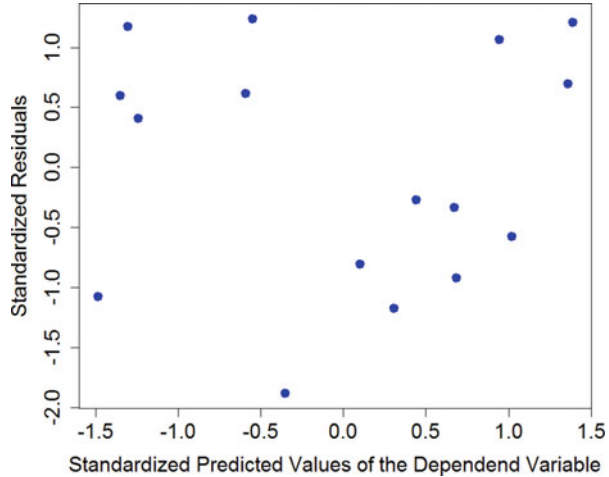
Fig. 5 Graphical representation of the regression equations in the case of the autocorrelation



the prediction of values with a high variance such that these observations receive a greater weight. The coefficients are still unbiased but are no longer efficient because they do not have the smallest estimation error (Stock and Watson 2015, p. 206).

We can detect heteroscedasticity by making a graphical comparison of the residuals with each of the independent variables or the dependent variable or by applying the Goldfeld-Quandt test, the Breusch-Pagan test, or the White test (Gujarati 2003, p. 400; Leeflang et al. 2000, p. 335). In our numerical example, the graphical comparison in Fig. 6 does not show a relationship between the standardized residuals and the standardized predicted values of the dependent variable, which indicates that heteroscedasticity is not present.

Fig. 6 Graphical inspection of heteroscedasticity



Unfortunately, heteroscedasticity is often not eliminated by collecting additional variables because the assumption of equal variances frequently does not make sense, as in our example of the estimation of market shares. It might help to transform variables in the regression equation (e.g., dividing the regression equation by the independent variable that causes the heteroscedasticity; Leeflang et al. 2000, p. 338). Alternatively, we can use a weighted linear regression analysis (“weighted least squares”) because it resolves the different weighting that results from heteroscedasticity (Gujarati 2003, pp. 415). When the data set consists of a large number of observations, calculating heteroscedastic robust standard error (Wooldridge 2009, p. 271) with procedures such as White robust standard errors (Gujarati 2003, pp. 439; White 1980) or the Newey-West estimator (Greene 2008, p. 643; Gujarati 2003, p. 439) also can be effective.

Identification of Outliers

The goal of linear regression analysis is usually that all observations have a comparable influence on the result; thus, it is optimal to avoid situations in which very few observations strongly influence the result. So-called outliers, whose values differ substantially from others in the data set, can cause such disproportional influence. Therefore, to examine for the presence of outliers, the easiest method is a visual inspection of the distribution of the observed values or the distribution of the residuals. In addition, a wide range of statistical methods are also effective. For example, the Mahalanobis distance is based on the standardized squared values of the independent variables, while Cook’s distance analyzes changes of the residuals that occur when the considered observation is removed from the regression equation. Chatterjee and Hadi (1986) provide a good overview of these and other statistical methods.

In our numerical example, the inspection of the residuals, the Mahalanobis distance, and the Cook’s distance do not show conspicuous values. However, district 17, which we have not yet used, has a much higher quantity than other districts. The

```

Model Summary

Call:
lm(formula = Quantity ~ Person + Price + Advertising, data =
regdata_wl7, x = TRUE, y = TRUE)

Residuals:
    Min       1Q   Median       3Q      Max
-14644  -5243  -1448   3443  29168

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 286804.698  51573.602   5.561  0.000 ***
Person       4294.625   1864.581   2.303  0.038 *
Price       -4710.844   1024.335  -4.599  0.000 ***
Advertising    0.009     0.054   0.159  0.876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10840 on 13 degrees of freedom
Multiple R-squared:  0.6377,    Adjusted R-squared:  0.5541
F-statistic: 7.628 on 3 and 13 DF,  p-value: 0.003423

```

Fig. 7 Results of the linear regression analysis in the presence of an outlier

values of Cook's distance and Mahalanobis distance indicate that district 17 may represent an outlier. Including district 17 in our regression analysis causes a strong change in the coefficients and the significance levels compared with the previous results (Fig. 7). The corresponding elasticities for the number of salespersons, price, and advertising also change significantly (0.27, -2.64 , and 0.03, respectively). In addition, the R^2 drops to 63.77%.

If the outliers are due to imputation errors, their values are easy to adjust. Otherwise, the solution is more difficult and depends largely on the question explored with the regression analysis. If the resulting recommendations should hold for all observations, eliminating outliers is not very satisfactory. Typically, however, recommendations are designed for the majority of the observations, so eliminating the outliers is appropriate.

In our example, we recommend eliminating the outlier (district 17) because the marketing manager is probably more interested in making recommendations that apply to the majority of the considered districts. However, at the same time, the marketing manager should carefully think about why district 17 is so different from the other districts. In any case, researchers should always carefully and precisely report which observations are considered outliers to prevent the appearance of data manipulation (Laurent 2013).

Transformation of Variables

When modeling quantity, it is frequently beneficial to estimate nonlinear relationships. The following equation is such a relationship that occurs in a multiplicative regression equation (i.e., a multiplicative sales response function):

$$Q = \alpha \cdot \text{Person}^\beta \cdot \text{Price}^\gamma \cdot \text{Advertising}^\delta \quad (12)$$

In contrast to the linear regression equation outlined in Eq. 1, the characteristics of this multiplicative regression equation are such that it captures interaction effects between the marketing instruments and varying marginal returns for each marketing instrument (see Gujarati 2003, p. 175 for a discussion of other functional forms).

It is still possible to use linear regression analysis to estimate such a nonlinear relationship, but to do so, transforming the variables so that a linear relationship between the independent variables and the dependent variable in the estimated regression equation is necessary. In the multiplicative regression equation, such a linear relation occurs if we take the logarithm of all variables:

$$\ln(Q) = \ln(\alpha) + \beta \cdot \ln(\text{Person}) + \gamma \cdot \ln(\text{Price}) + \delta \cdot \ln(\text{Advertising}). \quad (13)$$

The linear regression equation is then as follows:

$$\text{LN_}Q_i = a' + \beta \cdot \text{LN_}P_{\text{erson}_i} + \gamma \cdot \text{LN_}P_i + \delta \cdot \text{LN_}A_{\text{D}_i} + e_i \quad (i \in I), \quad (14)$$

where the variables LN_Q_i, LN_Person_i, LN_P_i, and LN_AD_i are defined as follows:

$$\begin{aligned} \text{LN_}Q_i &= \ln(Q_i), \\ \text{LN_}P_{\text{erson}_i} &= \ln(\text{Person}_i), \\ \text{LN_}P_i &= \ln(\text{Price}_i), \\ \text{LN_}A_{\text{D}_i} &= \ln(\text{Advertising}_i), \text{ and} \\ \alpha' &= \ln(\alpha). \end{aligned}$$

We again ignore district 17. Figure 8 displays the results. Although the R² and the F-value are high, the R² only describes the goodness of fit of the logarithmic model shown in Eq. 14 and not that of the initial multiplicative model (Eq. 13). Therefore, calculating the R² of the initial model using the estimated values of the regression coefficients in Eq. 13 presents an alternative. The significance levels of all variables are comparable to those of the linear model. An important advantage of a multiplicative regression equation is that the regression coefficients represent the elasticities of the respective marketing instruments (Gujarati 2003, p. 176), which makes interpreting the results easier.

The results also show that all marketing instruments have the expected signs and plausible values. Inserting the estimated regression coefficients of Eq. 14 into Eq. 13 yields the following (rounded) results for the multiplicative regression equation:

$$\begin{aligned} Q &= \exp(16.98) \cdot \text{Person}^{0.40} \cdot \text{Price}^{-2.34} \cdot \text{Advertising}^{0.22} \\ &= 23,676,653 \cdot \text{Person}^{0.40} \cdot \text{Price}^{-2.34} \cdot \text{Advertising}^{0.22} \end{aligned} \quad (15)$$

Due to its more plausible properties, the multiplicative regression equation is more popular than the linear regression equation (see, e.g., Tellis 1988).

```

Call:
lm(formula = log(Quantity) ~ log(Person) + log(Price)
    + log(Advertising), data = regdata)

Residuals:
    Min       1Q   Median       3Q      Max
-0.10317 -0.02581  0.01209  0.03504  0.07761

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    16.981      1.493   11.372   0.000 ***
log(Person)      0.400      0.047    8.565   0.000 ***
log(Price)     -2.339      0.248   -9.440   0.000 ***
log(Advertising)  0.217      0.082    2.647   0.021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05198 on 12 degrees of freedom
Multiple R-squared:  0.9276,    Adjusted R-squared:  0.9095
F-statistic: 51.23 on 3 and 12 DF, p-value: 4.098e-07

```

Fig. 8 Results for the linear regression of the multiplicative regression equation

Implications of the Analysis

In this section, using our numerical example, we determine the optimal price and the amount of money that should be spend on advertising and salespersons. To do so, we rely on the estimated multiplicative regression equation (i.e., sales response function) shown in Eq. 16. We also ignore the marketing instrument mailings.

$$\text{Quantity} = 23,676,653 \cdot \text{Person}^{0.40} \cdot \text{Price}^{-2.34} \cdot \text{Advertising}^{0.22} \quad (16)$$

The profit function builds on this sales response function. Costs per unit are \$30 and the costs per salespersons are \$120,000 per year. Thus, profit is:

$$\text{Profit} = (\text{Price} - 30) \cdot 23,676,653 \cdot \text{Person}^{0.40} \cdot \text{Price}^{-2.34} \cdot \text{Advertising}^{0.22} - 120,000 \cdot \text{Person} - \text{Advertising} \quad (17)$$

Its derivation leads to:

$$\frac{\partial \text{Profit}}{\partial \text{Price}} = 23,676,653 \cdot \text{Person}^{0.40} \cdot (-1.34) \cdot \text{Price}^{-2.34} \cdot \text{Advertising}^{0.22} - 30 \cdot 23,676,653 \cdot \text{Person}^{0.40} \cdot (-2.34) \cdot \text{Price}^{-3.34} \cdot \text{Advertising}^{0.22} \quad (18)$$

$$\frac{\partial \text{Profit}}{\partial \text{Person}} = (\text{Price} - 30) \cdot 23,676,653 \cdot 0.40 \cdot \text{Person}^{-0.60} \cdot \text{Price}^{-2.34} \cdot \text{Advertising}^{0.22} - 120,000 \quad (19)$$

$$\frac{\partial \text{Profit}}{\partial \text{Advertising}} = (\text{Price} - 30) \cdot 23,676,653 \cdot \text{Person}^{0.40} \cdot \text{Price}^{-2.34} \cdot 0.22 \cdot \text{Advertising}^{-0.78} - 1 \quad (20)$$

By setting the derivatives to 0 and solving for the three marketing instruments, we can determine the optimal price and spending for salespersons as well as advertising budget. For our example, we obtain the following results (calculation with unrounded values):

$$\begin{aligned} \text{Price}^* &= 3652.40 \\ \text{Person}^* &= 36603,070.53 \\ \text{Advertising}^* &= 36327,402.06. \end{aligned}$$

The optimal value leads to (again calculating with unrounded values) a quantity of 67,398 units, revenue of \$3,531,459.94, profit contribution (i.e., profit before considering the budget for advertising and salespersons) of \$1,509,505.85, and profit of \$579,033.27.

Endogeneity

Endogeneity is currently one of the most important topics in linear regression analysis. Endogeneity means that one or more of the independent variables correlate with the residuals and therefore influences the relationship between the dependent variable and the independent variable. Its existence leads to a systematic distortion of the estimated regression coefficient. It can occur for several reasons, such as an omitted independent variable, simultaneity in the variables, measurement error in an independent variable, autocorrelation with delayed dependent variable, or self-selection.

We use our numerical example to discuss this fundamental problem. To do so, we first compute the residuals resulting from the linear regression analysis and the regression coefficients that are depicted in Fig. 2. We then add a column with these residuals to Table 1 and sort all districts according to the size of the residuals, which are defined as the difference between the observed and predicted (also called estimated) value of the dependent variable ($e_i = y_i - \hat{y}_i$). Positive residuals represent districts in which the company was more successful than an “average” district. Table 4 presents the results. For ease of exposition, we ignore the variable mailings and the outlier district 17 for the subsequent considerations.

A fundamental assumption of the linear regression analysis is that the correlation between each of the independent variables and the residuals is equal to 0 ($\text{corr}(x_{i,k}, u_i) = 0$). In line with our initial assumption that the company’s marketing is naive (i.e., random), this assumption appears realistic.

Now, more realistically, we assume that the company knows in which districts it is particularly successful and also selects its marketing instruments in such a way

Table 4 Districts with sales information sorted by the residuals

District	Quantity	Salespersons	Price	Advertising	Residuals
15	81,410	6	52	363,501	5,047.79
11	107,836	6	45	359,511	4,922.35
3	70,830	4	50	297,909	4,788.31
4	101,192	6	45	271,884	4,341.61
6	105,369	7	47	367,644	2,834.12
5	78,319	6	51	299,919	2,523.77
13	67,817	4	50	288,303	2,439.98
7	68,564	3	47	241,362	1,660.99
9	88,834	7	49	296,100	-1,085.46
2	91,735	5	46	370,062	-1,352.74
8	95,523	7	46	244,575	-2,328.75
1	81,996	7	49	228,753	-3,263.45
10	89,511	5	46	372,498	-3,745.29
14	59,207	6	54	289,470	-4,367.71
12	83,310	7	50	324,837	-4,765.38
16	71,431	3	46	361,974	-7,650.14

that it intensifies its marketing activities in the districts where it is already particularly successful. Thus, the company sets the highest of its 16 advertising budgets in the districts with the highest (positive) residuals, the second highest advertising budget in the district with the second highest residuals, and so on. We assume that the regression coefficients and residuals (see the results in Fig. 2) still apply and consider them the true parameters and the true residuals. Of course, given the newly allocated advertising budgets, the quantities in each district change. Yet, the total quantity of all 16 districts and the total advertising budget remain unchanged. However, the correlation between the true residuals (see Table 5) and the advertising budget is now 0.965.

Running a linear regression analysis with the data presented in Table 5 now leads to the results depicted in Table 6 (in the row “Endogeneity for advertising”). The coefficient for advertising is now more than twice as high (0.149 vs. 0.069; see Table 6). The reason for this increase in value is that the coefficient now reflects its own effect and the effect of the systematic allocation of the advertising budgets to districts with higher quantity. The estimate of the regression coefficient is therefore inconsistent and, thus, biased.

Endogeneity is difficult to detect because we do not observe the true residuals. Stated differently, the true relationship between advertising budget and the true residuals displayed in Table 5 with a correlation of 0.965 is unknown. We only observe the residuals that result from the linear regression analysis with the observations displayed in Table 5. The correlation of these residuals with the advertising budget is, by definition, 0.

The analogous procedure for the other two marketing instruments, price and salespersons, results in similar effects (see Table 6). Here, we assign the lowest

Table 5 Quantity of modified allocation of advertising budgets across districts

District	Quantity	Salespersons	Price	Advertising (endogenous)	True residuals
15	82,032.54	6	52	372,498	5,047.79
11	108,566.07	6	45	370,062	4,922.35
3	75,655.24	4	50	367,644	4,788.31
4	107,531.34	6	45	363,501	4,341.61
6	104,976.67	7	47	361,974	2,834.12
5	82,442.41	6	51	359,511	2,523.77
13	70,344.93	4	50	324,837	2,439.98
7	72,615.79	3	47	299,919	1,660.99
9	88,959.17	7	49	297,909	-1,085.46
2	86,617.28	5	46	296,100	-1,352.74
8	98,629.46	7	46	289,470	-2,328.75
1	86,116.50	7	49	288,303	-3,263.45
10	82,549.12	5	46	271,884	-3,745.29
14	56,100.54	6	54	244,575	-4,367.71
12	77,534.03	7	50	241,362	-4,765.38
16	62,212.91	3	46	228,753	-7,650.14

True regression equation:

$$\text{Quantity} = 210,159.444 + 6723.478 * \text{Person} - 3,832.503 * \text{Price} + 0.069 * \text{Advertising} + \text{residual}$$

Table 6 Regression results when endogeneity for the variables advertising, price, and salespersons is present

		Advertising	Price	Salespersons	Intercept
Without	Coefficient	0.069	-3,832.503	6,723.478	210,159.444
Endogeneity	Standard error	0.024	4,44.013	840.997	23,729.909
	p-value	0.013	0.000	0.000	0.000
	R ²	0.919			
Endogeneity For advertising	Coefficient	0.149	-3,733.320	6,381.400	182,421.558
	Standard error	0.005	101.972	195.945	5245.791
	p-value	0.000	0.000	0.000	0.000
	R ²	0.996			
Endogeneity For price	Coefficient	0.068	-5,318.478	6,579.367	283,245.887
	Standard error	0.004	73.352	140.682	3917.548
	p-value	0.000	0.000	0.000	0.000
	R ²	0.999			
Endogeneity For salespersons	Coefficient	0.072	-3,679.197	9,460.590	186,682.532
	Standard error	0.008	144.168	272.228	8040.702
	p-value	0.000	0.000	0.000	0.000
	R ²	0.995			

prices and highest number of salespersons to the district with the highest residuals, and so forth. The correlation between the true residuals and price is -0.984 , and between true residuals and the number of salespersons is 0.940 . We still assume that the other two independent variables have the values shown in Table 4.

Again, it is clear that the size and significance level of the regression coefficients increase substantially when the values of the respective variables are set systematically, that is, according to the quantities of the districts. Ebbes et al. (chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#)) provide a more detailed discussion of solutions in such situations.

Further Topics

As mentioned at the beginning of this chapter, linear regression analysis aims to relate a metric-scaled dependent variable with one or more metric-scaled independent variables. In practice, however, researchers often aim to include nonmetric information – for example, discrete variables such as gender, income groups, or shopping locations – or nonlinear (e.g., U-shaped) relationships.

Dummy variables are frequently used to include discrete information. For representing k characteristics, we require $k - 1$ dummy variables. The most common type of coding is indicator coding, in which each of the $k - 1$ dummy variables is represented by 1 (present) or 0 (not present). The k^{th} characteristic that is not represented by 1 is then the reference characteristic. The estimated $k - 1$ coefficients represent the difference of the respective characteristic with respect to the reference characteristic. We can thus interpret the coefficients as the difference in the intercepts of the characteristics. An alternative approach for the coding of dummy variables is so-called effect coding (for more details, refer to Hair et al. 2014, p. 173).

It is also possible to include nonlinear curve-shaped relationships in the regression by using appropriate transformation of the respective variables (Albers 2012). An example of such a relationship is a diminishing marginal effect of a marketing instrument such as advertising. In many situations, several types of data transformations are appropriate for linearizing a curvilinear relationship (Hair et al. 2014, pp. 174–175). Examples of direct approaches are arithmetic transformations such as taking the square root or the logarithm of a variable. Another method is the use of polynomials, most often, only the first (x_i) and second order (x_i^2) of the independent variables. Polynomials can help represent complex relationships, thereby making it possible to interpret and statistically test each coefficient individually. The independent variables would then describe the overall effect of the polynomial.

The discussion of nonlinear relationships thus far reflects the relationship of the dependent variable and an independent variable; however, situations in which a second independent variable influences the effect of an independent variable on the dependent variable, or a moderating or an interaction effect, are also common. To represent such an interaction effect, it is necessary to create a new variable by multiplying the

two independent variables. Thus, if two variables $x_{i,1}$ and $x_{i,2}$ interact – in other words, the variable $x_{i,2}$ moderates the influences of variable $x_{i,1}$ (or vice versa) – then we add the product of the two variables to the regression equation. The coefficient of these two variables ($x_{i,1} \cdot x_{i,2}$), also called the moderator, reflects the change in the effect of the variable $x_{i,1}$ if the variable $x_{i,2}$ changes. For a further discussion of the misspecification and interpretation of interaction effects, refer to Irwin and McClelland (2001).

For our numerical examples in this chapter, we use either cross-sectional (Table 1) or longitudinal (Table 2) data. However, many marketing studies use panel data (also called pooled data) (Wooldridge 2009, p. 10), which combines longitudinal and cross-sectional data. Their analysis requires the consideration of specific issues (e.g., the consideration of structural differences between the individual cross-sectional data) that are beyond the scope of this chapter. For further details, refer to Hsiao (2014).

Furthermore, researchers should consider whether the data have a hierarchical structure. In that case, hierarchical or multilevel regressions are recommended. Not accounting for a hierarchical data structure in turn leads to an error in the estimation of the coefficients. For a good introduction to the topic of multilevel modeling, see Snijders and Bosker (2012).

Software

Software for estimating a regression analysis is available in many forms. Most spreadsheet programs (e.g., Microsoft Excel) enable the calculation of some simple analyses, but these programs are limited with respect to graphical representation of the results as well as the number of statistical and econometric tests. In addition, they offer only a limited range of functions to transform data easily, examine different variants of the regression analysis, detect heteroscedasticity and autocorrelation, identify outliers, and apply nonlinear regression analysis.

More sophisticated programs such as SPSS (now owned by IBM) allow for performing these analyses easily. Although SPSS is easy to use, it does not include all methods for the detection and management of autocorrelation and heteroscedasticity (e.g., the Durbin's h statistic). Other statistical programs such as SAS, STATA, EViews, and LIMDEP provide a higher functionality, although they are slightly less user-friendly. Furthermore, none of these programs are available free of charge. An open source program, PSPP, is intended as an alternative to IBM SPSS. In principle, matrix-oriented programs such as R, MATLAB, or Gauss also allow for applying all econometric methods because the user works directly with matrices. However, these programs are less easy to use and have a considerably greater learning curve. R's advantage is that it is open source and available free of charge. In summary, most spreadsheet programs are sufficient for the occasional computation of regression analysis, but for more detailed analyses, researchers should use one of the aforementioned statistical or matrix-oriented programs.

Summary

We use a numerical example to describe in detail one of the most important statistical methods, linear regression analysis. It examines the linear relationship between a dependent variable and one or more independent variables but can be easily used to also estimate nonlinear relationships if they can be linearized, as is the case for a multiplicative sales response functions (and many other functions). We use an extensive numerical example to illustrate all aspects that we cover in this section (autocorrelation, multicollinearity, heteroscedasticity, outlier detection, endogeneity, optimization of marketing mix), and we provide the code (for R, SPSS, and STATA) that we used to perform all calculations on our website (www.skiera.de). In addition, we also provide an Excel spreadsheet that contains all calculations, which should help in even better understanding of all calculations. Yet, we would like to highlight that Excel is not a good environment to conduct these calculations.

References

- Albers, S. (2012). Optimizable and implementable aggregate response modeling for marketing decision support. *International Journal of Research in Marketing*, 29(2), 111–122.
- Albers, S., Mantrala, M. K., & Sridhar, S. (2010). Personal selling elasticities: A meta-analysis. *Journal of Marketing Research*, 47(5), 840–853.
- Assmus, G., Farley, J. W., & Lehmann, D. R. (1984). How advertising affects sales: A meta-analysis of econometric results. *Journal of Marketing Research*, 21(1), 65–74.
- Bijmolt, T. H. A., van Heerde, H., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42(2), 141–156.
- Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regressions. *Statistical Science*, 1(3), 379–416.
- Greene, W. H. (2008). *Econometric analysis* (6th ed.). Upper Saddle River: Pearson.
- Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). New York: McGraw Hill.
- Hair, J. F., Black, W. C., Babin, J. B., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Upper Saddle River: Pearson.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Thousand Oaks: Sage.
- Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (1990). *Market response models: Econometric and time series analysis*. Boston: Springer.
- Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). Cambridge: Cambridge University Press.
- Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*, 38(1), 100–109.
- Koutsoyiannis, A. (1977). *Theory of econometrics* (2nd ed.). Houndmills: MacMillan.
- Laurent, G. (2013). EMAC distinguished marketing scholar 2012: Respect the data! *International Journal of Research in Marketing*, 30(4), 323–334.
- Leeflang, P. S. H., Wittink, D. R., Wedel, M., & Neart, P. A. (2000). *Building models for marketing decisions*. Berlin: Kluwer.
- Lodish, L. L., Abraham, M. M., Kalmenson, S., Livelsberger, J., Lubetkin, B., Richardson, B., & Stevens, M. E. (1995). How TV advertising works: A meta-analysis of 389 real world split cable T. V. advertising experiments. *Journal of Marketing Research*, 32(2), 125–139.
- Pindyck, R. S., & Rubinfeld, D. (1998). *Econometric models and econometric forecasts* (4th ed.). New York: McGraw-Hill.

- Sethuraman, R., Tellis, G. J., & Briesch, R. A. (2011). How well does advertising work? Generalizations from meta-analysis of brand advertising elasticities. *Journal of Marketing Research*, 48(3), 457–471.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage.
- Stock, J., & Watson, M. (2015). *Introduction to econometrics* (3rd ed.). Upper Saddle River: Pearson.
- Tellis, G. J. (1988). The price sensitivity of selective demand: A meta-analysis of econometric models of sales. *Journal of Marketing Research*, 25(4), 391–404.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838.
- Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th ed.). Mason: South-Western Cengage.



Logistic Regression and Discriminant Analysis

Sebastian Tillmanns and Manfred Krafft

Contents

Introduction	330
Discriminant Analysis	331
Foundations and Assumptions of Discriminant Analysis	331
Discriminant Analysis Procedure	333
Logistic Regression	343
Foundations and Assumptions of Logistic Regression	343
Logistic Regression Procedure	348
Applied Examples	357
Research Question and Sample, Model, Estimation, and Model Assessment	357
Discriminant Analysis	357
Logistic Regression	359
Conclusion	364
References	366

Abstract

Questions like whether a customer is going to buy a product (purchase vs. non-purchase) or whether a borrower is creditworthy (pay off debt vs. credit default) are typical in business practice and research. From a statistical perspective, these

Dr. Sebastian Tillmanns is an Assistant Professor at the Institute of Marketing at the Westfälische Wilhelms-University Münster. Prof. Dr. Manfred Krafft is Director of the Institute of Marketing at the Westfälische Wilhelms-University Münster. The book chapter is adapted from Frenzen and Krafft (2008). We thank Linda Hollebeck for her copy-editing.

S. Tillmanns (✉)

Westfälische Wilhelms-Universität Münster, Muenster, Germany

e-mail: s.tillmanns@uni-muenster.de

M. Krafft

Institute of Marketing, Westfälische Wilhelms-Universität Münster, Muenster, Germany

e-mail: m.krafft@uni-muenster.de

questions are characterized by a dichotomous dependent variable. Traditional regression analyses are not suitable for analyzing these types of problems, because the results that such models produce are generally not dichotomous. Logistic regression and discriminant analysis are approaches using a number of factors to investigate the function of a nominally (e.g., dichotomous) scaled variable. This chapter covers the basic objectives, theoretical model considerations, and assumptions of discriminant analysis and logistic regression. Further, both approaches are applied in an example examining the drivers of sales contests in companies. The chapter ends with a brief comparison of discriminant analysis and logistic regression.

Keywords

Dichotomous Dependent Variables · Discriminant Analysis · Logistic Regression

Introduction

Decision makers and researchers regularly need to explain or predict outcomes that have only one of two values. For example, practitioners might aim to predict, which customers will respond to their next mailing campaign. In this case, the outcome is either response or no response. Variables with such a nominal scaling can be described as dichotomous or binary. Table 1 lists several marketing-based examples. When approaching these issues systematically, it becomes obvious that traditional ordinary least squares (OLS) regression analyses are not suitable for analyzing these types of problems, because the results that such models produce are generally not dichotomous. For example, if product purchase (coding of the dependent variable $Y = 1$) and non-purchase ($Y = 0$) are considered, an OLS regression could predict purchase probabilities of less than zero or above one, which cannot be interpreted. In addition, a binary dependent variable would violate the assumption of normally distributed residuals and would render inferential statistical statements more difficult (Aldrich and Nelson 1984, p. 13). Discriminant analysis and logistic regression are suitable methodologies for the analysis of problems in which the dependent variable is nominally scaled.

Table 1 Potential applications of discriminant analysis and logistic regression in marketing

Subject of investigation	Grouping
Direct marketing/mail order business	Order (yes – no)
Nonprofit marketing	Donation (yes – no)
Retailing	Purchase/non-purchase (choice)
Staff selection (salespeople, franchisees)	Successful/unsuccessful employee
Customer win-back	Defected customers return (yes – no)
Branding	Repurchase (yes – no)
Product policy	New product success versus failure
Innovation adoption	Adopter versus non-adopter

Although both logistic regression and discriminant analysis are suitable for explaining or predicting dichotomous outcomes, discriminant analysis is optimal for classifying observations when the independent variables are normally distributed (conditional on the dependent variable Y). In this case, the discriminant analysis is more efficient, because it utilizes information about the distribution of the independent variables, which the logistic regression does not (Agresti 2013, p. 569). However, if the independent variables are not normally distributed (e.g., because they are categorically scaled) logistic regression is more appropriate, because it makes no assumptions about the independent variables' distribution. Further, in comparison to discriminant analysis, extreme outliers affect logistic regression less. Generally, logistic regression is preferred to discriminant analysis (e.g., Agresti 2013; Osborne 2008). Nevertheless, one way to validate the results of a logistic regression is to see whether an alternative method can replicate the findings and discriminant analysis can serve as such an alternative (Cambell and Fiske 1959; Anderson 1985).

In section “[Discriminant Analysis](#),” we discuss the basic objectives, theoretical model considerations, assumptions, and different steps of discriminant analysis. Analogously, the third section addresses the logistic regression. An applied example is presented in section “[Applied Examples](#),” in which both methods are used to explain which factors drive the use of sales contests in companies. This chapter ends with a conclusion and a brief comparison of discriminant analysis and logistic regression.

Discriminant Analysis

Foundations and Assumptions of Discriminant Analysis

Discriminant analysis is a multivariate (separation) method for the analysis of differences between relevant groups. It can help determine which variables explain or predict the observations' membership of specific groups. Depending on the investigation objectives, this enables the approach to undertake either diagnostic or predictive analyses.

The four key objectives of discriminant analysis can be described as follows (Hair et al. 2010, p. 350; Aaker et al. 2011, p. 470):

- Determining the linear combinations of independent variable(s) that lead to the best possible distinction between groups by maximizing the variance *between* the groups relative to the variance observed *within* the groups. This linear combination is also called a discriminant function or axis. By inserting the independent variables' attribute values into the discriminant function, researchers can calculate a discriminant score for each observation.
- Examining whether, based on the group means obtained through the discriminant scores (centroids), the groups differ significantly from each other.

- Identifying the variables that contribute most effectively to the explanation of intergroup differences.
- Allocating new observations for which the attribute values are known to the relevant group (either classification or prediction).

Based on the distance between the group centroids, it is possible to determine the statistical significance of a discriminant function. Undertaking such analyses requires a comparison of the distribution of the groups' discriminant scores. The less the distributions overlap, the more the groups based on the discriminant function differ. Figure 1 illustrates this basic concept of discriminant analysis according to the distributions of the two groups A and B: besides showing the separation value Y^* (i.e., the critical discriminant score) through which the examined observations can be classified, the discriminant axis is also shown next to the two group centroids \bar{Y}_A and \bar{Y}_B . The area of overlap (i.e., the white areas under the curves) between the two distributions corresponds to the proportion of misclassified observations in group A (i.e., to the right-hand side of Y^*) and group B (i.e., to the left-hand side of Y^*).

Discriminant analysis is considered a dependency analysis, which distinguishes between nominally scaled dependent variables and metrically scaled independent variables. When comparing discriminant analysis with regression and analysis of variance (ANOVA), the following analogies and differences are specifically found:

- Discriminant analysis is based on a model structure similar to the one deployed in multiple regression. The main difference lies in the nominal level of the dependent variable's measurement. In regression analysis, a normal distribution of the error terms is assumed, and the independent variables are known or pre-determined. In discriminant analysis, the reverse applies: the independent variables are assumed to have a multivariate normal distribution, while the dependent variable is fixed; that is, the groups are defined a priori (Aaker et al. 2011, p. 471).
- If a reverse relationship existed, such that the variables of interest are dependent on a group membership, an approach such as a one-factorial multivariate analysis

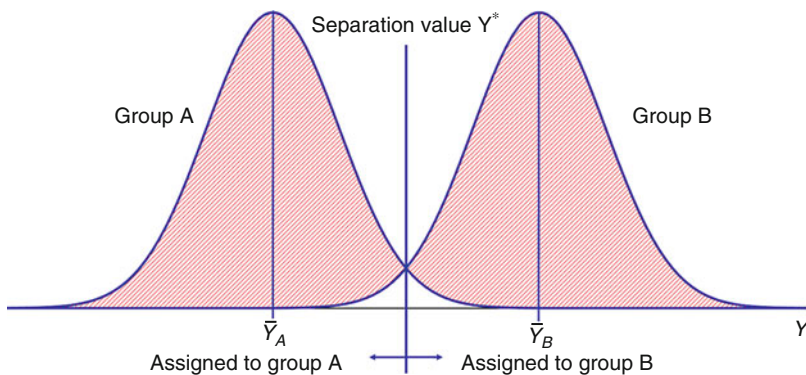


Fig. 1 Distributions of the discriminant scores of two groups A and B

of variance (MANOVA) would be appropriate. Accordingly, discriminant analysis can also be regarded as the reversal of a MANOVA.

To apply discriminant analysis, certain assumptions have to be complied with. For a comprehensive overview of the assumptions of discriminant analysis, see Hair et al. 2010, pp. 354 et seq. and Tabachnick and Fidell 2013, pp. 384 et seq. As in multiple regression analysis, this includes the absence of multicollinearity and autocorrelation. Further, the observed relationships should be of a linear nature. Finally, the independent variables should be normally distributed. Nonetheless, discriminant analysis is relatively robust regarding the violation of the normality assumption, provided the violation is not caused by outliers, but by skewness. If the group sizes are very different, the sample should be sufficiently large to ensure robustness: based on a conservative recommendation, robustness is assumed if the smallest group includes 20 cases and only a few (up to five) explanatory variables (Hair et al. 2010, p. 375; Tabachnick and Fidell 2013, p. 384). The equality of the independent variables' variance-covariance matrices in the different groups is the most important assumption for the application of discriminant analysis. Again, inferential statistics are relatively robust against a violation of this assumption, given that the group sizes are not too small or too uneven. However, the classification accuracy might not be robust even if reasonable group sizes are adopted, because in such cases observations are frequently misclassified into groups with greater variance. If classification is a key objective of the analysis, the homogeneity of the variance-covariance matrices should always be tested (Tabachnick and Fidell 2013, pp. 384 et seq.). Finally, all independent variables are required to be at least interval-scaled, since the violation of this assumption leads to unequal variance-covariance matrices.

Discriminant Analysis Procedure

The application of discriminant analysis involves several steps, which we discuss in the following.

Step 1: Problem and Group Definition As mentioned, discriminant analysis can be used to explain the differences between groups in a multivariate manner. However, it can also serve as a procedure to classify observations with known attribute values, but unknown group membership. To apply discriminant analysis, the dependent variable has to be defined first; that is, the groups to be analyzed need to be determined. The dependent variable can be drawn directly from the specific examination context (e.g., product purchasers vs. non-purchasers) or from the findings of preceding analyses. For example, customer segments identified by means of cluster analysis might be further investigated by undertaking discriminant analysis. The cluster analysis might be used to reveal groups and a subsequent discriminant analysis might use either the same or different variables for further analytical purposes. In the first case, the aim of the investigation would be to verify the clustering variables' adequacy in terms of their discriminatory meaning. In the

second case, the groups generated by means of cluster analysis are explored in greater depth. For instance, an initial cluster analysis generates consumer segments based on their purchase behavior. A subsequent discriminant analysis might then be used to explain the segment-specific differences in consumer purchase patterns by means of psychographic variables.

If, in its original form, the dependent variable is based on a metric scale, it can be classified into two or more groups (e.g., low/medium/high), which can be analyzed by the subsequent discriminant analysis. The group definition is also related to the number of groups to be analyzed. The simplest case is two-group discriminant analysis (e.g., ordering vs. non-ordering in mail order businesses). However, if the classification involves more than two groups, a multigroup discriminant analysis should be used. In determining the number of groups, the number of observations per group should be at least 20. Consequently, combining several groups into a single category might be appropriate.

Step 2: Model Formulation As part of discriminant analysis a discriminant function should be estimated, which ensures an optimal separation between the groups and an assessment of the discrimination capability of the applied variables. The discriminant function is a linear combination of the applied variables and generally adheres to the following scheme:

$$Y_i = \beta_0 + \sum_{j=1}^J \beta_j X_{ji}, \quad (1)$$

where

Y: Discriminant score

β_0 : Constant

β_j : Discriminant coefficient of independent variable j

By using the attribute values X_{ji} for each observation i , the discriminant function generates an individual discriminant score Y_i .

Step 3: Estimation of the Discriminant Function The parameters β_j are estimated such that the calculated discriminant scores provide an optimal separation between the groups. This result requires a discriminant criterion that measures the intergroup differences. The estimation procedure is then carried out such that the discriminant criterion is maximized.

Considering the distance between the group centroids for the purpose of evaluating the differences between the groups might initially seem obvious. However, also the variance of the discriminant scores within a group has to be considered. While a larger distance between two group centroids improves the distinction of the groups, the distinction is made more difficult when the variance of the groups' discriminant scores increases. This situation is illustrated in Fig. 2, where two pairs of groups A and B are represented as distributions on the discriminant axis. The centroids of the pairs of groups A ($\bar{Y}_A = -2$) and B ($\bar{Y}_B = 2$) are identical, but the

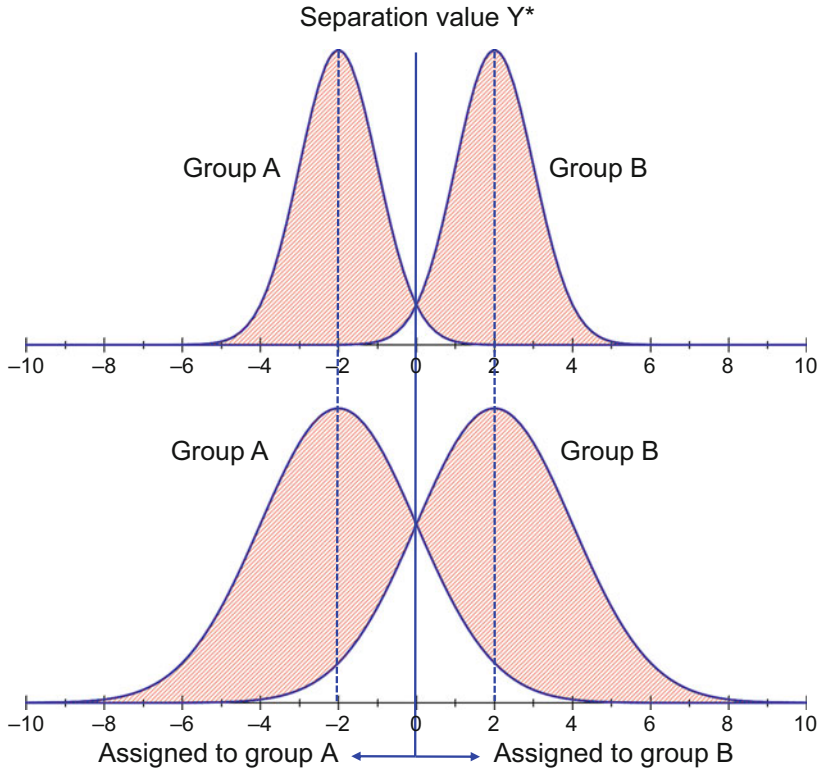


Fig. 2 Groups with different centroids and variances

standard deviations of the discriminant scores differ (standard deviations upper distributions = 1; standard deviations lower distributions = 2). Based on a comparison of the pairs shown at the top and at the bottom of Fig. 2, we can observe that, although the lower groups indicate identical centroids, they also have greater variances – in this case, generating a lower discrimination.

Thus, assessments of discriminant analysis performance should aim at optimizing both the distance of the group centroids and the spread of the observations. To achieve optimal separation efficiency between the groups, we refer to the well-established regression- and ANOVA-based principle of variance decomposition:

$$SS = SS_b + SS_w$$

$$\begin{aligned} \text{Total variance} &= \text{Between – group variance} + \text{Within – group variance} \quad (2) \\ &= \text{Explained variance} + \text{Unexplained variance} \end{aligned}$$

Thus, a discriminant function should be determined such that the group means (centroids) differ significantly from one another, if possible. For this purpose, we refer to the following discriminant criterion:

$$\Gamma = \frac{\text{Variance between the groups}}{\text{Variance within the groups}} \quad (3)$$

This criterion can be made more precise and converted into an optimization problem as follows:

$$\Gamma = \frac{\sum_{g=1}^G I_g (\bar{Y}_g - \bar{Y})^2}{\sum_{g=1}^G \sum_{i=1}^{I_g} (Y_{gi} - \bar{Y}_g)^2} = \frac{SS_b}{SS_w} \rightarrow \text{Max}_{b_j}! \quad (4)$$

where

G: Number of groups studied

I: Group size

In order to account for different group sizes, the variance between the groups is multiplied by the respective group size I . The discriminant coefficients β_j ($j = 1, \dots, J$) have to be determined so that the discriminant criterion Γ is maximized. The maximum value of the discriminant criterion

$$\gamma = \text{MAX}\{\Gamma\} \quad (5)$$

is referred to as the eigenvalue, because it can be determined mathematically by solving the eigenvalue problem (for further details refer to Tatsuoaka 1988, pp. 210 et seq.).

In the multigroup case (i.e., with more than two groups), more than one discriminant function and more than one eigenvalue must be determined. The maximum number of discriminant functions is given by $K = \text{Min}\{G-1, J\}$. Generally, the number of independent variables J exceeds the number of groups, so that the number of groups usually determines the number of discriminant functions to be estimated. The following applies to the order of the respective discriminant functions and their corresponding eigenvalues:

$$\gamma_1 \geq \gamma_2 \geq \gamma_3 \geq \dots \geq \gamma_K.$$

This way, the first discriminant function explains the majority of the independent variables' variance. The second discriminant function is uncorrelated (orthogonal) to the first and explains the maximum amount of the remaining variance once the first function has been determined. Since additional discriminant functions are always calculated in order to explain the remaining variance's maximum share, the functions' explanatory power declines gradually.

The following measure represents a discriminant function's relative importance by deriving the respective eigenvalue share (ES):

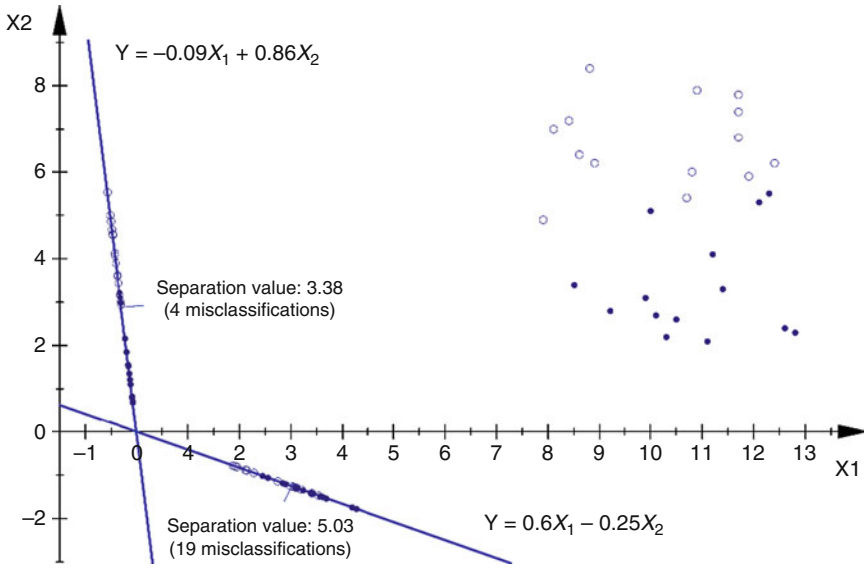


Fig. 3 Graphical representation of a two-group discriminant analysis

$$ES_k = \frac{\gamma_k}{\gamma_1 + \gamma_2 + \dots + \gamma_K} \tag{6}$$

The eigenvalue share equals the share of variance explained by the k^{th} discriminant function relative to the variance explained by all discriminant functions K combined. The eigenvalues always add up to one.

Figure 3 illustrates the discriminant analysis’s estimation principle by means of a graphic representation. In this case, the determination is based on a two-group discriminant function. It is assumed that two attribute values X_1 and X_2 are available for each of the subjects examined, which are shown in the scatter plot in Fig. 3, in which filled circles represent the members of group A and empty circles members of group B. In this example the optimal discriminant function is $Y = -0.09X_1 + 0.86X_2$. Based on this function four members of group A are misclassified as members of group B. The arbitrarily selected discriminant function $Y = 0.6X_1 - 0.25X_2$ shows a worse classification performance by having 19 misclassifications in total. A projection of the respective attribute value combinations on the discriminant axes corresponds to the examined subjects’ discriminant values.

Step 4: Assessment the Discriminant Function Performance Researchers have two basic approaches at their disposal to assess a discriminant function’s performance. The first approach compares the classification of observations with the observations’ actual group membership. This approach is explained later in this chapter in relation to logistic regression, where a specific example is also

provided. It is also specifically applicable in the context of the discriminant analysis.

A second fundamental way of assessing the discriminant function's quality is based on the discriminant criterion described above. The eigenvalue γ represents the maximum value of the discriminant criterion and, thus, forms a starting point for assessing the discriminant function's quality or discriminant power. Since γ has the disadvantage of not being standardized to values between zero and one, other metrics based on the eigenvalue have been established for quality assessment, including the canonical correlation coefficient (c):

$$c = \sqrt{\frac{\gamma}{1 + \gamma}} = \sqrt{\frac{\text{Variance explained}}{\text{Total variance}}} \quad (7)$$

In the two-group case, the canonical correlation equals the simple correlation between the estimated discriminant values and the binary grouping variables. The most common criterion for testing the quality of the discriminant function is Wilks's lambda (Λ):

$$\Lambda = \frac{1}{1 + \gamma} = \frac{\text{Unexplained variance}}{\text{Total variance}} \quad (8)$$

As can be seen from the above formula, Wilks's lambda is an inverse measure of goodness; that is, lower (higher) values imply a better (worse) discriminant power of the discriminant function. Wilks's lambda can be transformed into a test statistic, thereby enabling inferential statements on the diversity of groups. By applying the transformation

$$V = - \left[N - \frac{J + G}{2} - 1 \right] \ln \Lambda, \quad (9)$$

where

N: Number of observations

J: Number of independent variables

G: Number of groups

Λ : Wilks's lambda

the resulting test statistic is approximately χ^2 -distributed with $J(G-1)$ degrees of freedom. Thus, a statistical significance testing of the discriminant function can be performed. Given that the test statistic increases with lower values of Λ , higher values imply a greater diversity between the groups.

In the case of multigroup discriminant analysis, each discriminant function can be evaluated by means of the above measures on the basis of their respective eigenvalue. Here, the k^{th} eigenvalue γ_k measures the proportion of the explained variance which can be attributed to the k^{th} discriminant function. There is a clear analogy to

the principal components analysis procedure in which the components are extracted from the independent variables (like the “main components”) in the form of discriminant functions (Tatsuoka 1988).

All discriminant functions and their eigenvalues are considered to assess the overall differences between the groups. The multivariate Wilks’s lambda, which is calculated by multiplying the univariate lambdas, is a measure that can capture this:

$$\Lambda = \prod_{k=1}^K \frac{1}{1 + \gamma_k}, \quad (10)$$

where

K: Number of possible discriminant functions

γ_k : Eigenvalue of the k^{th} discriminant function

To statistically check whether the groups differ significantly from one another, a χ^2 -distributed test statistic can be generated by using transformation (9).

Other test statistics representing approximations of the F-distribution and Wilks’s lambda are available and can be applied to test the significance of group-related differences. These statistics cannot only be applied to a MANOVA but also to discriminant analysis (e.g., Hotelling’s trace, Pillai’s trace, and Roy’s GCR). Other measures, such as Rao’s V and Mahalanobis’s D^2 are particularly applied in the context of stepwise discriminant analyses (Tabachnik and Fidell 2013, p. 399; Hair et al. 2010, p. 435).

As with all statistical tests, a statistically significant test result does not necessarily imply a substantial difference. In a sufficiently large sample size, even small differences are likely to have statistical significance. Consequently, the absolute values of the mean differences between the groups, the canonical correlation coefficients, and Wilks’s lambda, should not be overlooked. For the sake of interpretability, it is advisable to limit analyses to two or three discriminant functions, even if further discriminant functions prove to be statistically significant.

Step 5: Examination and Interpretation of the Discriminant Coefficients If testing the discriminant function(s) result(s) in a sufficient discriminatory power between the groups in step 4, the independent variables can be examined. Assessing the importance of individual independent variables can be used to, first, explain the key differences between the groups, thereby contributing to the interpretation of group-based differences. Second, unimportant variables can be removed from the model if the goal is to specify a parsimonious model. As an alternative to simultaneously including all variables into the model for estimation, a stepwise estimation may be undertaken. Here, only variables are included in the discriminant function one at a time, which contribute significantly to the discriminant function’s improvement, depending on the level of significance that the researcher specifies.

Individual variables’ discriminatory relevance can be checked by using univariate and multivariate approaches. As part of a *univariate* assessment, each of the

variables can be tested separately to determine whether their mean values between the groups differ significantly from one another. Likewise, undertaking discriminant analyses of each of the variables based on Wilks's lambda serves to isolate their discriminant power. The F-test can also be used here. The result then corresponds to a one-factorial analysis of variance with the groups as factor levels.

A review of the univariate discriminatory relevance is insufficient if there are potential interdependencies between the variables. For example, while a particular variable may have little discriminatory meaning when viewed in isolation, it may significantly contribute to an increase of discriminant power in combination with other variables.

Multivariate assessments of individual variables' discriminatory power and of their importance as part of the discriminant function can be undertaken by using the standardized discriminant coefficients. These represent the influence of the independent variables on the discriminant variable and are defined as follows:

$$\beta_j^* = \beta_j \cdot s_j, \quad (11)$$

where

β_j : Discriminant coefficient of independent variable j

s_j : Standard deviation of independent variable j

Standardization allows for assessments independent of independent variables' scaling and their meaning. The higher the absolute value of a standardized coefficient, the greater the discriminatory power of the associated variable. The unstandardized discriminants are, however, required to calculate the discriminant scores (Hair et al. 2010, p. 381).

Deriving the correlation coefficients between the values of the respective independent variables and the discriminant scores is another alternative method to interpret the independent variables' influences. These correlation coefficients are called discriminant loadings, canonical loadings, or structure coefficients. Compared to (standardized) discriminant coefficients, potential multicollinearity between the independent variables affects them less. Therefore they often provide benefits in terms of an unbiased interpretation of the independent variables. As a rule of thumb, loadings exceeding a magnitude of 0.4 indicate substantially discriminatory variables (Hair et al. 2010, pp. 389 et seq.). The identification of variables with sufficiently high loadings allows for creating profiles of groups in terms of these variables and to identify differences between the groups. The signs of discriminant weights and loadings reflect the groups' relative average profile.

A sufficiently large sample size is required to obtain stable estimates of the standardized discriminant coefficients and discriminant loadings. As a guiding value, a minimum of 20 observations per independent variable is required (Hair et al. 2010, p. 435; Aaker et al. 2011, p. 477).

Step 6: Prediction In discriminant analysis, several alternative prediction approaches are available. Predictions can be based on classification functions, the distance concept, or the probability concept (Backhaus et al. 2016, pp. 246 et seq.). Classification functions and the probability concept allow the consideration of a priori probabilities. Such probabilities can reflect theoretical knowledge about different group sizes before any prediction is conducted. The consideration of a priori probabilities is especially useful, if the examined groups differ in their size. Furthermore, the probability concept allows to allocate specific costs to the misclassification of observations into a certain group. Specifically, this builds on the distance concept and thus should be addressed last.

Classification Functions

Fisher's classification functions can be used to predict an observation's group-membership. They require the variances in the groups to be homogeneous and one classification function has to be determined for each group. Thus, in a two-group case, two functions have to be determined:

$$\begin{aligned} F_{1i} &= \beta_{01} + \sum_{j=1}^J \beta_{j1} X_{ji} \\ F_{2i} &= \beta_{02} + \sum_{j=1}^J \beta_{j2} X_{ji} \end{aligned} \quad (12)$$

For the classification of an observation, the value of each function F_i has to be calculated. An observation is assigned to the group for which it yields the maximum value.

As mentioned above, classification functions allow the consideration of a priori probabilities $P(g)$. A priori probabilities must add up to one and are implemented in the classification function as follows:

$$F_g := F_g + \ln P(g) \quad (13)$$

It is also possible to determine individual probabilities $P_i(g)$ for each observation i .

Distance Concept

According to the distance concept, an observation i needs to be assigned to a group g , such that the distance to the centroid is minimized (i.e., to which it is closest on the discriminant axis). This corresponds to determining whether an observation lies either left or right of the critical discriminant score Y^* (see Fig. 1). The squared distance is the measure usually deployed in the K -dimensional discriminant space between observation i and the centroid of group g :

$$D_{ig}^2 = \sum_{i=1}^I (Y_{ki} - \bar{Y}_{kg})^2, \quad (14)$$

where

Y_{ki} : Discriminant score of observation i according to discriminant function k

\bar{Y}_{kg} : Centroid of group g regarding discriminant function k

It is, however, not necessary to consider all possible K discriminant functions to execute the classification. Usually, it is sufficient to limit the analysis to the significant or relevant discriminant functions, which substantially facilitates the calculation.

Classification based on the distance concept requires the variances in the groups to be nearly homogeneous. This assumption can, for instance, be checked using Box's M as a test statistic. When the assumption of homogeneous variances in the groups is violated, modified distance measures need to be calculated.

Probability Concept

The probability concept, which builds on the distance concept, is the most flexible approach to the classification of observations. It allows the consideration of a priori probabilities and different misclassification costs in the examined group. Without these modifications, the probability concept generates the same results as the distance concept.

Regarding the classification of observations with the probability concept, a priori probabilities and conditional probabilities are combined in order to derive a posteriori probabilities according to the Bayes theorem:

$$P(g|Y_i) = \frac{P(Y_i|g)P_i(g)}{\sum_{g=1}^G P(Y_i|g)P_i(g)}, \quad (15)$$

where

$P(g|Y_i)$: A posteriori probability, that an observation is in group g , given a discriminant score Y_i is observed.

$P(Y_i|g)$: Conditional probability, that a discriminant score Y_i is observed, given that it appears in group g .

$P_i(g)$: A priori probability, that an observation is in group g

An observation i is assigned to group g , for which the value of $P(g|Y_i)$ is maximized. For example, if $G = 2$, observation i is assigned to group 1 if

$$\frac{P(Y_i|g_1)P_i(g_1)}{\sum_{g=1}^2 P(Y_i|g)P_i(g)} > \frac{P(Y_i|g_2)P_i(g_2)}{\sum_{g=1}^2 P(Y_i|g)P_i(g)}. \quad (16)$$

The a posteriori probabilities can be calculated on the basis of the observations' distances to the group centroids (see Tatsuoka 1988, pp. 358 et seq. for more details). Another advantage of the probability concept lies in the possibility of explicitly incorporating the *costs of a misclassification* in the decision rule. The field of medical diagnostics can be used as an example: the consequences of not diagnosing a malignant disease are certainly more fatal than coming up with an erroneous diagnosis. Hence, the expected value of the costs involved might be considered in such calculations:

$$E_g(C) = \sum_{h=1}^G C_{gh}P(h|Y_i). \quad (17)$$

The costs, which are quantified by C_{gh} , arise if an observation is assigned to group g , though it belongs to group h . Thus, an observation i is assigned to the group g with the lowest expected costs $E_g(C)$.

Marketing decisions in which misclassification costs could play a major role are for example new product introductions. In this regard, high costs may arise due to misclassifications of products as a “success” or a “flop” in terms of either their launch or non-launch. Further, in the mail order business, substantial misclassification costs may result from customers either being sent or not sent a catalogue of relevant product assortments.

Logistic Regression

Foundations and Assumptions of Logistic Regression

The application of logistic regression has become increasingly popular in recent years. There is little difference between logistic regression and discriminant analysis with respect to their objectives and applications. On the one hand, logistic regression may be used to examine the variables and the specific degree to which they contribute to explaining group membership (diagnosis). On the other hand, it allows for classifying new observations into groups (prediction). The following section describes the basics underlying logistic regression's estimation method. Initially, the measures with which to assess the entire model are explained only generically, because an example in section “[Logistic Regression](#)” is used to clarify the quality measures as well as the options for interpreting the coefficients. This example draws on a study that Mantrala et al. (1998) conducted to address the (non-)adoption of sales contests. The dependent variable takes two values:

$$Y = \begin{cases} 1, & \text{if sales contests are used,} \\ 0, & \text{if sales contests are not used.} \end{cases} \quad (18)$$

Compared to a linear regression, linking one or more independent variables to a dependent variable that can only take one of two values (i.e., 0 and 1) is more complicated. However, this linkage can be established through a logistic regression model, which can be derived as either a latent variable model or a probability model (see Long and Freese 2014). In the following, both approaches are explained.

In the latent variable model, a non-observed (i.e., latent) variable Y^* is assumed, which is related to the observed independent variables in the following way:

$$Y_i^* = \beta_0 + \sum_{j=1}^J \beta_j X_{ji} + \epsilon_i, \quad (19)$$

where

β_0 : Constant

β_j : Coefficient of independent variable j

ϵ_i : Error term

Equation 19 is identical to a linear regression, in which the dependent variable can range from $-\infty$ to ∞ , but it differs with regard to the dependent variable, which is non-observable. In order to transform the continuous non-observable dependent variable into a dichotomous one (i.e., one that can only take the values 0 and 1), the following linkage is established:

$$Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

Thus, Y_i is assigned a value of 1 if Y_i^* takes positive values and a value of 0 if Y_i^* is smaller or equal to 0. In our abovementioned example, Y_i^* can be viewed as the propensity to adopt (as opposed to the non-adoption of) sales contests.

To illustrate the latent variable model, we assume a single independent variable in the following. For any given value of X of this variable, we can link the observable dependent variable Y_i to its non-observable counterpart through the following equation:

$$P(Y = 1|X) = P(Y^* > 0|X) \quad (21)$$

Equation 21 represents the probability of an event occurring (e.g., the use of a sales contest), with either $Y = 1$ or $Y^* > 0$ representing the occurrence of the event, which is conditional on the value X of the independent variable. If Y^* is substituted by Eq. 19 (and restricted to a single independent variable), Eq. 21 results in:

$$P(Y = 1|X) = P((\beta_0 + \beta_1 X_{1i} + \epsilon_i) > 0|X) \quad (22)$$

It can be rearranged in:

$$P(Y = 1|X) = P(\epsilon_i > -(\beta_0 + \beta_1 X_{1i})|X) \tag{23}$$

In this equation, the probability of the event occurring depends on the distribution of the error ϵ_i . Depending on the assumptions about this distribution, either a probit or a logit model can be derived. Similar to logit models, probit models are capable of modeling a dichotomous dependent variable. While a normal distribution with a variance of 1 is assumed for the probit model, a logistic distribution with a variance of $\frac{\pi^2}{3}$ is assumed for the logit model. In the logistic regression model, this leads to the following equation:

$$P(Y = 1|X) = \frac{e^{(\beta_0 + \beta_1 X_{1i})}}{1 + e^{(\beta_0 + \beta_1 X_{1i})}} \tag{24}$$

In a more general form, which allows for more than one independent variable, this becomes:

$$P_i = \frac{e^{Z_i}}{1 + e^{Z_i}} = \frac{1}{1 + e^{-Z_i}}, \tag{25}$$

where Z_i serves as a linear predictor of the logistic model for the i^{th} observation with $Z_i = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} + \dots + \beta_{ji} X_{ji}$.

The probability model is an alternative to the latent variable model, which allows the logistic model to be derived without referring to a latent variable (see Theil 1970; Long and Freese 2014). In order to derive a model with an outcome ranging from 0 to 1, the probabilities of an event occurring have to be transformed into odds:

$$\text{Odds}(Y = 1) = \frac{P(Y = 1)}{P(Y = 0)} = \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \tag{26}$$

Odds indicate how often something happens relative to how often it does not happen (i.e., $Y = 1$ vs. $Y = 0$). The odds therefore represent the chance of an event occurring. The natural logarithm of the odds is called logit (logistic probability unit), which ranges from $-\infty$ to ∞ , and is linear:

$$\ln(\text{Odds}(Y = 1)) = \beta_0 + \sum_{j=1}^J \beta_j X_{ji} \tag{27}$$

The logistic function, which is an S-shaped curve, has the advantageous characteristic that even for infinitely small or large values of the logit Z_i the resulting values of P_i are never outside the interval of $[0,1]$ (Hosmer et al. 2013, pp. 6 et seq.).

The nonlinear characteristics of Eq. 25 are represented in Fig. 4, in which the exponent X_i is systematically varied from -9 to +9. The figure shows the

prediction of an indifferent result if the sum of the weighted factors (i.e., Z_i) equals zero. Its symmetry at the turning point of $P_i = 0.5$ is another essential feature of the logistic function. The constant β_0 of the linear predictor Z_i moves the function horizontally, while higher coefficients lead to a steeper slope of the logistic function. Negative signs of the coefficient β_j change the origin of the curve, which corresponds to the dotted line in Fig. 4 (Menard 2002, pp. 8 et seq.; Agresti 2013, pp. 119 et seq.).

As mentioned, in empirical research the actual (non-)entry of an event and not its probability of entry is observed. The logistic regression approach regarding the occurrence of an event ($Y_i = 1$) and its opposing event ($Y_i = 0$) can thus be expressed as follows for each observation i :

$$P_i(Y) = \left(\frac{1}{1 + e^{-Z_i}} \right)^{Y_i} \left(1 - \frac{1}{1 + e^{-Z_i}} \right)^{1-Y_i} \tag{28}$$

Thus, the information required to establish the probabilities regarding the z-values (logits) can be calculated using Eq. 25.

The coefficients of the logistic model β_j can now be estimated by maximizing the likelihood (L) of obtaining the empirical observation values of all possible cases. Since the observed values Y_i represent realizations of a binomial process with the probability P_i , which vary depending on the expression of X_{ji} , we are able to set up the following likelihood function, which has to be maximized:

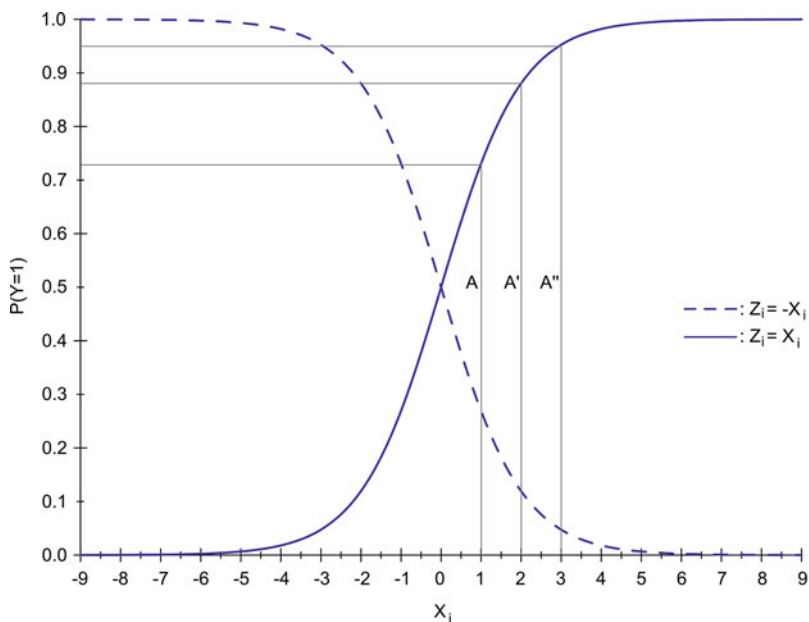


Fig. 4 Progression of the logistic function curves

$$L = \prod_{i=1}^N \left(\frac{1}{1 + e^{-Z_i}} \right)^{Y_i} \left(1 - \frac{1}{1 + e^{-Z_i}} \right)^{1-Y_i} \rightarrow \text{Max!} \quad (29)$$

Compliance with specific assumptions is required to apply logistic regression usefully. Compared to discriminant analysis, logistic regression offers the advantage that the assumptions of multivariate normality and identical variance-covariance matrices between the groups are not required. It should be ensured that the independent variables do not exhibit multicollinearity, errors are independent, and the linearity of the logit is given (Field 2013, pp. 768 et seq. and pp. 792 et seq.).

The Variance Inflation Factor (VIF) is commonly derived to assess multicollinearity. Empirical research often quotes critical VIF values that must not be exceeded. Nevertheless, it should be noted that multicollinearity problems might already arise with very small VIFs (Baguley 2012). Hence, in order to control for multicollinearity, researchers should exclude their specific independent variables from the model one at a time and verify whether the remaining independent variables experience a substantial change regarding their level of significance or the direction of their effect.

In logistic regression, the outcome variable is categorical and, hence, there is no given linear relationship between independent and dependent variables. Nevertheless, the assumption of linearity in logistic regression assumes a linear relationship between any metric independent variable and the logit of the outcome variable. A significant interaction term between an independent variable and its log transformation indicates that this assumption has been violated (Field 2013, p. 794; for further approaches to test this assumption refer to Hosmer et al. 2013, pp. 94 et seq.).

In addition, one should ensure that outliers and other influential observations do not affect the estimated results. Some statistical packages, like SAS, offer a huge variety of outlier statistics for logistic regression, which might be used to alter results in different directions. Hence, researchers should always report whether substantial changes occur in their results when outliers have been removed.

Finally, a reasonable sample size must be used, since maximum likelihood estimation is adopted with this method and requires a large number of observations based on its asymptotic properties. With respect to a study by Peduzzi et al. (1996) it is recommended that the ratio of the group size to the number of independent variables should be at least 10:1 for the smallest group. For instance, in a sample where 25% of the observations have an outcome of $Y = 1$ (representing the least frequent outcome) and 4 independent variables are available, 160 observations are needed (160 observations * 0.25 (share of the observations with the least frequent outcome) = 40 observations; 40 observations / 4 variables = 10 observations for each independent variable). Even though this recommendation is widely accepted in literature, Agresti (2013, p. 208) and Hosmer et al. (2013, p. 408) note that this is merely one guideline and models that violate it should not necessarily be neglected. For example, Vittinghoff and McCulloch (2006) conclude in their simulation study that at least 5-9 observations for each independent variable in the smallest group should be sufficient. A more stringent threshold is provided by Aldrich and Nelson (1984), who view 100 degrees of freedom as a minimum for a valid estimation.

Logistic Regression Procedure

The application of logistic regression comprises different steps, which are addressed in more detail below.

Step 1: Problem and Group Definition As stated, similar to two-group discriminant analysis, logistic regression is suitable for the multivariate explanation of differences between groups or for the classification of group-based observations for prediction purposes. Logistic regression is therefore generally appropriate when a single categorical variable is used as a dependent variable. When more than two groups are given, ordered or multinomial logistic regressions can be applied, which represent generalizations of binary logistic regression. If the dependent variable has a metric scale, either those observations that are furthest from one another can be coded as 0 and 1 or the metric variables can be classified into multiple groups (e.g., low, medium, high). This approach is called the “polar extreme approach” (Hair et al. 2010, p. 352).

Step 2: Model Formulation Model formulation may be used to assess which of the independent variables should be analyzed. In contrast to discriminant analysis, no metric scale is required of the explanatory variables. Rather, dummy or categorical variables can be considered. It is important to clarify whether these nonmetric variables are indicator or effect coded. With indicator coding, the reference category is assigned the value of 0 and the resulting coefficient should be interpreted as a relative effect of a category compared to the reference category. Effect-coded variables may be interpreted as a category’s relative influence compared to the average effect of all the categories. Unlike in indicator coding, one of the categories is assigned the value of -1. The coefficients of the categories of effect-coded variables add up to zero, so that the reference category’s coefficient can be calculated from the coefficients of the other categories (for further details, see Hosmer et al. 2013, pp. 55 et seq.).

Generally, the selection of the independent variables should be based on sound logical or theoretical considerations. Nevertheless, especially in the context of Big Data, a large number of interdependent independent variables with unknown quality are often available. In this case, theoretical considerations might not be possible. However, a limited set of variables still needs to be selected to avoid overfitting. Overfitting often occurs when a large number of variables are included in a prediction model. This might result in a good prediction performance, if a logistic regression model is applied to the observations used for its estimation. Nevertheless, as soon as a prediction is conducted for observations outside the original sample, the prediction performance is likely to suffer. Thus, researchers and practitioners might therefore reduce the dimensions of their data by applying a principal component analysis and only including a limited number of factors into their logistic regression model. Alternatively, variable selection approaches can be used. Tillmanns et al. (2017) provide an extensive comparison of different approaches that can handle a large number of independent variables in models predicting a binary outcome.

Finally, a sufficiently large number of observations is required (see section “[Foundations and Assumptions of Logistic Regression](#)”), which should be distributed as evenly as possible between the analyzed groups. If a sufficiently large sample is given, it also makes sense to split the observations into an estimation sample and a holdout sample. The regression function estimated on the basis of the estimation sample can then be used to cross-validate the results with those from the holdout sample.

Step 3: Function Estimation Before the regression model can be estimated, the assumptions underlying logistic regression need to be checked (see section “[Foundations and Assumptions of Logistic Regression](#)”). If these assumptions are met, the model can be estimated. As in multiple regression analysis or discriminant analysis, this can be done by using a stepwise procedure (“forward/backward elimination”) or by the simultaneous entry of each of the independent variables into the estimation equation (“enter”). If the logistic regression is used to test hypotheses, the latter method is required. In preliminary estimation runs, the potential presence of influential observations or outliers should also always be checked. In this process, the use of Cook’s distance is recommended. In cases where the sample is very unbalanced (i.e., one outcome is very rare), researchers might consider ReLogit (Rare Events Logistic Regression) as an alternative approach (for a detailed explanation see King and Zeng 2001).

The maximization of the likelihood function shown in Eq. 29 is achieved with statistical packages, such as SAS or SPSS, using the Newton-Raphson algorithm. The principle underlying maximum likelihood estimation is to select the estimates of parameter β_j in a stepwise iterative analytical approach, such that the observation of the estimated value is assigned a maximum likelihood.

Figure 5 illustrates two logistic functions fitted to two different samples. On the ordinate, the dependent variable Y is shown, wherein $P(Y = 1)$ represents the probability of occurrence of an event. On the abscissa, the values of an independent variable X_i are represented. The logistic function reflects the predicted (estimated) probabilities for the occurrence of the event under the independent variables’ different values. The actual observation values are marked by points. In the upper part (a), the logistic function is suitably adapted to the observed data: high independent variable values correspond to the occurrence of the event, and vice versa. However, in the lower part (b), the logistic function is not suitably adapted to the observed data, which is expressed by the large overlap between the two groups in the central region of the abscissa. Entering a cutoff point of 0.5 for classifying observations shows that, in example (a), four observations are misclassified whereas eight observations are misclassified in example (b).

Step 4: Assessment of the Model Performance Before starting with the interpretation of the individual coefficients, it is important to first investigate whether an estimated logistic regression model is in fact suitable. When reviewing this issue, one cannot rely on the traditional measurements and tests used in linear regression analysis (such as the coefficient of determination or F-values), because the

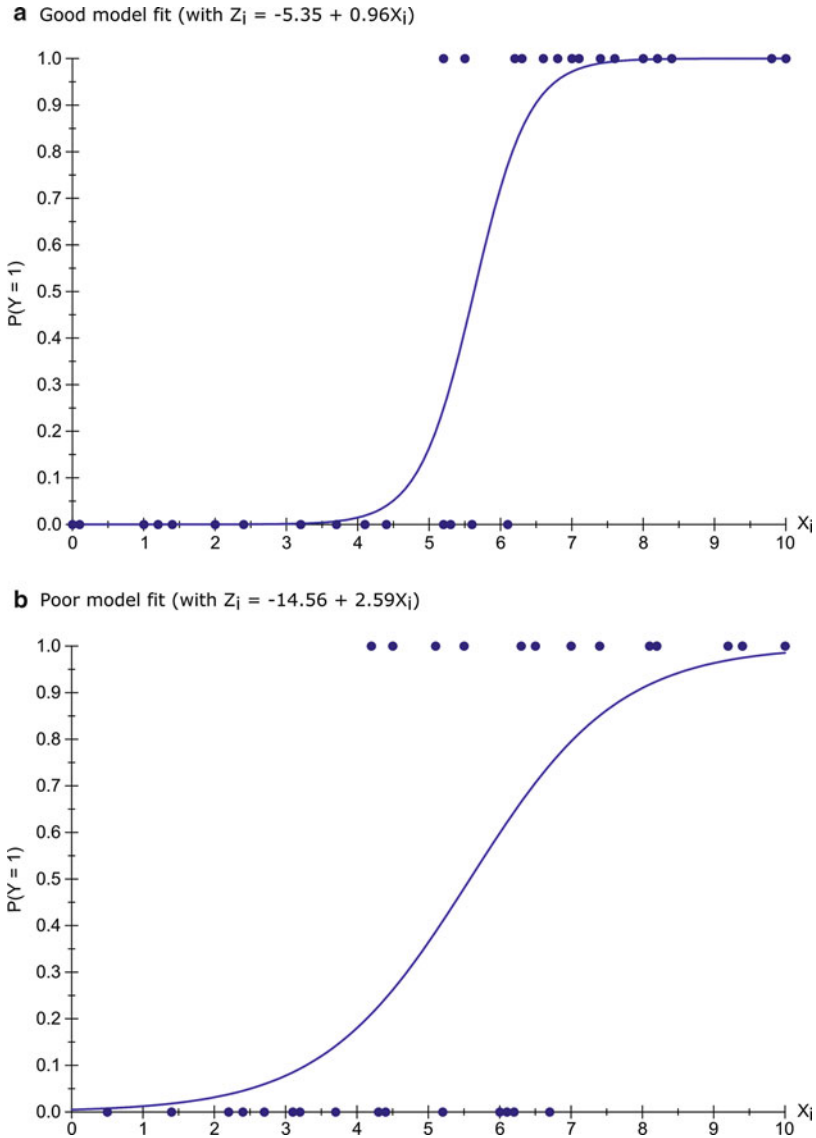


Fig. 5 Fit of the logistic function with different samples

coefficients obtained from logistic regression are determined by using maximum likelihood estimation. Goodness-of-fit in maximum likelihood applications is usually assessed by using *deviance* (or $-2LL$), which is calculated as $-2 \cdot \log(\text{likelihood})$. A perfect fit of the parameters is equivalent to a likelihood of 1, corresponding to a deviance of 0 (Aldrich and Nelson 1984, p. 59; Hosmer et al. 2013, p. 155).

In terms of interpretation, the likelihood is comparable to that of least squares errors in conventional regression analysis. Further, $-2LL$ is used, because it is asymptotically χ^2 -distributed with $(N-p)$ degrees of freedom, where N is the number of observations, and p the number of the parameters. Good models reflecting a “high” likelihood of close to 1 result in a deviance that is close to 0, while a bad fit is mirrored in high deviance values. Deviance values are unlimited in positive range.

Whether a deviance has to be considered as rather “high” or “low” depends on the particular sample used and the analysis deployed. The value of the deviance can be used to test a hypothesis H_0 , which posits that the overall model shows a perfect fit with the data. With either low deviance values or high significance, H_0 cannot be rejected and the model can be deemed as showing a good fit with the data (Menard 2002, pp. 20 et seq.).

In addition to the deviance, *likelihood ratio tests* and pseudo- R^2 statistics can be applied, which provide additional indicators permitting assessments of the full model fit compared to the null model, which only comprises an estimated value for β_0 , i.e., the intercept of the linear predictor Z_i . In this process, the deviance is applied, which can also be used for comparing incremental differences of different models. The absolute difference between the deviance of the null model and the full model provides an asymptotically distributed χ^2 value, which can be tested against the null hypothesis that the full model coefficients are not significantly different from 0. Thus, a likelihood-ratio test can be conducted, which is comparable to the F-test in linear regression analysis. This test statistic is called the Model Chi-Square. High χ^2 values and low significance levels suggest that the final model’s coefficients are significantly different from 0. For this test, the 5% level is usually set as the critical level of significance (Hosmer et al. 2013, p. 40; Menard 2002, p. 21; Tabachnik and Fidell 2013, pp. 448 et seq.).

In logistic regressions, the deviances of the full and the null model can be used to calculate McFadden’s R^2 . This pseudo- R^2 metric is expressed in the following equation:

$$\text{McFadden's } R^2 = 1 - \frac{LL_1}{LL_0}, \tag{30}$$

where

LL_1 : Natural logarithm of the likelihood of the full model.

LL_0 : Natural logarithm of the likelihood of the null model

There are additional pseudo- R^2 statistics, which are also based on a comparative goodness-of-fit of the null model and the full model. Similarly, the R^2 statistics of Cox and Snell, as well as of Nagelkerke suggest that higher values correspond to enhanced model fit, with the maximum value of 1 corresponding to perfect model fit.

Assessments of the model performance can also be undertaken on the basis of the attained classification results, where the predicted values P_i are compared with the actual (observed) values Y_i . As part of the Hosmer-Lemeshow test, the accuracy of

predictions is checked against the null hypothesis, which posits that the difference between the predicted and observed values is equal to zero (Hosmer et al. 2013, pp. 157 et seq.). These observations are divided into 10 groups of approximately equal size based on the predicted values P_i . By means of a Chi-Square test, the extent to which the observed and predicted frequencies differ is checked. A low χ^2 value of the test statistic, coupled with a high significance level, implies a good model fit.

Further, to assess the predictive accuracy of logistic models, the confusion matrix or classifications matrix is considered. In this 2*2 matrix, the predicted group memberships on the basis of the logistic regression model are compared to the empirically observed group memberships (Menard 2002, p. 31). The correctly classified items are then available on the main diagonal, while the misclassified observations appear off the main diagonal. The proportion of correctly classified elements – as facilitated by logistic regression (which is called the match quality or the hit ratio) – should be higher than the matches obtained from a random allocation. This is a limitation, since the matches attained by using the model are inflated when the parameter estimation of logistic regression and the calculation of the hit rate are based on the same sample (Morrison 1969, p. 158; Hair et al. 2010, p. 373; Afifi et al. 2012, p. 265). Using the estimated coefficients from a calibration sample is therefore expected to generate lower hit rates for other (holdout) samples. With approximately equal groups represented by the dependent variable, using the maximum chance criterion (MCC) to assess the classification performance is recommended. By applying the MCC metric, the classification performance is based on the share of the larger groups within the total sample (Morrison 1969, p. 158; Hair et al. 2010, p. 384; Aaker et al. 2011, p. 478). Nevertheless, if two groups of rather unequal size are considered, adopting the MCC metric is inadequate.

The proportional chance criterion (PCC) is particularly recommended when analyzing two groups of unequal size or when seeking a classification that is equally as good for both groups. The PCC is equivalent to a random hit rate of $\alpha^2 + (1-\alpha)^2$, where α represents the proportion of a group to the total number of observations (Morrison 1969, p. 158; Hair et al. 2010, p. 384; Aaker et al. 2011, p. 478). Whether to use the PCC, MCC, or a different classification criterion depends on the particular subject of investigation. For example, it may be meaningful to minimize the misclassification of only one of the two groups, when assessing credit risks or failure of new products.

The Receiver Operator Characteristic (ROC) is another well-established measure for assessing a logistic regression model's classification performance (see Hosmer et al. 2013, pp. 173 et seq. for a more detailed discussion). Within this analysis, the sensitivity (i.e., the true positive rate: the probability that a certain outcome is predicted to occur, given that the outcome occurred) and specificity (i.e., the true negative rate: the probability that a certain outcome is predicted not to occur, given that the outcome did not occur) are derived. Classification performance depends on the choice of an appropriate cutoff point (i.e., the probability that is necessary to assign a predicted outcome of either 0 or 1 to an object). Generally, an outcome of 1 is predicted for probabilities greater than 0.5. Nevertheless, the choice of an appropriate cutoff point has implications for the

classification performance in terms of sensitivity and specificity. Hence, it is worthwhile deriving a cutoff point that maximizes both sensitivity and specificity. The ROC curve considers all possible cutoff points by plotting sensitivity versus 1 - specificity and represents the likelihood that an object with an actual outcome of 1 has a higher probability than an object with an actual outcome of 0. If the logistic regression model has no predictive power, the curve has a slope of one, resulting in an area under the ROC curve of 0.5. Higher values indicate a good predictive power and typically range between 0.6 and 0.9 (Hilbe 2009, p. 258). Hosmer et al. (2013, p. 177) provide a general rule for evaluating classification performance regarding the area under the ROC curve: $ROC = 0.5$: no discrimination; $0.7 \leq ROC < 0.8$: acceptable discrimination; $0.8 \leq ROC < 0.9$: excellent discrimination; $ROC \geq 0.9$: outstanding discrimination.

Table 2 provides an overview of the different criteria, which can be used to assess the performance of a logistic regression.

Step 5: Examination and Interpretation of the Coefficients If the total model fit is considered acceptable, one can start testing and interpreting the coefficients in terms of their significance, direction, and relative importance. It should be noted that the parameter estimates of a logistic regression are much more difficult to interpret than those attained with linear regression. In linear regression, the coefficient corresponds to the absolute change in the dependent variables with a one-unit increase in the independent variables. The nonlinear nature of the logistic function makes the interpretation more difficult. For a demonstration, refer to Fig. 4 and suppose two variables X_1 and X_2 . For example, Z_i might increase from A to A' or from A' to A'' because of an increase of X_1 by one unit. The resulting change of the probability for $Y = 1$ depends heavily on the starting point of the increase – A or A'. These starting points in turn depend on the value of the other variables in the model, like X_2 in our example. Thus, in logistic regression coefficients represent only the change in the dependent variable's logit with a one-unit change in the independent variables (Aldrich and Nelson 1984, p. 41; Hair et al. 2010, p. 422; Agresti 2013, p. 163). The logit is the natural logarithm of the “chance of winning;” that is, the ratio of the probability that the dependent variable

Table 2 Acceptable ranges for logistic regression performance measures

Criterion	Range of (acceptable) values
Deviance (-2LL)	Deviance close to 0; significance level close to 100%
Likelihood-ratio test (“Model χ^2 ”)	Highest possible χ^2 value; significance level < 5%
Hosmer-Lemeshow test	Lowest possible χ^2 value; significance level close to 100%
Proportional chance criterion (PCC)	Classification should be better than the proportional chance: $\alpha^2 + (1-\alpha)^2$, with α = relative size of a group
Area under the ROC curve	$ROC = 0.5$: no discrimination $0.7 \leq ROC < 0.8$: acceptable discrimination $0.8 \leq ROC < 0.9$: excellent discrimination $ROC \geq 0.9$: outstanding discrimination

is equal to 1 divided by its counter probability (see Eqs. 26 and 27). For interpretation purposes, it may be easier to use the “odds ratios,” or effect coefficients, which are obtained by means of the e^β transformation (Hosmer et al. 2013, pp. 50–51). Specifically, odds ratios show how the odds change with a one-unit increase in the independent variables. The odds describe the ratio of the probability of occurrence of an event to its counter probability (i.e., chance) as displayed in Eq. 26. Odds range from 0 to ∞ , whereby values below one indicate that the chance of an event occurring becomes lower with increasing values of the independent variable and values above one indicate the opposite.

To verify the *significance* of the independent variables, either the Wald statistic or the likelihood-ratio test can be used. The confidence interval of individual coefficients can be determined based on the χ^2 -distributed Wald statistic, which is calculated from the square of the ratio of the coefficient and the variable’s standard error. This formula applies to metric variables with one degree of freedom only. For categorical variables, the variable’s degrees of freedom have to be considered in addition. With the likelihood-ratio test, the full model is tested against a reduced model, which is reduced by the variable under consideration. The significance test is performed on the basis of the difference between both models’ deviances, which again follows a χ^2 -distribution.

The *direction* of the variables’ effects with significant coefficients can be interpreted directly. As can be seen in Fig. 4, negative signs imply that the probability P_i decreases, while positive signs imply increasing probabilities with higher values of the variable under consideration. Statements regarding the *relative importance* of each variable can be made based on the aforementioned “odds ratios.” Their level is, however, dependent on the scaling of the variables. Furthermore, a constant change in the odds ratios does not result in a constant change in probabilities, and the magnitude of the effect on the probabilities is not symmetric around one (Hoetker 2007). Hence, alternative interpretations are discussed in the following.

In order to derive the relative importance of each predictor, different options are available. First, a *standardized coefficient* can be calculated to reflect the strength of the effect independent from the scaling of the independent variables. This effect strength can be interpreted similar to the standardized coefficients used in linear regression, as this metric specifies the number of standard deviations by which the logit changes when the independent variable increases by one standard deviation (for further details refer to Menard 2002, pp. 51 et seq.).

Marginal effects, i.e., the partial derivative of the logit function for an independent variable X_j , represent another alternative (Leclere 1992, pp. 771 et seq.). The marginal effect of X_j can be derived by applying the following formula:

$$\frac{\partial P_i}{\partial X_j} = \frac{e^{-\left(\beta_0 + \sum_{j=1}^J \beta_j X_j\right)}}{\left(1 + e^{-\left(\beta_0 + \sum_{j=1}^J \beta_j X_j\right)}\right)^2} \beta_j \quad (31)$$

Equation 31 can be reduced to:

$$\frac{\partial P_i}{\partial X_j} = \beta_j P_i (1 - P_i) \tag{32}$$

Regarding the relative importance of different variables, it has to be considered that marginal effects depend on the scale of the examined variables. Furthermore, marginal effects vary across different values of an independent variable as can be seen easily in our example in Fig. 4. Because of the different gradients of the curve in points A, A', and A'', the marginal effects at these points are substantially different. The issue regarding marginal effects' dependence on the scaling of independent variables can be resolved by standardizing the independent variables on the one hand, while on the other hand using their elasticities to interpret the logistic regression coefficients.

Elasticities are easier to interpret than the coefficients' scale-variant partial derivatives, because they are dimensionless and quantify the percentage change in the probability P_i under a 1% change in the respective independent variable of the logistic model. The elasticity of the probability P_i regarding infinitesimally small changes of X_j is obtained by adoption of the following equation (Leclere 1992, p. 772):

$$\varepsilon_{j,i} = \frac{X_j}{P_i} \frac{\partial P_i}{\partial X_j} = \frac{X_j}{P_i} \frac{e^{-\left(\beta_0 + \sum_{j=1}^J \beta_j X_j\right)}}{\left(1 + e^{-\left(\beta_0 + \sum_{j=1}^J \beta_j X_j\right)}\right)^2} \tag{33}$$

The equation can be simplified to

$$\varepsilon_{j,i} = \frac{X_j}{P_i} \frac{\partial P_i}{\partial X_j} = X_j (1 - P_i) \cdot \beta_j. \tag{34}$$

The elasticity thus results from the multiplication of the partial derivative of probability P_i (with respect to the independent variable X_j) with X_j divided by P_i . For X_j the mean value \bar{X}_j is often applied (Leclere 1992, pp. 773 et seq.). As with partial derivatives, elasticities depend on the initial values of P_i and the independent variables' values. However, the lacking dimensionality of elasticities allows direct comparisons of various independent variables' relative influence on the probability P_i .

From the perspective of non-econometricians, *sensitivity analysis* may be an appealing way of interpreting logistic regression models. In order to conduct a sensitivity analysis, the probability P_i is examined for different values of the independent variable. Usually, the value of a single independent variable is varied systematically (e.g., +10%, +20%, ..., -10%, -20%), while the other variables are kept constant. The difference between the initially estimated and

the resulting new probability P_i can then be interpreted as the relative importance of each significant independent variable for P_i . The advantage of sensitivity analysis lies in the visualization of the absolute effect of the independent variables' different values on the probability P_i (Leclere 1992, pp. 772 et seq.). In section "[Interpretation of the Coefficients](#)," we provide an example for such a sensitivity analysis.

Several authors emphasize that the interpretation of *interaction effects* in logistic regression is often insufficient, as the marginal effect of an interaction between two variables is not simply the coefficient of their interaction and the associated level of significance (e.g., Hoetker 2007). Ai and Norton (2003) show that the interaction effect of two independent variables X_1 and X_2 is the cross-derivative of P_i with respect to each. For two continuous variables X_1 and X_2 this results in the following equation:

$$\begin{aligned} \frac{\partial^2 P_i}{\partial X_1 \partial X_2} &= \beta_{12} P_i (1 - P_i) \\ &+ (\beta_1 + \beta_{12} X_2)(\beta_2 + \beta_{12} X_1) (P_i (1 - P_i)(1 - 2P_i)) \end{aligned} \quad (35)$$

Based on Eq. 35, Norton et al. (2004, p. 156) emphasize four important implications for the interpretation of interaction effects in logistic regression. First, eq. (35) shows that the interaction effect is not equal to the coefficient of the interaction β_{12} and an interaction can even exist, if $\beta_{12} = 0$. Second, the statistical significance of an interaction must be derived for the entire cross-derivative and not just the coefficient of the interaction. Third, the interaction effect is conditional on the independent variables. Fourth, Eq. 35 consists of two additive terms, which can take different signs. Accordingly, the interaction effect may have different signs for different values of the independent variables and therefore, the sign of the interaction coefficient does not necessarily represent the sign of the interaction effect. Thus, we recommend to derive the interaction effects for each observation and plot it against the predicted values of the dependent variable in order to reveal the full interaction effect (see e.g., Norton et al. 2004 for an example).

Step 6: Prediction At the beginning of section "[Logistic Regression](#)," we pointed out that logistic regression can also be used for prediction purposes. As in discriminant analysis, a holdout sample can be applied to validate the prediction performance of a logistic regression. Alternatively, cross-validation can be undertaken by means of the U-method or the jackknife method (Hair et al. 2010, pp. 374 et seq.; Aaker et al. 2011, p. 478). Examples for applying logistic regression in a prediction context include the check of creditworthiness, where financial service providers first analyze good and poor credit agreements, in order to then be in a position to assess current loan applications and their associated risks. The methods presented in the preceding sections are illustrated below in an application to sales contests among sales people.

Applied Examples

Research Question and Sample, Model, Estimation, and Model Assessment

Mantrala et al. (1998) examine the use of sales contests (SCs) in the USA and Germany. Although SCs represent commonly used motivational tools in practice, they are often neglected in research. The authors therefore develop hypotheses based on institutional economics that investigate the influence that the sales force size (x_1), the ease and adequacy of output measurement (x_2), the replaceability of sales people (x_3), and the length of the sales cycle (x_4) has on the probability that SCs are used. This might help managers to determine, whether their firm should use sales contests according to market standards. In addition to 118 observations from the USA, the study considers 270 German cases. Of the 270 original cases, 39 have missing values for at least one of the four variables. The remaining 231 observations are made up of 91 sales forces that do not use SCs (39.39%) and 140 sales forces that do deploy SCs as part of their incentive system (60.61%). The latter data set is also described in Krafft et al. (2004) and forms the basis of the following example. In our example, 231 observations are used to estimate four parameters, such that 226 degrees of freedom remain. The data set thus meets the most stringent criteria in terms of sample size and the ratio of sample size to parameters to be estimated. Table 3 provides an overview of the independent variables’ mean values and whether they differ significantly with regard to the use of sales contests. The analyses are based on one-factorial analyses of variances and provide initial evidence that the independent variables can discriminate between the groups in our sample. This initial analysis is especially suitable for a discriminant analysis, which requires a metric scale for the independent variables.

Discriminant Analysis

Model Estimation and Model Assessment

The estimation of the discriminant function yields an eigenvalue of 0.255. As mentioned in section “Discriminant Analysis Procedure,” higher values indicate a

Table 3 Determinants of the adoption of sales contests

Independent variable	Means			F	
	Total (n = 231)	No SC (n = 91)	SC (n = 140)		
Sales force size (x_1)	269.16	42.12	416.74	13.58	***
Adequacy of output measures (x_2)	0.51	0.44	0.55	18.99	***
Replaceability of sales people (x_3)	7.98	6.89	8.69	11.83	***
Length of the sales cycle (x_4)	12.27	19.08	7.85	24.85	***

*** $p < .001$

higher quality of the discriminant function. Since the measure is not standardized to values between zero and one, Wilks's lambda is a more appropriate measure to indicate a discriminant function's discriminant power. Wilks's lambda is an inverse measure of goodness, such that lower values represent a better discriminant power. In the given application, Wilks's lambda yields a value of 0.797 and indicates that the discriminant function significantly discriminates between companies that use sales contests and those that do not use them ($\chi^2(4), p < 0.001$). A classification matrix can be applied to derive the predictive power of the discriminant function and is depicted in Table 4. Given the group sizes in our example, the proportional chance criterion (PCC) is 52.25% (i.e., $0.61^2 + (1 - 0.61)^2$). In relation to the PCC, the hit rate of 69.70% (i.e., $(61 + 100) / 231$) can be considered sufficient.

Interpretation of the Coefficients

Even though Table 3 provides initial evidence about the discriminatory meaning of the variables in our example, particularly by highly significant F statistics, the interdependencies between the variables are not considered and a multivariate assessment might provide more valuable insights. Examining the standardized discriminant coefficients and the discriminant loadings, which are displayed in Table 5, is one such approach.

In our example, the standardized discriminant coefficients and discriminant loadings indicate that the length of the sale cycle and adequacy of output measures are the best predictors for the use of sales contests. Notably, slight differences can be observed with regard to the relative importance of both variables. Because

Table 4 Classification matrix regarding the (non-)adoption of sales contests after applying a discriminant analysis

Observed group membership	Predicted group membership		Correct classifications (%)
	No sales contests	Adoption of sales contests	
No sales contests	61	30	67.03
Adoption of sales contests	40	100	71.43
Total			69.70

Table 5 Discriminant coefficients and loadings

Independent variable	Standardized discriminant coefficients	Discriminant loadings
Sales force size (x_1)	0.443	0.482
Adequacy of output measures (x_2)	0.531	0.570
Replaceability of sales people (x_3)	0.311	0.450
Length of the sale cycle (x_4)	-0.526	-0.652

discriminant loadings are superior to discriminant coefficients in terms of an unbiased interpretation (see section “[Discriminant Analysis Procedure](#)”), we recommend to take them into account for assessing variables’ relative importance.

Logistic Regression

Model Estimation and Model Assessment

According to the Cook’s distance statistic, we do not find any influential outliers in our data set. After seven iterations of the logistic regression, no further improvement in the model likelihood could be achieved (improvement < 0.01%). The full model has a deviance of 232.109 (-2LL of the null model: 309.761). This relatively low value in combination with a significance level of 1.000 for the deviance indicates a good model fit. This is also confirmed by a high significance level of the Hosmer-Lemeshow test. The logistic model exhibits a likelihood ratio value of 77.652, which is highly significant ($p < 0.001$). Thus, compared to the null model, the inclusion of the four independent variables results in a significantly improved model fit. A LL_1 value of -116.055 is obtained for the full model, while an LL_0 value of -154.880 is observed for the null model. The resulting McFadden’s R^2 of 0.2507 for the final model may be viewed as relatively good, given that only four variables are used.

Considering the proportional chance criterion (PCC) of 52.25%, an overall hit rate of 75.32% of correctly classified observations ((58 + 116) / 231) indicates a good prediction performance within the calibration sample. In addition, the hit rate of 63.74% for the (smaller) group of sales forces without SCs indicates a suitable classification performance. Table 6 shows the classification table of the logistic regression in our example.

The predicted group memberships in Table 6 are based on a cutoff point of 0.5 for the predicted probability regarding the use of SCs. Consequently, if the predicted probabilities are larger than or equal to 50%, the observations are assigned to the SC group. In our example, a true positive rate (i.e., sensitivity) of 0.83 (i.e., 116 / (116 + 24)) and a true negative rate (i.e., specificity) of 0.64 (i.e., 58 / (58 + 33)) are achieved. Especially if the groups under consideration are unequal in size or if a misclassification is associated with costs, applying a different cutoff point might be more reasonable. For example, misclassification costs might appear if firms predict households’ responses to a direct mailing campaign and

Table 6 Classification matrix regarding the (non-)adoption of sales contests after applying a logistic regression

Observed group membership	Predicted group membership		Correct classifications (%)
	No sales contests	Adoption of sales contests	
No sales contests	58	33	63.74
Adoption of sales contests	24	116	82.86
Total			75.32

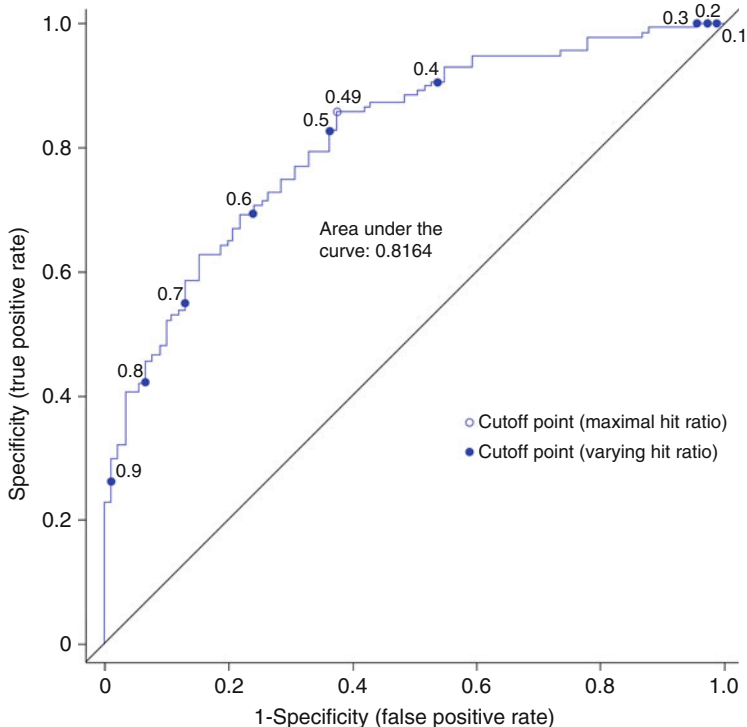


Fig. 6 Receiver Operator Characteristic (ROC) curve

seek to avoid reactance (i.e., costs) by those who are not interested in the offering, but who would be misclassified as potential respondents. In this case, the firm would try to limit the false positive rate. The Receiver Operator Characteristic (ROC) curve in Fig. 6 visualizes the trade-off between the true positive rate and the false positive rate, when different cutoff points are used in our example. If higher cutoff points are chosen, the false positive rate decreases (e.g., from 0.36 (i.e., $33 / (33 + 58)$) to 0.24 (i.e., $22 / (22 + 69)$) if a cutoff point of 0.6 instead of 0.5 is chosen). Choosing a cutoff point of 0.49 would maximize the hit ratio to 0.77 in the given setting. Generally, the area under the curve is frequently used as a meaningful measure to evaluate the predictive performance of logistic regression models. In the given example, it yields a value of 0.8164, indicating that the logistic regression provides an excellent discrimination.

Interpretation of the Coefficients

Table 7 shows the coefficients of the four variables included in the logistic regression, as well as the constant. With the exception of the significance level and the direction of the effect of each variable, a direct interpretation of the coefficients of the logistic regression model is not possible. As an indicator of the change in odds

Table 7 Elasticities and sensitivity analysis of the significant variables regarding the use of sales contests

Independent variable	Logit-coefficient	SE	Odds Ratio	Means			Elasticity (based on no SC mean)	Probability in % that SCs are used when ... is included in the logistic function ^a		
				No SC	Total	SC		No SC mean	Total mean	SC mean
Sales force size	0.0080 ^b	0.002	1.008	42.12	269.16	416.74	0.1977	41.07	80.95	93.23
Adequacy of output measures	3.0574 ^b	0.897	21.272	0.44	0.51	0.55	0.7920	41.07	46.09	49.41
Replaceability of sales people	0.1110 ^b	0.043	1.117	6.89	7.98	8.69	0.4506	41.07	44.03	45.98
Length of the sales cycle	0.0258 ^c	0.012	0.974	19.08	12.27	7.85	-0.2905	41.07	45.38	48.22
Constant	2.3121 ^b	0.652								

No SC: Non-deployment of sales contests, SC: Deployment of sales contests, SE: Standard error

^aThe respective mean value of the SC sample is replaced only for the variable under consideration (*ceteris paribus*)

^bsignificant at the 0.1 % level

^csignificant at the 5 % level

resulting from a one unit change in the predictor, the odds ratios (see Table 7) are helpful for interpreting the coefficients (see Field 2013, pp. 766 et seq. for a detailed explanation).

In our example, the odds of the use of SCs are defined as follows:

$$\text{odds} = \frac{\text{probability, that SCs are used}}{\text{probability, that no SCs are used}} \quad (36)$$

The probability of using SCs (i.e., $P(\text{SCs})$) is derived by applying the logistic regression coefficients in our example to formula (25), while the probability, that SCs are not used is simply $1-P(\text{SCs})$:

$$P(\text{SCs}) = \frac{1}{1 + e^{-(2.3121+0.008x_1+3.0574x_2+0.1110x_3-0.0258x_4)}} \quad (37)$$

The odds ratio represents the odds before and after a one unit change of the predictor variable:

$$\text{odds ratio} = \frac{\text{odds after a one unit change}}{\text{original odds}} \quad (38)$$

In our example, the odds ratio for sales force size is derived by $e^{0.0080} = 1.008$ and indicates that an increase of the salesforce by one salesperson increases the odds of using SCs by 1.008 or 0.8%. If the length of the sales cycle is increased by one month, the odds of using SCs are reduced by 0.974, equivalent to 2.6% ($e^{0.0258}$). As mentioned in section “[Logistic Regression Procedure](#),” odds ratios are dependent on the scaling of the variables. This can easily be seen when comparing the magnitude of the odds ratios with the mean values in Table 7. Further, a change of odds ratios does not result in a constant change of probabilities. Thus, in order to derive insights on the relative importance of the determinants of the use of SCs, we calculated the variables’ elasticities based on the means of the non-SC observations. If these means are used in the logistic regression function, we observe a probability of $P = 41.07\%$ that SCs are used.

As elasticities are dimensionless, they can be compared directly with each other at the absolute level. However, just like the partial derivatives, they are valid only for a certain point examined in the logistic probability function (e.g., the means of the non-SC observations in our application) and vary when other points (e.g., the means of the SC observations in our application) are examined.

All elasticities shown in Table 7 exhibit absolute values that differ substantially from 0; that is, the influence of the independent variables on the probability that SCs are deployed is comparatively high. However, there are substantial differences between the elasticities: in terms of magnitude, the strongest influence stems from the variable “adequacy of output measures,” followed by “replaceability of sales people,” “length of the sales cycle,” and “sales force size.”

In the event that $P = 41.07\%$, it may be deduced that with a 1% increase in the size of the sales force (i.e., from 42.12 to 42.54 sales people), the probability P_i that SCs are used increases by 0.20% to approximately 41.15%. If the length of the 19.08 week sales cycle is increased by 1% to approximately 19.27 weeks, the probability P_i decreases by -0.29% to 40.95%. For the other two significant variables used in our example, the effect of a 1% change in the influencing factors can be calculated analogously.

For sales managers it might be revealing to know, how the probability that SCs are used changes when the values of the independent variables are changed by a certain level (e.g., “what effect is observed on the probability that SCs are used, when the length of the sales cycle increases from 10 to 15 weeks?”). In our example of a sensitivity analysis in Table 7, the mean values of the non-SC observations are selected as the baseline again. If only the mean values of the non-SC observations are used in the logistic function, we derive a probability of 41.07% that SCs are used. For our sensitivity analysis, we now replace the mean values of the non-SC observations with the mean values of the SC observations one at a time for each independent variable. The values of the other variables are at the same time kept constant at the means of the non-SC observations. The estimated probabilities shown in Table 7 indicate that the largest influence on the probability that SCs are used emanates from “sales force size.” If the number of sales people increases from 42.12 (i.e., mean for sales forces that do not use SCs) to 416.74 (i.e., mean value for sales forces that deploy SCs), the estimated probability to use SCs rises from 41.07% to 93.23%. Notably, the variable is ranked as least important with regard to its elasticity. This can be explained by the substantial difference between the means of the non-SCs and SCs observations, which exhibit a ratio of almost 1:10. While elasticities are good at indicating a variable’s influence for a rather limited range of values, sensitivity analyses are useful to reveal probability changes for larger changes of the independent variable. The variable “adequacy of output measures” also exerts a substantial, although clearly smaller influence on the change in the probability that SCs are used. If this standardized multiple-item variable (which is normed between 0 and 1) equals 0.55 (i.e., the mean of cases where SCs are used), rather than equaling 0.44, the probability of using SCs increases from 41.07% to 49.41%. While a substantial impact is also exerted by the “length of the sales cycle” on the estimated change in the dependent variables, the variable “replaceability of sales people” exerts the lowest influence on P_i .

Logistic regression can also be used for prediction or cross-validation purposes. As part of cross-validation, the estimated coefficients can be applied to a holdout sample. In this case, only a part of the sample is used for calibrating the logistic function, which is then used to predict the actual outcomes of the remaining sample. Comparing the predicted outcomes with the actual outcomes of the holdout sample provides a good indication about the predictive power of the logistic function. Especially in direct marketing, it is crucial to know which customers will respond

Table 8 Data for a pharmaceutical company regarding the use of sales contests

Independent variable	Logit coefficient	Observation at the sample company
Sales force size	+0.0080	18
Adequacy of output measures	+3.0574	0.39 ^a
Replaceability of sales people	+0.1110	5 ^a
Length of the sales cycle (in weeks)	-0.0258	20

^aFor further information on these scales refer to Krafft et al. (2004) and the literature cited there

to a direct marketing instrument (e.g., direct mailings) in order to select the most promising customers. In this case, choosing the logistic function with the best out-of-sample prediction performance is pivotal for the success of a marketing campaign. Evaluations of the predictive performance, which are based on the calibration sample, might be particularly misleading when using models that suffer from overfitting (see section “[Logistic Regression Procedure](#)” and Tillmanns et al. 2017 for further explanations).

However, the estimated function can also be used to facilitate management decisions: companies considering whether or not to use SCs might use the findings of the logistic regression model presented in our application. Based on the attribute values given in a certain company, managers can derive the probability whether SCs are used by companies with similar attribute values. As an illustration, we use the observed attributes of a pharmaceutical company (see Table 8). The attributes of this company are then applied to Eq. 25 with the coefficients of the logistic regression model, which results in a probability of $P_i = 28.05\%$ that SCs are used. Since the observed company currently does not use SCs, the current (non-)use of sales contests corresponds to the typical practice as observed in the sample.

Conclusion

In the preceding sections, we addressed the fundamentals of discriminant analysis and logistic regression. Table 9 provides a compact overview of both methods in terms of their essential characteristics.

In principle, both methods are suitable for research questions in which the dependent variable has a categorical scale level with two or more groups. They might be applied for either classification or prediction purposes.

Compared to discriminant analysis, logistic regression has a number of key benefits, which relate particularly to the comparatively high robustness of the estimation results. For example, logistic regression allows to conduct analyses even in cases where the assumptions of discriminant analysis are violated, such as for the analysis of nonmetric independent variables (Hosmer et al. 2013, p. 22). As in linear regression, categorical variables can be analyzed using dummy variables, while in discriminant analysis, such variables would violate the assumption of homogeneous variances in the groups (Hair et al. 2010, p. 341 and p. 426).

Table 9 Overview of logistic regression and discriminant analysis

Logistic regression	Discriminant analysis
<i>Objectives</i>	
<ul style="list-style-type: none"> ■ Identification of variables that contribute to the explanation of group membership ■ Prediction of group membership for out-of-sample observations 	<ul style="list-style-type: none"> ■ Determination of linear combinations of the independent variables that optimize the separation between the groups and minimize the misclassification of observations ■ Identification of variables that contribute to explaining differences between groups ■ Predicting the group membership for out-of-sample observations
<i>Estimation principle</i>	
Maximum likelihood approach	Maximization of the variance between the groups, relative to the variance within the groups
<i>Scaling of the variables</i>	
<ul style="list-style-type: none"> ■ Dependent variable: nominal scale ■ Independent variables: metric and/or nominal scales 	<ul style="list-style-type: none"> ■ Dependent variable: nominal scale ■ Independent variables: metric scales
<i>Assessment of the significance and strength of influence of the independent variables</i>	
<ul style="list-style-type: none"> ■ Wald test, likelihood ratio test ■ Odds ratio, standardized coefficients, partial derivatives, elasticities, sensitivity analyses 	<ul style="list-style-type: none"> ■ F-test (univariate ANOVA) ■ Standardized discriminant coefficients, discriminant loadings
<i>Interpretation of the coefficients</i>	
Coefficients represent the effect of a one-unit change in the independent variables on the logit	Discriminant coefficients and weights reflect the relative average group profile
<i>Sample size (recommendations)</i>	
<ul style="list-style-type: none"> ■ A minimum of 10 observations per independent variable in the smallest group ■ Large samples sizes are recommended because of the asymptotic properties of the maximum likelihood estimates in the model parameters 	<ul style="list-style-type: none"> ■ A minimum of 20 observations per group ■ A minimum of five observations per independent variable. Better: 20 observations per independent variable to attain stable estimates for the standardized coefficients and weights
<i>Assumptions/recommendations</i>	
<ul style="list-style-type: none"> ■ Nonlinear relationships ■ No multicollinearity ■ Errors are independent ■ Linearity of the logit 	<ul style="list-style-type: none"> ■ Linear relationships ■ No multicollinearity ■ Multivariate normal distribution of the independent variables ■ Homogeneity of the variance-covariance matrices of the independent variables

Further, the assumption of multivariate normally distributed independent variables, which is required for the use of discriminant analysis, is frequently not met in practical applications. In these situations, logistic regression is also preferable. The same holds for studies in which the analyzed groups have very different sizes (Tabachnik and Fidell 2013, p. 380). Another important advantage of logistic regression is that through the regression procedure, asymptotic t-statistics can be

provided for the estimated coefficients. The confidence intervals obtained in discriminant analysis are, on the other hand, not interpretable (Morrison 1969, pp. 157 et seq.).

Compared to discriminant analysis, logistic regression is thus an extremely robust estimation method. However, it should not be concluded that logistic regression is always the best choice. Discriminant analysis might provide more efficient estimates with higher statistical power if group sizes do not turn out to be too unequal and in cases where the assumptions of discriminant analysis are met (Press and Wilson 1978, p. 701; Tabachnik and Fidell 2013, p. 380 and p. 443). Further, because of the asymptotic properties of the maximum likelihood estimation, the use of logistic regression often requires larger sample sizes than discriminant analysis. Additionally, researchers should examine whether the assumed nonlinear development of the probability P_i is suitable for the specific research context. If, for example, a linear change of P_i is more appropriate, the logistic regression should be avoided. Instead, researchers should check whether linear approaches such as the linear probability model (LPM) or discriminant analysis are more appropriate (Aldrich and Nelson 1984).

References

- Aaker, D. A., Kumar, V., Day, G. S., & Leone, R. P. (2011). *Marketing research* (10th ed.). New York: Wiley & Sons.
- Afifi, A., May, S., & Clark, V. A. (2012). *Practical multivariate analysis* (5th ed.). Boca Raton: Taylor & Francis.
- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken: Wiley-Interscience.
- Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1), 123–129.
- Aldrich, J., & Nelson, F. (1984). *Linear probability, logit, and probit models*. Beverly Hills: SAGE.
- Anderson, E. (1985). The salesperson as outside agent or employee: A transaction cost analysis. *Marketing Science*, 4(3), 234–254.
- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2016). *Multivariate Analysemethoden* (14th ed.). Berlin: Springer.
- Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave MacMillan.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
- Field, A. (2013). *Discovering statistics using IBM SPSS* (4th ed.). London: Sage.
- Frenzen, H., & Krafft, M. (2008). Logistische Regression und Diskriminanzanalyse. In A. Herrmann, C. Homburg, & M. Klarmann (Eds.), *Handbuch Marktforschung – Methoden, Anwendungen, Praxisbeispiele* (pp. 607–649). Wiesbaden: Gabler.
- Hair, Jr, J., Black, W. C., Babin, B. J., Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River: Pearson Prentice Hall.
- Hilbe, J. (2009). *Logistic regression models*. Boca Raton: CRC Press.
- Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal*, 28(4), 331–343.
- Hosmer, D., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken: Wiley.

- King, G., & Zeng, L. (2001). Explaining rare events in international relations. *International Organization*, 55(3), 693–715.
- Krafft, M., Albers, S., & Lal, R. (2004). Relative explanatory power of agency theory and transaction cost analysis in German salesforces. *International Journal of Research in Marketing*, 21(4), 265–283.
- LeClere, M. (1992). The interpretation of coefficients in models with qualitative dependent variables. *Decision Sciences*, 23(3), 770–776.
- Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using stata* (3rd ed.). College Station: Stata Press.
- Mantrala, M., Krafft, M., & Weitz, B. (1998). Sales contests: An investigation of factors related to use of sales contests using German and US survey data. In P. Andersson (Ed.), *Proceedings Track 4 – Marketing management and communication, 27th EMAC conference, Stockholm*, pp. 365–375.
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks: Sage.
- Morrison, D. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, 6(2), 156–163.
- Norton, E. C., Wang, H., & Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *Stata Journal*, 4(2), 154–167.
- Osborne, J. W. (2008). *Best practices in quantitative methods*. Los Angeles: Sage.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12), 1373–1379.
- Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), 699–705.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.
- Tatsuoka, M. M. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan.
- Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology*, 76(1), 103–154.
- Tillmanns, S., Ter Hofstede, F., Krafft, M., & Goetz, O. (2017). How to separate the wheat from the chaff: Improved variable selection for new customer acquisition. *Journal of Marketing*, 80(2), 99–113.
- Vittinghoff, E., & McCulloch, C. E. (2006). Relaxing the rule of ten events per variable in logistic and cox regression. *American Journal of Epidemiology*, 165(6), 710–718.



Multilevel Modeling

Till Haumann, Roland Kassemeyer, and Jan Wieseke

Contents

Introduction: Relevance of Multilevel Modeling in Marketing Research	370
Fundamentals of Multilevel Modeling	372
The Conceptual Relevance of Multilevel Modeling	372
The Statistical Relevance of Multilevel Modeling	375
Types of Constructs and Models in Multilevel Modeling	378
Process of Multilevel Modeling: The Two-Level Regression Model	381
Step 1: Baseline Model	382
Step 2: Adding Independent Variables at Level 1	383
Step 3: Adding Independent Variables at Level 2	383
Step 4: Testing for Random Slopes	384
Step 5: Adding Cross-Level Interaction Effects	385
Assumptions of Multilevel Modeling	386
Model Estimation & Assessing Model Fit	386
Variable Centering	389
Sample Size Considerations	390
Multilevel Structural Equation Modeling	392
Software for Estimating Multilevel Models	396
Example: Building and Estimating a Two-Level Model	398
Conclusions	402
Cross-References	402
Appendix	403
References	404

T. Haumann (✉)

South Westphalia University of Applied Sciences, Soest, Germany
e-mail: haumann.till@fh-swf.de

R. Kassemeyer

Marketing Group, Warwick Business School, University of Warwick, Coventry, UK
e-mail: roland.kassemeyer@wbs.ac.uk

J. Wieseke

Sales Management Department, University of Bochum, Bochum, Germany
e-mail: jan.wieseke@rub.de

Abstract

Many phenomena in marketing involve multiple levels of theory and analysis. Adopting a multilevel lens to marketing phenomena can often yield richer and more rigorous results. However, the consideration of multiple levels of theory and analysis often leads to the challenge to cope with nested data structures in which a lower level unit of analysis is nested within a higher level unit of analysis. Explicitly acknowledging such nested data structures is important as its analysis with single level analysis techniques may result in biased results and thus incorrect conclusions because nested data structures often violate assumptions of conventional single level analysis techniques. A methodological approach which explicitly accounts for multiple levels of analysis and thus the nested structure of data is referred to as multilevel modeling. This chapter attempts to help researchers and practitioners interested in investigating multilevel phenomena by providing an introduction to multilevel modeling. It therefore describes the theoretic fundamentals of multilevel modeling by outlining the conceptual and statistical relevance of multilevel modeling. Furthermore, it provides guidance how to build a multilevel regression model using a step-by-step approach. The chapter also discusses how to assess the fit of multilevel models, how to center variables at different levels of analysis, and how to determine the sample sizes to adequately estimate multilevel models. Moreover, it offers insights how the logic of multilevel regression analysis could be expanded to multilevel structural equation modeling, discusses different statistical software packages that can be employed to estimate multilevel models, and provides a detailed example of building and estimating a multilevel model.

Keywords

Random coefficient modeling · Hierarchical linear modeling · Nested data structures · Hierarchical data · Between variance · Within variance · Random intercept · Random slope · Cross-level interaction · Intraclass correlation coefficient · Group mean centering · Grand mean centering

Introduction: Relevance of Multilevel Modeling in Marketing Research

Many phenomena in marketing which raise the interest of marketing researchers and practitioners involve multiple levels of theory and analysis. For example, marketing researchers and practitioners may be interested investigating which skills salespeople do need to service their customers best (Homburg et al. 2011), how marketing and/or sales managers may behave to increase their employees' performance (Wieseke et al. 2009), or to determine the effectiveness of different instruments of the marketing mix in managing brands across different countries (Steenkamp et al. 2010). Indeed, all of these questions pertain to multiple levels – i.e., the salesperson and the customer level, the manager and the employee level, and the customer and

the country level. Recognizing the multilevel nature of marketing phenomena can be important for at least two reasons.

First, the recognition and inclusion of different levels of theory often has the potential to provide richer and more rigorous insights of the studied phenomenon. It may provide richer insights because acknowledging the multilevel nature of a phenomenon allows the identification of contextual factors which influence the phenomenon under investigation. Or as Hitt et al. (2007, p. 1385) put it: “Using a multilevel lens reveals the richness of social behavior; it draws our attention to the context in which behavior occurs and illuminates the multiple consequences of behavior traversing levels of social organization. For management to continue advancing as a field in which scholars seek to explain the behaviors of individuals, groups, and organizations, we must expand our theories and empirical investigations to encompass these multilevel effects.” Originally referring to management, this conclusion also strongly pertains to marketing research. Adopting a multilevel lens may also provide more rigorous insights because it is associated with collecting data at multiple levels. Such multilevel data collections can assure that each variable is measured at the adequate level (e.g., salesperson variables at the salesperson level and customer variables at the customer level) (Klein et al. 1994) and often result in multisource data sets benefiting from advantageous characteristics such as a lower susceptibility to common method bias (Johnson et al. 2014).

Second, it is important to account for the multilevel nature of marketing phenomena because acknowledging multiple levels of analysis may have consequences for the adequate methodological approach. Specifically, multilevel phenomena are characterized by hierarchical data structures in which lower entities are nested within higher entities (e.g., customers nested within salespersons or employees nested within managers) (Heck and Thomas 2015). To acknowledge such nested data structures is important because they may violate assumptions of conventional single level analysis techniques and thus may result in biased standard errors and in spuriously significant results (Hox et al. 2018; Wieseke et al. 2008).

A methodological approach to adequately account for the complexity associated with hierarchical nested data structures is referred to as multilevel or random coefficient modeling (Raudenbush and Bryk 2002; Hox et al. 2018). It can conceptually be regarded as a system of hierarchical regression equations. In such a hierarchical system of regression equations, one equation captures the influence of variables at the lower level, while one or more equations refer to the influences at the higher level (Hox et al. 2018). Thereby, multilevel modeling can help to assure the integrity between multilevel theory and multilevel analysis and thus help to unfold the full potential of a multilevel lens to marketing phenomena.

This chapter attempts to help researchers and practitioners interested in investigating multilevel phenomena by providing an introduction to multilevel modeling. We begin with discussing the “[Fundamentals of Multilevel Modeling](#)” and describe its conceptual and statistical relevance. Furthermore, we discuss core types of constructs and models in multilevel contexts. Then, the section “[Process of Multilevel Modeling: The Two-Level Regression Model](#)” begins with offering a detailed description of the steps of estimating a two-level regression model and provides

information on how the model fit of a multilevel model is assessed. Afterwards, we discuss different approaches of centering variables at lower and higher levels and provide insights into sample size requirements for adequately analyzing multilevel models. The Section “[Multilevel Structural Equation Modeling](#)” extends the multilevel regression approach and explains how to apply multilevel structural equation modeling. In Section “[Software for Estimating Multilevel Models](#)” we provide information on core software packages for multilevel analysis and in section “[Example: Building and Estimating a Two-Level Model](#)” we offer a detailed example of building and estimating a multilevel model. Finally, we conclude this chapter by offering an overview of the most important terms introduced in this chapter together with their definitions.

Fundamentals of Multilevel Modeling

The Conceptual Relevance of Multilevel Modeling

The adequate analysis of nested data structures is important because such data structures permeate marketing research (Wieseke et al. 2008). Specifically, marketing research in fields such as relationship marketing, international marketing, personal selling, sales management, services, organizational research in marketing, research based on secondary data (e.g., research at the interface of marketing and finance), longitudinal marketing phenomena, or meta-analyses attempts to address research questions which involve nested data structures and multiple level of analysis. Table 1 provides an overview of exemplary nested data structures in the aforementioned fields of marketing research and additionally offers examples of papers that conducted a multilevel approach to analyze their data.

Most of the nested data structures presented in Table 1 and thus the conceptual relevance of multilevel modeling primarily stems to large extent from the hierarchical structure of (marketing) organizations. In this respect Klein et al. (1994, p. 198) state that “By their very nature, organizations are multilevel. Individuals work in dyads, groups, and teams within organizations that interact with other organizations both inside and outside the industry. [...]. To examine organizational phenomena is thus to encounter levels issues.” In the following, we outline the conceptual relevance of multilevel modeling in marketing research by highlighting how different entities are integrated in the conceptual structure of a marketing organization and how they are thus nested in higher level entities (e.g., Hitt et al. 2007; Wieseke et al. 2008). Figure 1 illustrates an exemplary hierarchical structure in marketing organizations.

Figure 1 visualizes customers as the centroid of the exemplary multilevel structure. As salespeople’s behavior is crucial for customer perceptions of the selling interaction (e.g., Homburg et al. 2009a), the company’s products (e.g., Goff et al. 1997), and the organization as a whole (e.g., Homburg et al. 2011), customers are nested within salespeople. Consequently, the aforementioned customer perceptions depend in part on the salespeople by whom they are served. The study of Mikolon

Table 1 Exemplary nested data structures in marketing research

Research area	Exemplary nested data structures ^a		Exemplary papers
	Lower level	Higher level	
Relationship Marketing	Customers	Organizations	Homburg et al. 2009b; Maxham et al. 2008; Netemeyer et al. 2012; Palmatier 2008
International Marketing	Individuals (e.g., salespeople, customers)	Countries Cultures	Hohenberg and Homburg 2016; Steenkamp et al. 2010; Walsh et al. 2014
Personal Selling	Customers	Salespeople	Homburg et al. 2011; Mikolon et al. 2020; Wieseke et al. 2014
Sales Management	Salespeople	Leaders Territories Sales Teams	Ahearne et al. 2010; Auh et al. 2014; Mathieu et al. 2007; Wieseke et al. 2009; Van der Borgh et al. 2019
Service Research	Customers Frontline Employees	Service Providers Organizations	Brady et al. 2012; Donovan et al. 2004; Mikolon et al. 2015
Organizational Research in Marketing	Frontline Employees	Work Groups Subsidiaries	de Jong et al. 2004; Homburg et al. 2011; Liao and Chuang 2007; Wieseke et al. 2012
Marketing-Finance Interface	Firms	Industries	Anderson et al. 2004; Groening et al. 2016; Gruca and Rego 2005; Josephson et al. 2016; Larivière et al. 2016; Misangyi et al. 2006
Longitudinal Marketing Phenomena	Observations in Time	Individuals (e.g., salespeople, customers)	Boichuk et al. 2014; Fu et al. 2010 Lam et al. 2013
Meta-Analysis in Marketing Research	Effect Sizes/ Relationships	Samples/ Studies	Edeling and Fischer 2016; Krasnikov and Jayachandran 2008; Roschk and Hosseinpour 2020; Troy et al. 2008

Note: ^aIncluded levels of analysis (i.e., nestings) are only exemplary and for illustrative purposes. The cited article may include further other levels of analysis.

et al. (2015), for example, investigates data from customers nested within service providers. The study uses dyadic data of 310 customers that interacted with 108 service providers to explore the role of complexity in professional service encounters.

Further, Figure 1 illustrates salespeople’s complex integration within their organization. First, salespeople work in sales teams which are part of larger organizational units (e.g., departments), which in turn are nested in organizations which are nested in environments (e.g., industries). Consequently, salespeople are part of a larger collective that may share values, attitudes, cognitions, experiences, perceptions or behaviors (Kozlowski and Klein 2000). The study of Ahearne et al. (2010) provides an example of such a nested structure by investigating a data set comprising 1070 sales representatives nested within 185 selling teams.

Second, sales managers are usually responsible for sales teams and department managers for company departments. Therefore, salespeople are nested in their

Exemplary Multilevel Structure in Marketing Organizations



Fig. 1 Exemplary multilevel structure in marketing organizations. (Note: Adapted from Hitt et al. (2007))

direct supervisors, who again may be nested in their department managers. Department managers in turn may have to report to the organization's chief marketing officers at the top of the organization. As supervisors oftentimes have an enormous impact on their subordinate employees (e.g., Bass and Bass 2009), it is important to account for important supervisor characteristics when investigating processes at lower organizational levels. An example yielding support for the importance of supervisors in affecting their subordinates is provided by the study of Wieseke et al. (2009).

Furthermore, Figure 1 illustrates that customers, salespeople, sales teams, and departments are nested in organizations. Thus, researchers that are interested in investigating inter-organizational differences of organizational entities or customers will be confronted with the nesting of these entities in organizations. Palmatier's (2008) investigation of interfirm relational drivers of customer value reflects an example of the analysis of such data where business-to-business customers ($n = 466$) are nested in different organization ($n = 27$).

Finally, Figure 1 shows that organizations are nested in environments. Environments could be the industry, country, culture or market in which an organization operates. The study of Gruca and Rego (2005) provides an example as it investigates how consequences of customer satisfaction on shareholder value differentiate between industries. The authors' analyze data of 840 firm-year observations in 23 industries.

In addition to the hierarchical structure of marketing organizations, there are further study designs that are associated with a nested data structure (Deadrick et al. 1997). For example, longitudinal data or data of repeated measures can be considered as hierarchical because repeated measurements are nested within individuals (Hox et al. 2018). When investigating longitudinal data employing a multi-level modeling approach, the series of repeated measures can be modeled at the

lower level and the individual subjects at the higher level. The consideration of repeated measures of individuals over time allows researchers to investigate “the existence, nature, and causes of within-person [...] changes over time” (Deadrick et al. 1997, p. 748).

The study of Fu et al. (2010) provides an example for the application of multilevel modeling for the examination of longitudinal data. The study comprises survey data of 534 salespeople and corresponding performance data of salespeople’s daily sales during two product’s first several months in the market. The authors employ a growth curve model to examine how the continuous outcome daily sales changes over time and how it is affected by salesperson variables.

Another study design in which multilevel modeling can yield richer and more accurate results refers to meta-analytical investigations (a general introduction to meta-analyses in marketing research is provided by Bijmolt (2021) in the online version of this handbook). A meta-analysis reflects a systematic approach towards synthesizing a larger number of results from empirical studies to summarize findings of a specific research question (Glass 1976; Hox et al. 2018; Lipsey and Wilson 2001). Meta-analyses may be characterized by nested data structures as multiple effect sizes of relationships under investigation are nested within studies (Bijmolt and Pieters 2001). Adopting a multilevel approach recognizes these nestings and additionally allows to include differences between studies (e.g., research operational factors such as measurement approaches, environmental factors such as industries or cultures, or sample or manuscript related factors such as socio-demographical differences between samples or manuscript statuses [e.g., published vs. unpublished]). Methodologically, a multilevel approach to meta-analysis includes effect sizes of relationships under investigation at the lower level and between study characteristics at the higher level (Hox et al. 2018). Thereby, the study characteristics can be used as explanatory variables to explain differences in the investigated relationships between studies (Hox et al. 2018).

An example for applying multilevel modeling to conduct a meta-analysis reflects the study of Troy et al. (2008). In this study, the authors analyze 146 correlations of 25 studies of cross-functional integration and new product success. They use multilevel modeling to investigate moderating influences under which the examined relationship of cross-functional integration and new product success varies.

Thus, in sum, focusing solely on a single level of theory often impedes a profound understanding of complex phenomena prevalent in marketing research (Wieseke et al. 2008). Returning to the introductory thoughts, including the (multiple) relevant levels of analysis into one’s theoretical framework has the potential to increase the validity of its findings and thus its contribution to academic marketing research and to provide richer insights for marketing practitioners.

The Statistical Relevance of Multilevel Modeling

The following section describes the methodological challenges that exist when analyzing hierarchical data structures. Figure 2 illustrates a hierarchical data

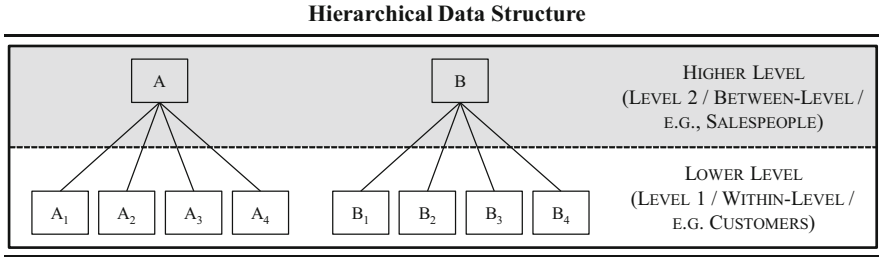


Fig. 2 Hierarchical data structure

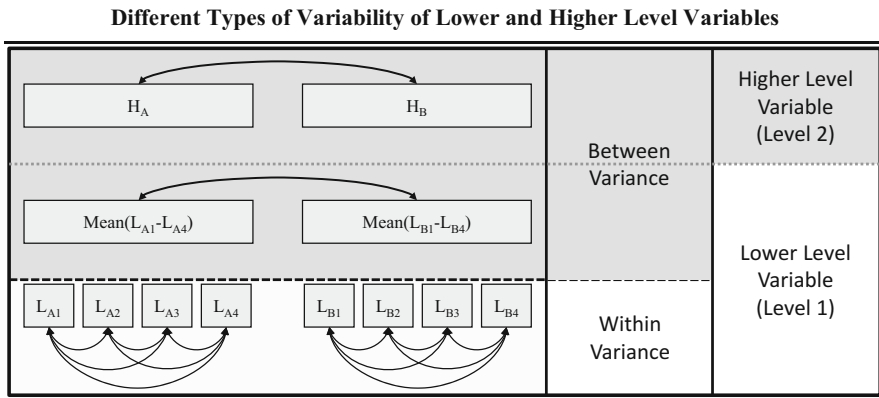


Fig. 3 Different types of variability of lower and higher level variables

structure with two levels. In hierarchical data structures, usually one entity at the higher level (A; e.g., a salesperson) relates to multiple entities at the lower level (A₁-A₄; e.g., consumers served by the same salesperson). In such nested data structures, the association between a specific higher level entity (A) and its nested lower level entities (A₁-A₄) is referred to as a cluster (cluster A).

In hierarchical data structures, in which multiple entities from the lower level are nested within a specific entity at the higher level, one can distinguish between different types of variability of variables at the lower and the higher level. Figure 3 visualizes the different types of variability of lower and higher level variables.

Specifically, in a two level context, lower level variables (L) are characterized by two types of variability, variability at the lower level (Level 1) which captures the variability of a variable within a cluster (*within variance* between L_{A1}-L_{A4} and L_{B1}-L_{B4} at Level 1) and variability at the higher level which captures the variability of the lower level variable between clusters (*between variance* between the cluster mean of cluster A and B at Level 2). In contrast, higher level variables (H) only possess variance at the higher level (*between variance* between H_A and H_B at Level 2).

Turning to a mathematical formulation, these types of variability can be expressed by decomposing the total variance of a variable in its parts at the lower level (i.e., the

within variance) and the higher level (i.e., the between variance). Consequently, the variance of a variable x can be written as:

$$V_T(x) = V_W(x) + V_B(x) \quad (1)$$

If x denotes a variable at the higher level, $V_T(x)$ equals $V_B(x)$ since higher level variables possess no lower level variability. However, variables at the lower level often possess meaningful variability at both levels.

This complex variance structure of lower and higher variables can have consequences for the analysis of data. First, lower level observations belonging to different clusters may not be independent from each other (which would be reflected in a meaningful between variance of the lower level variable). However, independence of observations is a core assumption of many single-level methods and a violation of this assumption could lead to biased standard errors and incorrect hypothesis tests (e.g., Hox et al. 2018; Wieseke et al. 2008).

Second, if one is interested in modeling the influence of a higher level variable at a lower level variable using a disaggregated approach (i.e., by disaggregating the higher level variable to the lower level), conventional single-level methods would erroneously assume the disaggregated higher level observations to be independent information and use the larger lower level sample size for hypothesizing testing rather than the adequate higher level sample size. Erroneously using the larger sample size may lead to downward-biased standard errors and thus again to incorrect hypotheses tests (e.g., Hox et al. 2018; Wieseke et al. 2008).

These potential biases of using single-level methods to analyze hierarchical, nested data structures clearly highlight the importance of adopting a methodological approach that is able to adequately handle such multilevel data structures. This is especially true if one is interested in analyzing models with variables from multiple levels.

If one is only interested in analyzing relationships between lower level variables which are however part of a nested data structure, the biases associated with employing a single level method depends on the degree of non-independence of observations and thus of the proportion of between variance of the lower level variables. This proportion of between variance (V_B) compared to the total variance (V_T) of a lower level variable is often referred to as intraclass correlation (ρ ; also referred to as intraclass correlation coefficient [ICC]) and is especially helpful in evaluating the necessity of adopting a multilevel modeling technique for investigating single level models testing relationships between lower level variables from a nested data structure. Using mathematical notation, the intraclass correlation for a two-level data structure can be expressed as:

$$\rho = \frac{V_B}{V_T} = \frac{V_B}{(V_W + V_B)} \quad (2)$$

where V_T reflects the total variance, V_B the between variance, and V_W the within variance of a lower level variable (to compute intraclass correlations for three-level

models see for example Hox et al. (2018)). The higher the intraclass correlation the higher the non-independence of observation, the higher the potential biases, and thus the more important it is to employ a multilevel modeling approach. Hox (2010) suggests 0.05, 0.10, and 0.15 as small, medium, and large values for the intraclass correlation. However, even for small intraclass correlations biases may be substantial if cluster sizes are large. Therefore, a second helpful metric to determine the necessity to adopt a multilevel modeling approach is the design effect which additionally includes information about the average cluster sizes.

The design effect reflects “the ratio of the variance of estimation obtained with the given sampling design, to the variance of estimation obtained for a simple random sample from the same population, and with the same total sample size” (Snijders and Bosker 2012, p. 287). The higher the design effect, the higher the variation between clusters and the more relevant it is to employ a multilevel modeling approach. Design effects can be calculated as a function of the intraclass correlation (ρ) and the average cluster size (c) (Muthén and Satorra 1995):

$$\text{deff} = 1 + (c - 1) * \rho \quad (3)$$

A commonly used rule of thumb is that a design effect greater than 2 indicates that the considered variable varies substantially between clusters and, thus, a multilevel approach is necessary to analyze the data. However, as it is only a rule of thumb, Lai and Kwok (2015) have shown that researchers should refrain from using this rule if they are interested in effects of higher-level predictors or investigate data with a cluster size that is less than 10.

Types of Constructs and Models in Multilevel Modeling

Depending on the specific research question, one may distinguish between different types of constructs and models in multilevel modeling. We first discuss constructs referring to individuals and collectives and then distinguish between different types of multilevel models.

Types of constructs in multilevel modeling. In multilevel modeling, one can distinguish between constructs that refer to an individual and constructs that refer to a larger collective. Individual-level constructs refer to characteristics, experiences, attitudes, perceptions, values, cognitions, or behaviors of individuals (e.g., a customer’s satisfaction or a salesperson’s customer orientation) and can be relevant at different levels of a multilevel model (Kozlowski and Klein 2000). For example, in a context with sales managers supervising multiple salespeople one may be interested whether a sales manager’s organizational identification at the higher level may affect the salespeople’s level of organizational identification at the lower level (Wieseke et al. 2009).

Beyond individual level constructs, multilevel models are especially helpful for studying larger collectives (e.g., teams, departments, organizations, or industries). Constructs describing such larger collectives can refer to their (1) global, (2) shared, or (3) configural properties (Kozlowski and Klein 2000).

First, global properties refer to descriptive or easily observable characteristics of a collective (e.g., the number of salespersons in a sales team or the type of sales context [B2C vs. B2B]). Such information can be retrieved from archival data or measured using a key informant approach (e.g., the sales manager responsible for the respective sales team).

Second, shared properties emerge from experiences, attitudes, perceptions, values, cognitions, or behaviors that are held in common by a collective of individuals (e.g., a service- or team-climate within a sales team) (Kozlowski and Klein 2000). Such constructs are typically assessed at the individual level and then aggregated to the adequate level of theory and analysis, i.e., the higher level. When employing shared properties it is important to assure that the theoretical foundation resides at the higher level, that the measurement refers to the entity at the higher level (e.g., the group), and that the empirical aggregation of the individual observations is justified. To justify the aggregation of lower level observations to a higher level, researchers often refer to the level of within group or agreement, r_{wg} or $r_{wg(j)}$ (James et al. 1984, 1993), the proportion of within to between variance of the variable, i.e., the intraclass correlation (ICC1), and the reliability of the group mean (ICC2) (Bliese 2000). A description of these criteria and their specific formulas appears in the Appendix. More detailed discussions about these and further criteria to justify aggregation can for example be found in Woehr et al. (2015), LeBreton and Senter (2008), LeBreton, James, and Lindell (2005), or Bliese (2000).

Third, configural properties also refer to experiences, attitudes, perceptions, values, cognitions, or behaviors. However, in contrast to shared properties, configural properties do not reflect agreement or common understanding but describe a pattern or variability between the individuals of the respective collective. A typical example for a configural property is a form of consensus within a collective. Ahearne et al. (2010), for instance, study how interpersonal climate consensus and the consensus regarding leadership empowerment behaviors in sales teams influence the effect of the respective shared properties (i.e., both group-level perceptions) on team potency. Another typical example for a configural property from service research refers to the consensus of service climate within a group of service employees (e.g., Schneider et al. 2002).

Types of models in multilevel modeling. Multilevel models may focus on relationships between constructs on a single level or on relationships across levels (Kozlowski and Klein 2000). In single-level models, researchers are solely interested in the relationships between constructs of one level of a nested data structure. Despite this focus on a single level, it might yet be advantageous to adopt a multilevel modeling approach. Specifically, even if one is solely interested in relationships between lower level variables (e.g., how individual salespersons' customer orientation affect their individual level of sales performance), it can be worthwhile to adopt a multilevel modeling approach. First, as mentioned before the nested data structure may violate the assumption of independence between observations of conventional single level analysis techniques. Second, if observations are not independent from each other, adopting a multilevel modeling approach can account for unobserved heterogeneity between the lower level observations (e.g.,

Hamaker and Muthén 2020). However, if one wants to analyze a single-level model and is solely interested in relationships between higher level variables and if the level of theory and analysis of the included constructs reside at this level and no higher levels exist, a single-level analysis technique can be employed (Kozlowski and Klein 2000).

In contrast to single-level models, cross-level models include relationships between variables of different levels. In cross-level models, a higher level predictor variable may either directly affect an outcome variable at the lower level or moderate a relationship between two variables at the lower level, which is typically referred to as a cross-level interaction. For example, a sales manager's empowering leadership may affect her/his followers' self-efficacy using a new sales technology system and may additionally moderate the relationship between salespersons' prior work experience and their technology self-efficacy (Mathieu et al. 2007).

Another type of multilevel models includes predictor and outcome variables from multiple levels but do not assume cross-level relationships. A specific type of such models, which gained some interest in multilevel research, are homologous models. Homologous models examine whether a process at the lower level (e.g., the individual level) is consistent with a similar process at the higher level (e.g., the team level) (Chen et al. 2005). For example, such a model may examine whether the effect of individual feedback on individual self-efficacy is consistent with the effect of team feedback on collective self-efficacy (Chen et al. 2005). Figure 4 offers a graphical overview of the discussed models for which employing a multilevel analysis technique is recommended.

Beyond these models, another widely used application of multilevel modeling is the analysis of longitudinal data. Longitudinal data has a nested structure such that repeated observations over time are nested in an entity – for example daily data on sales performance nested within a salesperson (Fu et al. 2010). Consequently, multilevel models to analyze longitudinal data describe the development of the outcome variable over time at the lower level. This development in its most simple form is characterized by a (random) intercept and slope parameter. Additionally, time-varying covariates may be added at the lower level. At the higher level, time-invariant covariates, that may help to explain the variance in the random parameters of the lower level, can be added to the model.

Consequently, such models can be used to effectively analyze panel data (please see ► [“Panel Data Analysis: A Non-Technical Introduction for Marketing Researchers”](#) in this handbook for a general introduction to panel data analysis and an example of multilevel modeling of panel data). Furthermore, such models may be easily extended to account for additional levels (e.g., salespeople in teams) or more complex patterns of the development of the outcome variable over time, such as quadratic or cubic patterns. This opportunity to analyze different patterns over time makes multilevel modeling also an alternative to structural equation modeling to estimate latent growth models (Hox et al. 2018). More detailed discussions of multilevel models to analyze longitudinal data and estimate latent growth models can for example be found in Hox (2011), Hedeker and Gibbons (2006), Duncan, Duncan, and Stryker (2013), Stoel and Garre (2011), and Hox et al. (2018).

Core Types of Multilevel Models

A: Single-level model – lower level predictor-outcome relationship with random intercept and slope	
Higher Level	
Lower Level	
Model Notation: L1: $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$; L2: $\beta_{0j} = \gamma_{00} + u_{0j}$; $\beta_{1j} = \gamma_{10} + u_{1j}$	
B: Cross-level model – direct effects of a lower and a higher level predictor on a lower level outcome variable	
Higher Level	
Model Notation: L1: $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$; L2: $\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$; $\beta_{1j} = \gamma_{10} + u_{1j}$	
C: Cross-level model including a cross-level interaction effect.	
Higher Level	
Model Notation: L1: $Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$; L2: $\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j}$; $\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j}$	
D: Homologous Model – similar model structure for the individual (I) level and the collective (C) level	
Higher Level	
Model Notation: L1: $Y_{ij}^l = \beta_{0j}^l + \beta_{1j}^l X_{ij}^l + e_{ij}$; L2: $\beta_{0j}^l = \gamma_{00} + u_{0j}$; $\beta_{1j}^l = \gamma_{10} + u_{1j}$; $Y_j^c = \beta_{0j}^c + \beta_{1j}^c Z_j^c + u_j^c$	

Fig. 4 Core types of multilevel models. (Notes: In this overview we exemplary focus on models with two levels. In all described models, we allow the intercepts and slopes to vary across higher-level entities. Multilevel models may of course include further interactions of variables residing at the same level, which are not included here. In describing the homologous model in Panel D we assume that the variables at the lower and higher level are measured separately)

Process of Multilevel Modeling: The Two-Level Regression Model

The major advantage of multilevel modeling is the opportunity to estimate effects between variables on different levels of analysis while accounting for the hierarchical data structure. However, due to consideration of multiple levels of analysis, the estimation of multilevel models oftentimes involves a high degree of complexity. Therefore, multilevel models are usually estimated in a stepwise process. In the following, we outline the five-step process of estimating a two-level regression model as suggested by Hox et al. (2018) (please refer to Skiera, Jochen Reiner, and Albers (2021) in this handbook for a general introduction to regression analysis). In describing each step, we adopt a notation in which we outline the equation at the lower level (i.e., Level 1 [L1]), at the higher level (i.e., Level 2 [L2]), and then provide an integrated regression equation (I). After explaining the stepwise process

of the two-level model, we briefly discuss the assumptions of the multilevel regression model. In addition to the statistical description of the stepwise-estimation of a multilevel model, the section “[Building and Estimating a Two-Level Model](#)” provides an example of this process in a marketing context.

Step 1: Baseline Model

Step 1 reflects the estimation of the baseline model. The baseline model (also referred to as intercept-only or random-intercept model) does not include any explanatory variables and does only include the intercept and the residuals at Level 1 and 2. Equation (4) describes the Level 1 model, Equation (5) the Level 2 model, and Equation (6) the integrated form of the baseline model of a two-level regression analysis.

$$L1 : Y_{ij} = \beta_{0j} + e_{ij} \quad (4)$$

$$L2 : \beta_{0j} = \gamma_{00} + u_{0j} \quad (5)$$

$$I : Y_{ij} = \gamma_{00} + e_{ij} + u_{0j} \quad (6)$$

In Equation (4) Y_{ij} represents the value of the dependent variable of observation i ($i = 1, \dots, n_j$; e.g., customers) in cluster j ($j = 1, \dots, J$; e.g., salespersons). β_{0j} reflects the (random) intercept-term at Level 1 which is allowed to vary between clusters as expressed in Equation (5) (see Fig. 4a for a graphical illustration of a random-intercept model). e_{ij} denote the Level 1 residuals which are assumed to have an expected mean of zero ($E(e_{ij}) = 0$) and a variance σ_e^2 ($Var(e_{ij}) = \sigma_e^2$). As the baseline model does not include any exploratory Level 1 variables this residual variance σ_e^2 reflects the within variance of the dependent variable Y_{ij} at Level 1.

Equation (5) represents the equation for the (random) intercept. Here, γ_{00} reflects the intercept value and u_{0j} the residuals of the intercept equation at Level 2. Analogously to the Level 1 residuals, the residuals of the random intercept equation at Level 2 are assumed to have an expected mean of zero ($E(u_{0j}) = 0$) and a variance τ_{00} ($Var(u_{0j}) = \tau_{00}$). Again, as the baseline model contains no exploratory variable, τ_{00} reflects the between variance of the dependent variable Y_{ij} at Level 2. Substituting Equation (5) into Equation (4) allows to derive the integrated form of the baseline model (Equation 6).

Although the baseline model does not include exploratory variables, it provides helpful and important information because it provides estimates of the variance of the dependent variable at Level 1 (the within variance; σ_e^2) and Level 2 (the between variance; τ_{00}). Thus, the information from the baseline model allows the calculation of the intraclass correlation, extending Equation (2):

$$\rho = \frac{V_B}{V_T} = \frac{V_B}{(V_W + V_B)} = \frac{\tau_{00}}{(\sigma_e^2 + \tau_{00})} \quad (7)$$

The baseline model thereby provides information about the non-independence of observations regarding the dependent variable and hence about the potential bias of ignoring the nested data structure and consequently the importance of adopting a multilevel modeling approach.

Step 2: Adding Independent Variables at Level 1

In step 2 of the model estimation, independent variables at the lower level are added to the model. Equation (8) describes the model at Level 1 with the inclusion of one independent variable, Equations (9) and (10) show the Level 2 model, and Equation (11) presents the integrated model equation:

$$\text{L1 : } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \quad (8)$$

$$\text{L2 : } \beta_{0j} = \gamma_{00} + u_{0j} \quad (9)$$

$$\text{L2 : } \beta_{1j} = \gamma_{10} \quad (10)$$

$$\text{I : } Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + e_{ij} + u_{0j} \quad (11)$$

New in the Level 1 equation is the term $\beta_{1j}X_{ij}$ which characterizes the effect of the Level 1 independent variable X_{ij} on the dependent variable Y_{ij} . The regression coefficient of the Level 1 variable can be substituted by γ_{10} (Equation 10)¹ as shown in the integrated equation (Equation 11). The equation of the random intercept remains unchanged (Equation 10) and can also be substituted in the integrated form. Of course, it is possible in this step to add multiple lower level variables.

Step 3: Adding Independent Variables at Level 2

In step 3, independent variables at the higher level are included in the model. Equation (12) presents the unchanged Level 1 equation, Equation (13) shows the new equation of the random intercept, Equation (14) depicts the unchanged equation of the regression coefficient β_{1j} , and Equation (15) presents the integrated equation of the two-level model.

$$\text{L1 : } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \quad (12)$$

¹Please note that although Equation (10) is labeled as a Level 2 equation where γ_{10} is a fixed effect reflecting the linear effect of the independent variable X_{ij} on the dependent variable Y_{ij} at Level 1 (Raudenbush and Bryk 2002). In step 4 we will allow this regression coefficient to vary between clusters which then results in Equation (18) characterizing a potentially meaningful Level 2 influence.

$$\text{L2} : \beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \quad (13)$$

$$\text{L2} : \beta_{1j} = \gamma_{10} \quad (14)$$

$$\text{I} : Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + e_{ij} + u_{0j} \quad (15)$$

In this model a new independent variable at Level 2, Z_j , has been added. This variable only varies between clusters (j) but not between Level 1 observations (i) as indicated by its subscript. Thus, Level 2 independent variables only possess between variance but no within variance and can therefore only explain the between variance component of the dependent variable.

The influence of independent variables at Level 2 is reflected in the intercept equation (β_{0j} ; Equation 13). Here, γ_{01} reflects the influence of Z_j on the dependent variable Y_{ij} . Again, this equation can be extended to include additional Level 2 variables. Furthermore, Equation (13) together with Equation (14) can be substituted into Equation (12) to derive the integrated form of the model (Equation 15), now additionally including the influence of the Level 2 independent variable.

Step 4: Testing for Random Slopes

In step 4, the random slope model investigates whether there is substantial variance in the regression coefficient of a Level 1 variable across Level 2 observations. Thus, the model evaluates whether the slopes significantly vary between clusters. Equations (16) and (17) reflect the unchanged Level 1 model and the random intercept model, respectively. Equation (18) describes the revised equation of the regression coefficient (β_{1j}) and Equation (19) presents the integrated equation of the two-level random slope model.

$$\text{L1} : Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \quad (16)$$

$$\text{L2} : \beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \quad (17)$$

$$\text{L2} : \beta_{1j} = \gamma_{10} + u_{1j} \quad (18)$$

$$\text{I} : Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + e_{ij} + u_{0j} + u_{1j}X_{ij} \quad (19)$$

New to the model in step 4 is that the regression coefficient β_{1j} is allowed to vary between clusters, which is reflected in adding the residual of the slope u_{1j} to the random slope equation (Equation 18) which results in the term $u_{1j}X_{ij}$ in the integrated form (Equation 19) (Fig. 4b illustrates a model with random slope parameter and Fig. 4c shows a model with random intercept and slope parameters). Analogously to the residual of the random intercept equation, the residual of the random slope equation at Level 2 is assumed to have an expected mean of zero ($E(u_{1j}) = 0$) and a variance τ_{11} ($\text{Var}(u_{1j}) = \tau_{11}$).

One can test whether this variance in the slope is meaningful by directly testing whether τ_{11} is significantly different from zero and/or by comparing the fit of a

model in which the slope is allowed to vary with a model in which the slope is fixed across clusters using a deviance test (see section “[Model Estimation & Assessing Model Fit](#)”). If the variance in the random slope is substantial, it is recommended to allow the slope to vary across clusters. Furthermore, one may then include higher level variables (i.e., variables at Level 2) to explain this variance, which leads to cross-level interaction effects, which we explain in step 5.

If there are multiple random slopes, we recommend testing their significance in a stepwise approach to prevent errors in the specification of the model and potential issues in the model estimation. Furthermore, in models including random intercept and random slope parameters, one should allow the intercept and the slope to covary ($Cov(u_{0j}; u_{1j}) = \tau_{01}$).

Step 5: Adding Cross-Level Interaction Effects

If there is substantial variance in the slope between clusters (or strong theoretical reasons suggest a cross-level interaction effect; cf. Snijders and Bosker 2012), one can continue with step 5 – the estimation of cross-level interaction effects. In step 5, independent variables at the higher level are added to the model to account for the variance of the random slope. Thereby one can explore whether higher level variables explain the variation between clusters of an effect of a lower level variable on the dependent variable (i.e., the variance in the slope between clusters).

Equations (20), (21), (22), and (23) describe this model. Equation (20) and (21) show the unchanged Level 1 equation and the equation of the random intercept. Equation (22) describes the revised equation of the random slope and Equation (23) presents the integrated form of the model if Equations (21) and (22) are substituted into Equation (20).

$$L1 : Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \quad (20)$$

$$L2 : \beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \quad (21)$$

$$L2 : \beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j} \quad (22)$$

$$I : Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + e_{ij} + u_{0j} + u_{1j}X_{ij} \quad (23)$$

New to the random slope equation (Equation 22) is the term $\gamma_{11}Z_j$, which describes the effect (γ_{11}) of the Level 2 variable Z_j on the random slope parameter β_{1j} . The interpretation of this effect becomes more intuitive if we derive the integrated form of the model by substituting Equations (21) and (22) into Equation (20). If $\gamma_{11}Z_j$ from Equation (22) is substituted into Equation (20) it is multiplied by X_{ij} resulting in $\gamma_{11}Z_jX_{ij}$. This expression clarifies that γ_{11} reflects a cross-level interaction effect, which indicates whether the relationship between the independent variable X_{ij} and the dependent variable Y_{ij} varies as a function of the Level 2 variable Z_j and thus determines whether Z_j moderates the relationship between X_{ij} and Y_{ij} across clusters.

Assumptions of Multilevel Modeling

Multilevel regression models share many assumptions of the linear multiple regression model² (Hox et al. 2018; Snijders and Bosker 2012) such as the correct model specification (e.g., with respect to the functional relationship and the absence of omitted variables), perfect reliability of included variables, or the absence of multicollinearity of Level 1 and Level 2 variables (Raudenbush and Bryk 2002; Hox et al. 2018). Additional assumptions of the multilevel regression refer to the residuals at both levels and the complex variance structure. We therefore present the assumptions referring to the residuals of the two-level model presented in step 5 in Equations (24), (25), (26), and (27):

$$e_{ij} \sim iid N(0; \sigma_e^2) \quad (24)$$

$$u_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim iid N(0; \mathbf{T}) \text{ with } \mathbf{T} = \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \sim iid N \quad (25)$$

$$Cov(e_{ij}; u_{0j}) = 0 \quad (26)$$

$$Cov(e_{ij}; u_{1j}) = 0 \quad (27)$$

Specifically, Equation (24) describes that residuals at Level 1 are assumed to be independently and identically normally distributed (iid) with an expected mean of zero and a variance of σ_e^2 . Equation (25) highlights that residuals at Level 2, u_{0j} and u_{1j} , are assumed to be multivariate normal distributed with expected means of 0, variances of τ_{00} and τ_{11} and a covariance of $\tau_{01} = \tau_{10}$. Furthermore, Equations (26) and (27) posit that residuals at Level 1 should be independent from residuals at Level 2. Additionally, the residuals at both levels should be homoscedastic (see Snijders and Bosker 2012 for details, potential relaxations, and guidance on how to test this assumption) and independent from the predictor variables at the respective level (Raudenbush and Bryk 2002). Furthermore, analogously to standard regression analysis, the regressors at both levels should be uncorrelated with the error terms to avoid potential problems associated with endogeneity (please see the chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#) for a general introduction on how to deal with endogeneity in marketing research).

Model Estimation & Assessing Model Fit

Usually multilevel models are estimated by maximum likelihood estimation techniques. Maximum likelihood methods produce estimates for the population

²Note that we focus here on assumptions of multilevel models which are estimated using a maximum likelihood estimator. Other estimation techniques can be helpful if these assumptions are not fulfilled (see section [“Model Estimation & Assessing Model Fit”](#) and Hox et al. 2018).

Graphical Illustration of Random Intercept and Slope Models

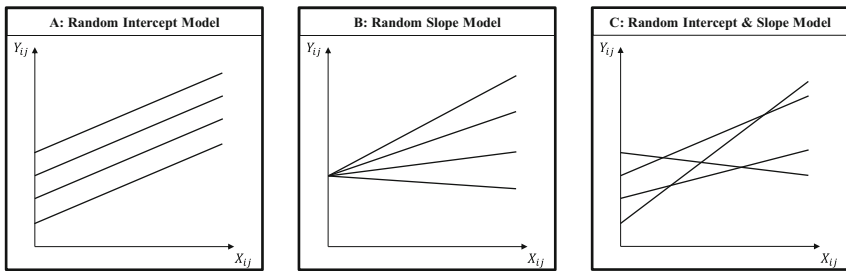


Fig. 5 Graphical illustration of random intercept and slope models

parameters by minimizing the difference between the variance-covariance matrix of the model and the empirical variance-covariance matrix (i.e., the variance-covariance matrix of the observed data; Hox et al. 2018). The estimation process is an iterative process, in which in each step the improvement of the model fit is evaluated by comparing the model fit of the new model with the fit of the previous model. The model fit is assessed by the (Log-) Likelihood. Usually, the differences in the (Log-) Likelihood are relatively large in the beginning and become smaller by every iterative step. If the improvement of model fit becomes very small, the estimation is finished and the parameter estimates of the last step are used to conduct significance tests. Eliason (1993) provides a general overview of maximum likelihood estimation techniques and Heck and Thomas (2015) and Hox et al. (2018) provide overviews of maximum likelihood estimation techniques for multilevel models.

An alternative to employing a maximum likelihood estimator, is the Bayesian estimation of multilevel models (e.g., Hamaker and Klugkist 2011). Bayesian estimation of multilevel models can especially be helpful when dealing with small samples and non-normality (Hamaker and Klugkist 2011). Detailed discussions of Bayesian estimation of multilevel models can for example be found in Depaoli and Clifton (2015), Gelman et al. (2014), Hamaker and Klugkist (2011), and Gelman and Hill (2006).

After estimating the parameter estimates, the significance of regression coefficients can be tested. To test the significance of regression coefficients, regression coefficients are divided by their respective standard error. In most statistical software packages, the resulting test statistic follows the standard normal or t -distribution, depending on the specific software (Hox et al. 2018).

As outlined in the previous section, multilevel models are usually investigated in a stepwise procedure. As every model has an individual model fit, model comparisons assess whether each step improves the fit of the model to the data. Model comparisons in multilevel modeling usually base on the deviance. The deviance value between two models is calculated by the difference between the respective Log-Likelihood values of each model multiplied by -2 .

Generally, models with a lower deviance fit better than models with a higher deviance. Thus, to test whether a specific model which is nested in a more general model shows a better model fit, we can compare its deviances by calculating the difference of the deviance between the more general model (M_0 with a deviance D_{M_0}) and the model of interest (M_1 with a deviance D_{M_1}):

$$D_{Diff} = D_{M_0} - D_{M_1}; df = p_{M_1} - p_{M_0} \quad (28)$$

The difference in the deviances (D_{Diff}) of this test, which is referred to as deviance difference test or likelihood ratio test, follows the chi-square distribution. The degrees of freedom reflect the difference of the number of parameters estimated in the model of interest (p_{M_1}) and the more general model (p_{M_0}). The deviance difference test is especially helpful if the stepwise procedure of multilevel modeling is employed to assess the improvement in model fit from one step to the next.

However, as mentioned before, the deviance test can be only applied to compare nested models. If the models that should be compared are not nested, relative model comparisons can be conducted with the help of other criteria, such as the Akaike information criterion (Akaike 1973) or the Schwarz information criterion (Schwarz 1978).

Furthermore, it is often of great interest to know how much variation in the dependent variable is explained by the independent variables. In multiple regression analysis, the squared multiple correlation coefficient R^2 measures the explained proportion of variance in the dependent variables. This logic can also be transferred to multilevel regression analysis. However, as the variance structure in multilevel models is not limited to one level of analysis, the evaluation of the explained proportion of variance in the dependent variable becomes more complex and needs to be assessed on multiple levels of analysis. For example, in a two-level model at least two coefficients of determination have to be calculated. One coefficient of determination for the lower level of analysis and one for the higher level of analysis.

In the following, we provide an application-oriented description of the approach of Snijders and Bosker (1993), to provide a suitable multilevel version of R^2 . Snijders and Bosker (1993) treat the proportional decreases in the estimated variance components in the baseline model as analogs of R^2 -values. Consequently, to calculate multilevel R^2 -values the variance component of the baseline model is compared to the variance component of the comparison model (Hox et al. 2018; Raudenbush and Bryk 2002). The following equation can be used to calculate the R^2 -value for a dependent variable at the lower level, where b denotes the baseline and m the comparison model:

$$R_{L1}^2 = \frac{\sigma_b^2 - \sigma_m^2}{\sigma_b^2} \quad (29)$$

Further, to assess the R^2 -value for the random intercept, the following equation can be used, where, again, b reflects the baseline model and m the comparison model:

$$R_{L2,int}^2 = \frac{\tau_{00|b}^2 - \tau_{00|m}^2}{\tau_{00|b}^2} \quad (30)$$

Finally, the following equation shows how the R^2 -value for the random slope can be calculated by comparing the baseline model b with the comparison model m :

$$R_{L2,slope}^2 = \frac{\tau_{11|b}^2 - \tau_{11|m}^2}{\tau_{11|b}^2} \quad (31)$$

Variable Centering

The use of centering independent variables to establish a zero point on scales that otherwise lack such a value or to investigate interaction effects is relatively common in ordinary least squares regression (Aiken et al. 1991; Enders and Tofghi 2007). Most importantly, independent variables are centered to ensure that the intercept of the regression model is interpretable as the expected value of the dependent variable, if all independent variables have their mean value. As multiple regression models are invariant under linear transformations, the transformation of variables changes the estimated parameters in a similar way. Consequently, it is always possible to recalculate the untransformed estimates (Hox et al. 2018). Due to the hierarchical structure of multilevel data, multilevel models are only invariant for linear transformations when the model does not include random slopes, which vary at the higher level. Thus, centering becomes more complex when investigating multilevel models.

Given the nested data structure in multilevel models with lower level observations nested within higher level observations, we can distinguish different forms of centering. The traditionally most prevalent two approaches are grand mean centering and group mean centering (for other centering approaches, such as latent mean centering, see Asparouhov and Muthén (2019)). When grand mean centering is applied, the grand mean value of a variable is subtracted from all observations of that variable in the dataset (i.e., $X_{ij} - \bar{X}$ or $Z_j - \bar{Z}$). When group mean centering is applied, one subtracts the group mean of a cluster j from all observations i of that respective cluster (i.e., $X_{ij} - \bar{X}_j$; with \bar{X}_j describing the group mean of cluster j). Consequently, for variables at the higher level only grand mean centering can be applied, whereas lower level variables can either be grand mean and group mean centered. As both techniques produce parameter estimates that can differ in their value and their meaning and can create differences in deviance values, the centering of variables in multilevel modeling has been discussed vibrantly in the methodological literature (e.g., Enders and Tofghi 2007; Hofmann and Gavin 1998; Kreft 1996; Kreft et al. 1995; Longford 1989; Paccagnella 2006; Raudenbush 1989; Wu and Wooldridge 2005). In this section, we follow the recommendations of Enders and Tofghi (2007) who focus on two-level cross-sectional data. Further recommendations on the centering of longitudinal data can be found for example in Biesanz et al. (2004) and Asparouhov and Muthén (2019).

In multilevel models, the independent variables of the higher level are usually centered on their grand mean. The decision whether lower level variables should be centered on their grand mean or on their group mean is more complex. According to Enders and Tofighi (2007) centering independent variables on the grand mean is appropriate if one primarily focuses on higher level effects and includes independent variables at the lower level only as control variables and when interaction effects between higher level variables are investigated. Group mean centering is appropriate if the lower level effect of the independent variable on the dependent variable is of substantial interest and when examining cross-level interaction effects and interaction effects that include lower level variables (Enders and Tofighi 2007). Both approaches can be applied if the focus is on the analysis of the differential effects of a variable at the lower level and the higher level. However, Enders and Tofighi (2007) highlight that these are only recommendations as the decision whether to perform grand mean or group mean centering cannot be solely based on statistical evidence. Therefore, it always depends on the individual research question whether grand mean or group mean centering is the appropriate method.

In addition, the appropriate centering of independent variables is important for the estimation of the multilevel model. Independent variables that are centered appropriately will increase the speed of the estimation and will lower the likelihood of convergence problems (Hox et al. 2018). Thus, especially, if independent variables have a high variation in their means and variances, the appropriate centering is important to ensure the convergence of the model estimation process.

Sample Size Considerations

In multilevel modeling decisions about adequate sample sizes are somewhat more complex than in conventional single level analysis. As multilevel models comprise observations on multiple levels of analysis, also decisions about adequate sample sizes refer to multiple levels. Therefore, questions arise about the minimum level-specific sample size to estimate unbiased parameters and standard errors and the potential biases caused by samples that are too small. Previous simulation studies provide some answers to these questions for two-level models.

For example, Maas and Hox (2005) investigated how the number of level two observations (*here*: number of groups; $n_j = 30, 50, 100$), the number of level one observations nested within each level two unit (*here*: group sizes; $n_{ij} = 5, 30, 50$), and the intraclass correlation ($\rho = 0.1, 0.2, 0.3$) influence the parameter estimates and standard errors in a simple two-level regression model with one predictor at each level and thus one direct effect at Level 1 and Level 2 and a cross-level interaction effect (i.e., $Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}$). Results on the basis of 27,000 simulated data sets (1,000 for each simulation condition) show that all parameter estimates (i.e., intercept, regression coefficients, and variance components) in each condition are largely unbiased (with an average parameter bias $<0.05\%$). Furthermore, the results show that also the standard errors of the intercept and regression coefficients are estimated accurately under each condition. However,

standard errors for level two variance estimates are substantially underestimated if the number of level two observations is too low. Specifically, standard errors of level two variance parameters are estimated about 15% too small with a number of 30 level two observations. With 50 observations at level two estimates are more acceptable and most accurate with 100 observations at level two. The intraclass correlation of the dependent variable had neither a substantial effect on the accuracy of the parameter estimates nor on the accuracy of the standard errors. Overall, Maas and Hox's (2005) results are in line with other simulation studies showing that a larger number of level two observations are needed to accurately estimate level two variance components (Hox et al. 2018).

In addition, Hox et al. (2018) formulate three rules of thumb for sufficient sample sizes in two-level regression models employing standard estimation techniques. These rules propose specific sample sizes depending on the part of the model the researcher is interested in. In line with the recommendation by Kreft (1996; see also Kreft and De Leeuw 1998), Hox et al. (2018) propose a 30/30 rule with 30 level two observations and 30 level one observations per each level two unit if one is only interested in the fixed part of the model (i.e., the direct effects of the level one and level two predictor variables). If one is interested in cross-level interactions, a 50/20 rule with 50 level two observations and 20 level one observations per level two unit is suggested. For researchers who are especially interested in (co-)variance and standard errors at level two, Hox et al. (2018) propose a 100/10 rule with 100 level two observations and 10 corresponding level one observations.

Whereas these rules of thumb offer sound advice, in practice, it may be difficult for researchers to meet the recommended sample sizes. In this respect, the discussed simulation study provides additional insights by showing that researchers may yield accurate regression coefficients and standard errors with smaller samples (Maas and Hox 2005). As mentioned in the section "[Model Estimation & Assessing Model Fit](#)," another potential way for researchers to cope with small samples may be the use of a Bayesian estimation technique, which does not rely on asymptotic results (Hamaker and Klugkist 2011).

Although a sample which is large enough to yield accurate results is a prerequisite for conducting any multilevel study, it does not guarantee that it is large enough to detect existing effects in the population as significant. This uncertainty leads to the question of sufficient sample sizes to assure a high statistical power. The statistical power of significance test refers to the probability to detect an existing effect in the population as significant by rejecting the null hypothesis (e.g., Cohen 1992). The mistake to not reject the null hypothesis in the presence of an effect in the population is known as type II error and occurs with the probability β . Therefore, statistical power is defined as $(1 - \beta)$ – the probability to reject a false null hypothesis (Cohen 1992). Cohen (1992) proposes a desired level of 0.80 to assure high statistical power of a significance test.

The decision about adequate sample sizes to achieve high statistical power in multilevel investigations a priori (i.e., before the data is collected) depends on the focus of the model and several additional assumptions about the data and different parameter estimates (e.g., Hox et al. 2018; Snijders 2005). Thus, general

recommendations about sufficient sample sizes which assure a high statistical power would be difficult (e.g., Snijders and Bosker 2012). However, software programs may help to determine adequate sample sizes based on the researcher's assumptions. For example, PinT (Power in Two-level designs) developed by Bosker, Snijders, and Guldemond (2003; see also Snijders and Bosker 1993) helps to decide about adequate sample sizes by providing approximate standard errors of regression coefficients for different combinations of level one and level two observations in two-level models. Furthermore, Mathieu et al. (2012) developed a multilevel Monte Carlo power tool, executable in R, which helps researchers to a priori estimate the power of their cross-level interactions. Together these resources can be very helpful in determining model specific samples sizes to assure a high statistical power when collecting multilevel data.

Multilevel Structural Equation Modeling

In explaining the fundamentals of multilevel analysis we so far focused on the basic multilevel regression model. In the following we extend this approach by integrating the logic of structural equation modeling techniques to multilevel analysis (please refer to Baumgartner and Weijters (2021) in this handbook for a general introduction to structural equation modeling and to Hughes and Ahearne (2010) or Hunter and Panagopoulos (2015) for examples of multilevel structural equation modeling in marketing research). This integration of structural equation modeling techniques to multilevel analysis, referred to as multilevel structural equation modeling, allows the accurate modeling of latent variables, the consideration of measurement error, and the simultaneous estimation of more complex relationships, such as the effects on multiple dependent variables or the modeling of causal chains (e.g., Heck and Thomas 2015; Hox et al. 2018). Thereby, multilevel structural equation modeling allows a more accurate assessment of relationships between variables and the investigation of more complex relationships in multilevel settings (Heck and Thomas 2015).

Single level structural equation modeling combines the advantages of factor analysis, defining latent variables by their observed indicators, and path analysis, allowing the investigation of complex causal relationships (Bollen 1989; Matsueda 2014; Muthén 2002). Corresponding with both types of advantages, the definition of latent variables by measured indicators is captured in the measurement model whereas the structural relationships are captured in the structural part of the model. The basic measurement model can be represented as:

$$y_i = v + \Lambda\eta_i + \varepsilon_i \quad (32)$$

where for observation i y_i is a $p \times 1$ vector of measured variables, v is a $p \times 1$ vector of intercepts terms, Λ is a $p \times m$ matrix of factor loadings, η_i is a $m \times 1$ vector of

latent variables, and ε_i is a $p \times 1$ vector of measurement errors. The basic structural model can be written as:

$$\eta_i = \alpha + B\eta_i + \zeta_i \tag{33}$$

where η_i is a $m \times 1$ vector of latent variables, α is a $m \times 1$ vector of intercept terms, B is a $m \times m$ matrix of structural regression coefficients among the latent variables, and ζ_i is a $m \times 1$ vector of latent variable regression residuals. Residuals ε_i from the measurement model and residuals ζ_i from the structural model are assumed to be multivariate normal with means of zero and variance/covariance matrices Θ and Ψ . Together the measurement and structural model imply a mean (μ) and covariance (Σ) structure which are employed to estimate the parameters using for example a maximum likelihood estimator.

Structural equation modeling techniques have only recently been more widely applied to multilevel data structures (Heck and Thomas 2015). In accommodating multilevel data structures, the model set up for a (multilevel) structural equation model is more complex as some parameters of the measurement and structural model are allowed to vary at the higher level (Heck and Thomas 2015; Preacher et al. 2010). In the case of two-level contexts, a multilevel structural equation model can therefore be characterized by measurement and structural models for both the lower and the higher level (Heck and Thomas 2015). The level one (within (W)) measurement and structural model can be expressed as:

$$y_{ij} = \Lambda_W \eta_{Wij} + \varepsilon_{Wij} \tag{34}$$

$$\eta_{Wij} = B_W \eta_{Wij} + \zeta_{Wij} \tag{35}$$

where, for observation i nested in level two unit j , y_{ij} is a $p \times 1$ vector of measured variables, Λ_W is a $p \times m$ matrix of level one factor loadings, η_{Wij} is a $m \times 1$ vector of latent variables, and ε_{Wij} is a $p \times 1$ vector of measurement errors at level one with means of zero and a variance/covariance matrix Θ_w (Asparouhov and Muthén 2007; Heck and Thomas 2015). In the structural part of the level one model, B_{Wij} is a $m \times m$ matrix of structural regression coefficients among the latent variables at level one and ζ_{Wij} is $m \times 1$ vector of latent variable regression residuals at level one with zero means and a variance/covariance matrix Ψ_w (Asparouhov and Muthén 2007; Heck and Thomas 2015).

The measurement and structural model at level two (between (B)) can be written as:

$$y_j = v_j + \Lambda_B \eta_{Bj} + \varepsilon_{Bj} \tag{36}$$

$$\eta_{Bj} = \alpha_j + B_B \eta_{Bj} + \zeta_{Bj} \tag{37}$$

Here, y_j is a $p \times 1$ vector of measured variables at level two, v_j is a $p \times 1$ vector of intercept terms of the measured variables, Λ_B is a $p \times m$ matrix of level two factor loadings, η_{Bj} is a $m \times 1$ vector of latent variables, and ε_{Bj} is a $p \times 1$ vector of measurement errors at level two with means of zero and a variance/covariance matrix Θ_B (Asparouhov and Muthén 2007; Heck and Thomas 2015). In the structural part of the level one model, B_{Bj} is a $m \times m$ matrix of structural regression coefficients among the latent variables at level two and ζ_{Bj} is $m \times 1$ vector of latent variable regression residuals at level two with zero means and a variance/covariance matrix Ψ_B (Asparouhov and Muthén 2007; Heck and Thomas 2015). As in the single level structural equation model, the measurement and structural models imply a mean (μ) and covariance (Σ_T) structure. However, in two-level structural equation models the covariance structure (Σ_T) can be decomposed in a level one (within) and a level two (between) component which are orthogonal and additive (Heck and Thomas 2015):

$$\Sigma_T = \Sigma_W + \Sigma_B \quad (38)$$

Multilevel structural equation models can be estimated employing a full information maximum likelihood estimator which allows the accommodation of missing data, unbalanced cluster sizes, and importantly random slopes (Hox et al. 2018; Preacher et al. 2010; Mehta and Neale 2005).³

The global fit of multilevel structural equation models can be evaluated assessing standard fit indices, such as the chi-square statistic, the comparative fit index (CFI; Bentler 1990), the Tucker-Lewis Index (TLI; Tucker and Lewis 1973), the root mean square error of approximation (RMSEA; Browne and Cudeck 1992; Steiger and Lind 1980), or the standardized root mean residual (SRMR; Bentler 1995). However, those indices which are based on the chi-square statistic apply to the entire model and thus comprise information about the model fit of both the level one and the level two model. Furthermore, as the sample size at level one is generally considerably larger than the sample size at level two, these global model fit indices are often dominated by model fit at level one. Given the confounding information in the standard global model fit indices, Hox et al. (2018) propose to assess the model fit separately for each level of analysis (see also Ryu and West 2009). To evaluate the fit indices for level one, one can estimate

³Sometimes a slightly different notation for multilevel structural equation models is employed (Asparouhov and Muthén 2008; Preacher et al. 2010, 2011). Following this notation, the measurement model can be expressed as: $Y_{ij} = v_j + \Lambda_j \eta_{ij} + K_j X_{ij} + \varepsilon_{ij}$. The level one structural model can be written as $\eta_{ij} = \alpha_j + B_j \eta_{ij} + \Gamma_j X_{ij} + \zeta_{ij}$ and the level two structural model can be expressed as $\eta_j = \mu + \beta \eta_j + \gamma X_j + \zeta_j$. This notation additionally includes exogenous covariates captured by the vectors X_{ij} and X_j respectively. Furthermore, elements of the matrices v_j , Λ_j , K_j , α_j , B_j , and Γ_j may vary between level two units as expressed by the level two subscripts (j) (for further details of this notation see Preacher et al. 2010).

an independence model (as a baseline model for the comparative fit indices) and the hypothesized model at level one with a saturated model at level two and then calculate the respective fit indices. Analogously, one can estimate an independence model and the hypothesized model at level two with a saturated model at level one to assess the fit indices for the level two model (Hox et al. 2018). In addition to the standard fit indices from structural equation modeling, the fit of nested models can be compared by employing the likelihood ratio test (see also “[Model Estimation & Assessing Model Fit](#)”; Mehta and Neale 2005). Furthermore, as in multilevel regression models, information-theoretic criteria such as Akaike’s Information Criterion (AIC, Akaike 1973, 1987) or the Bayesian Information Criterion (BIC, Schwarz 1978) can be employed to evaluate non-nested models (Mehta and Neale 2005).

With respect to the implementation of multilevel structural equation models some issues can be relevant. A first issue refers to the proportion of level two variance of latent level one variables. If one would like to examine the proportion of level two variance of a latent level one variable relative to its total variance (i.e., the counterpart of the intraclass correlation coefficient in multilevel regression analysis), the factor loadings have to be constrained to be invariant across both levels in order to make the common variance attributed to the latent factor directly comparable (Heck and Thomas 2015). In case of invariant factor loadings, the proportion of the between variance of a latent level one variable relative to its total variance can be expressed as:

$$\psi_B / (\psi_B + \psi_W) \quad (39)$$

where ψ_B refers to the proportion of the factor variance at level two (between) and ψ_W refers to the proportion the factor variance at level one (within).

A second relevant implementation issue refers to the centering of manifest variables in multilevel structural equation models. When a multilevel structural equation model is employed which implicitly partitions each measured level one variable into a latent level one (within) and level two (between) component, no explicit centering of observed predictor variables is required (Preacher et al. 2010). However, group mean centering of level one variables may be helpful for model convergence if the level two variance of a level one variable is essentially zero. Group mean centering of level one variables should be avoided, if the level two effects are of theoretical interest (Preacher et al. 2010).

A third issue when implementing a multilevel structural equation model refers to the residual variances of observed level one variables at level two. Residual variances at level two are often very small, reflecting a high reliability (Heck and Thomas 2015). In such occasions it can be necessary to fix these very small level two variances to zero in order to avoid estimation problems (Heck and Thomas 2015; Muthén and Muthén 1998–2017).

Overall, multilevel structural equation models offer several advantages such as the accurate modeling of latent variables, the consideration of measurement error and the simultaneous estimation of more complex relationships (which is especially

helpful to estimate less biased indirect effects in mediation models; Preacher et al. 2010). However, these advantages come at the cost of a higher model complexity which can make it challenging to generate a converging solution. In such cases it can be helpful to start with a smaller, less complex model and successively add additional variables or relationships. Nevertheless, it sometimes can be difficult to identify the exact source of the estimation problem, so that Heck and Thomas (2015, p. 179) wisely advise: “[...] that patience is virtue when working toward a solution.”

Software for Estimating Multilevel Models

In the last decades the number of multilevel studies in organizational and marketing research has increased substantially (Wieseke et al. 2008). This increase may in part also be traced back to the growing availability of software for estimating multilevel statistical models. Given the high number of different software packages that allow multilevel analyses, it is well beyond of the scope of this chapter to review each of these software packages. We therefore focus our brief review on some of the most widespread software packages for multilevel analysis⁴. In doing so we distinguish between software which has been specifically designed for multilevel analysis and general purpose software that allows the estimation of multilevel models.

One of the most widespread software packages that has been designed to conduct multilevel analysis is HLM (Raudenbush et al. 2019a). HLM, currently available in its eighth version, may be especially well suited for beginners due to its user friendly graphical display, which allows to specify models on a step-by-step basis. Furthermore, HLM is accompanied by a freely available, comprehensive manual (Raudenbush et al. 2019b). The theoretical background of many applications can additionally be found in the textbook by Raudenbush and Bryk (2002). In its current version, HLM allows multilevel analyses up to four levels, can estimate different types of models (i.e., univariate, multivariate, and cross-classified), allows different distributional properties of the outcome variables (e.g., normal, Bernoulli, Poisson binomial, multinomial, ordinal, and over-dispersion), and offers different estimation methods (e.g., REML, FML, PQL, AGH, and higher-order Laplace approximations to maximum-likelihood⁵) (Palardy 2011). This variety of modeling options makes HLM not only attractive for beginners but also for those who want to estimate more advanced multilevel models.

⁴More detailed reviews of many different software packages that allow the estimation of multilevel models can be found at the homepage of the Centre for Multilevel Modelling at the University of Bristol (www.bristol.ac.uk/cmm/learning/mmssoftware/)

⁵REML = Restricted maximum likelihood; FML = Full maximum likelihood; PQL = penalized quasi-likelihood; AGH = Adaptive Gauss-Hermite quadrature.

Another important program which was explicitly designed to estimate multilevel models is MLwiN (Rasbash et al. 2016a). MLwiN has been suggested to be “the most extensive multilevel package” (Snijders and Bosker 2012, p. 325), but has also been considered as less user-friendly than HLM “in the sense that it is not as directive and requires greater user knowledge” (Wieseke et al. 2008, p. 333). However, MLwiN comes with a wealth of information, including an extensive manual (Rasbash et al. 2016b), which can be obtained from the homepage of the Centre for Multilevel Modelling at the University of Bristol (www.bristol.ac.uk/cmm/). MLwiN may fit models up to five levels, has a great modeling flexibility, and allows a wide range of possible models (Snijders and Bosker 2012; Wieseke et al. 2008), including the estimation of multiple membership models, Bayesian analysis of multilevel models, or bootstrapping of standard errors in multilevel models.

The most widespread general purpose statistical software in the social sciences may be IBM SPSS. To estimate multilevel models in SPSS, one can use the routine MIXED which allows the estimation of models with up to three levels (Snijders and Bosker 2012). Furthermore, since Version 19 SPSS also allows the estimation of multilevel models with categorical outcomes (using the GENLIN MIXED routine) (Heck et al. 2014). However, multilevel modeling in SPSS has some limitations regarding the modeling flexibility (Heck et al. 2014). Detailed introductions for the estimation of multilevel models with continuous and categorical outcomes in SPSS are provided by Heck et al. (2012, 2014).

Another general-purpose software which offers a high modeling flexibility for the estimation of multilevel models is Mplus (Muthén and Muthén 1998–2017). Mplus is accompanied by an extensive user guide with detailed examples which help users to learn the Mplus command language. Furthermore, extensive resources like videos of short courses, a helpful discussion forum, and further examples can be obtained from the Mplus homepage (www.statmodel.com). Mplus currently allows the estimation of multilevel models up to three levels (four levels for longitudinal models in which time is the lowest level such as in a three-level latent growth model) and offers a variety of models, estimators, and algorithms. Based on the general latent variable framework, Mplus is especially well suited for the estimation of multilevel structural equation models (Hox et al. 2018; Muthén and Asparouhov 2011). Heck and Thomas (2015) offer a detailed introduction to multilevel modeling techniques employing Mplus.

Multilevel models can also be estimated employing the open source environment of R. The freely available program (www.r-project.org) is potentially the most flexible modeling environment for statistical computing. Multilevel models can be estimated employing different packages, such as nlme (Pinheiro et al. 2016) or lme4 (Bates et al. 2016). An introduction to multilevel modeling using nlme and lme4 is provided by Bliese (2016) and Finch et al. (2014). Multilevel models can also be estimated using other general purpose software like Stata or SAS. An extensive and detailed introduction for the estimation of multilevel models in Stata is provided by Rabe-Hesketh and Skrondal (2012). Introductions for estimating multilevel models in SAS are offered for example by Singer (1998), Bliese (2002), or Albright and Marinova (2010).

Conceptual Framework of the Example

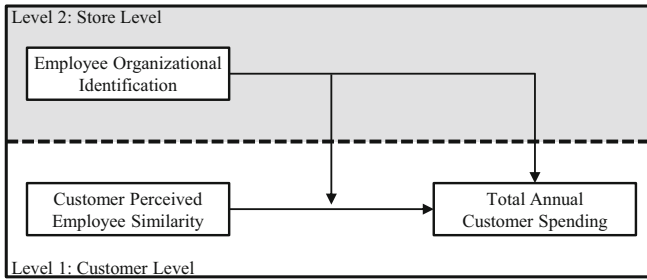


Fig. 6 Conceptual framework of the example. (Notes: Adapted from Netemeyer et al. 2012)

Example: Building and Estimating a Two-Level Model

In order to offer an example for building and estimating a multilevel model, we draw from the study of Netemeyer, Heilman, and Maxham (2012). In their study the authors examine the effects of customer perceived similarity to store employees and store employees' organizational identification (aggregated at the store level) on customer identification and customer spending in the context of a private label women's apparel retailer. Furthermore, they explored the cross-level interaction between employees' organizational identification and customer perceived employee similarity on both outcomes. For our example, we adopt the framework from Netemeyer et al. (2012) and focus on the direct and interactive effects of customer perceived employee similarity and employees' organizational identification on customer spending. Figure 6 presents the conceptual framework of the example.

To offer an example for the analysis strategy outlined in the section “[Process of Multilevel Modeling: The Two-Level Regression Model](#),” we created a fictitious data set comprising 10,000 observations at level one (i.e., customers) nested within 500 observations at level two (i.e., stores) to build and test the model presented in Fig. 4. Data for customer perceived similarity and store employees' organizational identification is created to reflect responses on seven-point scales. Data for total annual customer spending is created to reflect the total dollar amount of customer spending within one year⁶ (for the original measures see Netemeyer et al. 2012). In

⁶Note that we divided the total dollar amount of customer spending within 1 year by 100 to keep the (residual) variance estimates at a lower level, which is helpful to assure a smooth model estimation process.

line with the recommendations in the section “[Variable Centering](#),” we group mean centered customer perceived similarity and grand mean centered employees’ organizational identification. We estimated all models using Mplus 8.5 and employed a maximum likelihood estimator with robust standard errors (Muthén and Muthén 1998–2017). The data set and Mplus input and output files can be downloaded from the companion homepage of the handbook. In addition, we also provide the R code employing the lme4 package (Bates et al. 2016) on the companion homepage of the handbook.

As outlined in the section “[Process of Multilevel Modeling: The Two-Level Regression Model](#)” the starting point for testing a multilevel model is to estimate an intercept-only model without any predictor variables. Results of the intercept-only model presented in Table 2 (Model 1) show an estimated mean customer spending amount of 17.628 (reflecting 1,762.80\$). Furthermore, results of Model 1 are informative about the level one (σ_e^2) and level two variance (τ_{00}) of the dependent variable (i.e., customer spending). With the estimates of both variance components one may determine the intraclass correlation (ρ), by dividing the level two variance by the total variance (i.e., $(\frac{\tau_{00}}{\tau_{00} + \sigma_e^2})$). Results from Model 1 imply an intraclass correlation of 0.3349 meaning that 33.49% of the variance of customer spending can be explained at level two (i.e., the store level).

In a second step, one can now add the level one predictor variables. In our example, the level one predictor variable which is added to the model is customers’ perceived similarity with the store employees. Results from Model 2 in Table 2 show that customers’ perceived employee similarity has a positive significant effect on customer spending ($\gamma_{10} = 1.455, p < 0.01$). Furthermore, the results show that the level one variance decreases substantially from Model 1 to Model 2 (Model 1: $\sigma_e^2 = 12.259$, Model 2: $\sigma_e^2 = 8.955$), which reflects the variance explained by adding the level one predictor (here: customer perceived employee similarity; $R_{L1}^2 = \frac{\sigma_b^2 - \sigma_m^2}{\sigma_b^2} = \frac{12.259 - 8.955}{12.259} = .2695$). This finding is also underlined by a significant increase of model fit as suggested by the Satorra-Bentler-corrected log-likelihood difference test ($\chi^2 = 1252.4561, p < 0.01$) (Satorra and Bentler 1999).

In a third step, the level two predictors may be added to the model. Results from Model 3 in Table 2 show that the level two predictor variable, employee organizational identification, has a positive significant effect on customer spending ($\gamma_{01} = 1.538, p < 0.01$). Furthermore, the results show a substantial decrease in the estimate of the level two intercept variance component from Model 1 to Model 3 (Model 1: $\tau_{00} = 6.337$, Model 3: $\tau_{00} = 2.619$), which reflects the variance explained by adding the level two predictor (here: employee organizational identification; $R_{L2,int}^2 = \frac{\tau_{00|b} - \tau_{00|m}}{\tau_{00|b}} = \frac{6.337 - 2.619}{6.337} = .5867$). In addition, when comparing Model 3 and Model 2, we again find a significant increase in model fit ($\chi^2 = 410.0310, p < 0.01$).

In a fourth step one can now test whether the slope between customer perceived similarity and customer spending varies across level two units (here: stores), which would be necessary for a level two predictor variable to significantly explain the

**Cross-Level Interaction Plot:
Customer Perceived Employee Similarity x Employee Organizational Identification**



Fig. 7 Cross-level interaction plot: Customer Perceived Employee Similarity × Employee Organizational Identification. (Notes: Low/high levels of the interaction plot refer to one standard deviation below/above the mean)

variance in this slope and thus for a substantial cross-level interaction. Results from Model 4 in Table 2 show that the estimate of the variance component of the slope is significant ($\tau_{11} = 0.712, p < 0.01$) and thus fulfills the necessary condition for a potential cross-level interaction.

In a fifth and final step, one can now try to explain (some of) the variance in the random slope by a level two predictor variable (here: employee organizational identification). Results of Model 5 in Table 2 show that employee organizational identification explains variance in the random slope of customer perceived employee similarity on customer spending as reflected in a significant cross-level interaction effect between customer perceived employee similarity and employee organizational identification on customer spending ($\gamma_{11} = 0.483, p < 0.01$). Furthermore, results show a substantial decrease in the estimate of the random slope variance component (Model 4: $\tau_{11} = 0.712$, Model 5: $\tau_{11} = 0.344$), which reflects the variance explained in the random slope by the level two predictor (here: employee organizational identification; $R^2_{L2,slope} = \frac{\tau^2_{11|b} - \tau^2_{11|m}}{\tau^2_{11|b}} = \frac{.712 - .344}{.712} = .5169$). Furthermore, explaining the variance in the random slope leads to a significantly better model fit ($\chi^2 = 275.8173, p < 0.01$) which underlines the substantiveness of the cross-level interaction effect.

Plotting the cross-level interaction effect can yield further insights into the interplay between the level one and level two predictor variable in influencing the

Table 2 Results of Multilevel Analyses

	Model 1	Model 2	Model 3	Model 4	Model 5
	Intercept-Only Model	Level 1 Model	Level 2 Model	Test of Slope Variance	Full Multilevel Model
	L1: $Y_{ij} = \beta_{0j} + e_{ij}$ L2: $\beta_{0j} = \gamma_{00} + u_{0j}$	L1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ $\beta_{1j} = \gamma_{10}$	L1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ $\beta_{1j} = \gamma_{10}$	L1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{1j}$	L1: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ L2: $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$
Parameter	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
Intercept (γ_{00})	17.628** (0.116)	17.628** (0.116)	17.496** (0.078)	17.628** (0.080)	17.628** (0.078)
Similarity (γ_{10})	–	1.455** (0.044)	1.455** (0.044)	1.467** (0.045)	1.461** (0.036)
Employee OI (γ_{01})	–	–	1.538** (0.073)	1.215** (0.080)	1.538** (0.073)
Similarity x Employee OI ($\gamma_{11} X_{ij} Z_j$)	–	–	–	–	0.483** (0.046)
Variance Components					
σ_e^2	12.259** (0.497)	8.955** (0.345)	8.955** (0.345)	7.897** (0.195)	7.902** (0.196)
τ_{00}	6.173** (0.528)	6.337** (0.531)	2.619** (0.208)	2.836** (0.250)	2.671** (0.210)
τ_{11}	–	–	–	0.712** (0.127)	0.344** (0.060)
Model Fit					
$-2 \times \text{Log-Likelihood}$	54643.618	51659.516	51262.442	50657.356	50427.876
R^2_{L1}	–	0.2695	0.2695	0.3558	0.3554
$R^2_{L2,int}$	–	–	0.5867	0.5406	0.5673
$R^2_{L2,slope}$	–	–	–	–	0.5169
Δ Model fit ^a (df)	–	1252.4561** (1)	410.0310** (1)	501.8545** (2)	275.8173** (1)

Notes: Dependent variable: Customer Spending (in hundred USD)

* $p < 0.05$, ** $p < 0.01$, Est. = Estimate, SE = Standard Error, L1 = Level 1, L2 = Level 2

^aWe employed the correction originally proposed by Satorra and Bentler (1999) for the model comparisons between Model 1–2, 2–3, and 4–5 and the strictly positive correction (Satorra and Bentler 2010; Asparouhov and Muthén 2013) for the model comparisons between model 3–4.

Details can also be found at the Mplus homepage (statmodel.com). Model 1 = baseline model for R^2_{L1} and $R^2_{L2,int}$ calculations. Model 4 = baseline model for $R^2_{L2,slope}$ calculation

dependent variable. Figure 7 presents the interaction plot for the cross-level interaction effect between customer perceived employee similarity and employee organizational identification on customer spending. Specifically, Figure 7 further supports the results presented in Model 5 in Table 2, by indicating that the effect of customer perceived employee similarity on customer spending is stronger if employee organizational identification is high rather than if it is low (Table 2).

Conclusions

In the last decades the number of multilevel studies in marketing and management has substantially increased (Wieseke et al. 2008), reflecting a growing interest in the investigation of complex phenomena traversing different levels of analysis. With this chapter we wanted to provide an applied introduction how research pertaining to multiple levels of analysis can be conducted by employing multilevel modeling techniques. Therefore, we provided insights about the fundamentals of multilevel modeling discussing the conceptual and statistical relevance of multilevel modeling. Furthermore, we offered a step-by-step analysis strategy how to build and estimate multilevel models. We offered insights how to evaluate the goodness of fit in multilevel models and shed light on some important issues for the implementation of multilevel models, such as different approaches to the centering of predictor variables and recommendations for sufficient sample sizes. Moreover, we provided insights to more advanced multilevel modeling techniques, such as multilevel structural equation modeling, and offered a brief overview of different software packages that allow the estimation of multilevel models. Finally, we offered a detailed example of building and estimating a multilevel model. Overall, we hope that this chapter may be helpful for those who want to start adopting a multilevel lens to capture more of the complexity inherent to many phenomena in marketing and management research.

Cross-References

- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Panel Data Analysis: A Non-Technical Introduction for Marketing Researchers](#)
- ▶ [Regression Analysis](#)
- ▶ [Structural Equation Modeling](#)

Appendix

Appendix Key Terms and Definitions

Term	Definition
Nested/ hierarchical data structures	Data structures in which lower entities are nested within higher entities (e.g., customers nested within salespersons, employees nested within managers, or customers nested within countries) (Heck and Thomas 2015).
Intraclass correlation coefficient [also Intraclass correlation coefficient (1) – ICC(1)]	Measure of heterogeneity of a lower level variable between higher level entities. The intraclass correlation coefficient reflects the proportion of between to total variance of a lower level variable. $\rho = \frac{V_B}{V_T} = \frac{V_B}{(V_W + V_B)}$ where V_T reflects the total variance, V_B the between variance, and V_W the within variance of a lower level variable (Hox et al. 2018; Snijders and Bosker 2012).
Intraclass correlation coefficient (2) – ICC(2)	Measure of group mean reliability of a lower level variable across higher level entities. The ICC(2) is often expressed in terms of ICC(1) with $ICC(2) = \frac{k\rho}{1+(k-1)\rho}$ where k reflects the cluster size (Bliese 1998, 2000; Snijders and Bosker 2012). Recommended range for aggregation: $ICC(2) \geq .70 - .85$ (LeBreton and Senter 2008).
r_{WG}	Index to reflect the interrater agreement for a group regarding a single variable. $r_{WG} = 1 - \frac{S_X^2}{\sigma_E^2}$ where S_X^2 reflects the observed variance on the variable X and σ_E^2 reflects the expected variance of X if there is a complete lack of agreement among raters (LeBreton et al. 2005; LeBreton and Senter 2008). Recommended threshold for aggregation: $r_{WG} \geq .70$ (LeBreton and Senter 2008).
$r_{WG(J)}$	Index to reflect the interrater agreement for a group regarding multiple (J) essentially parallel items. $r_{WG(J)} = \frac{J(1 - \bar{S}_{X_j}^2 / \sigma_e^2)}{J(1 - \bar{S}_{X_j}^2 / \sigma_e^2) + (\bar{S}_{X_j}^2 / \sigma_e^2)}$ where $\bar{S}_{X_j}^2$ reflects the mean of the observed variances of the J essentially parallel items and σ_e^2 reflects the expected variance of X if there is a complete lack of agreement among raters (LeBreton et al. 2005; LeBreton and Senter 2008).
Grand mean centering	Subtracting the grand mean from the individual observations of a variable ($X_{ij} - \bar{X}$).
Group mean centering	Subtracting the group mean from the individual observations of a variable ($X_{ij} - \bar{X}_j$).
Random intercept	An intercept that is allowed to vary between higher level entities ($\beta_{0j} = \gamma_{00} + u_{0j}$).
Random slope	A slope that is allowed to vary between higher level entities ($\beta_{1j} = \gamma_{10} + u_{1j}$).
Cross-level interaction effect	Indicates whether a relationship between an independent and a dependent lower level variable varies as a function of a higher level variable and thus determines whether the higher level variable moderates the relationship between the independent variable and the dependent variable across clusters ($\gamma_{11}X_{ij}Z_j$).

References

- Ahearne, M., MacKenzie, S. B., Podsakoff, P. M., Mathieu, J. E., & Lam, S. K. (2010). The role of consensus in sales team performance. *Journal of Marketing Research*, 47(3), 458–469.
- Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks: Sage.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csake (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3), 317–332.
- Albright, J. J., & Marinova, D. M. (2010). Estimating multilevel models using SPSS, Stata, SAS, and R. Working Paper. *Indiana University*, 1–35.
- Anderson, E. W., Fornell, C., & Mazvancheryl, S. K. (2004). Customer satisfaction and shareholder value. *Journal of marketing*, 68(4), 172–185.
- Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Proceedings of the 2007 JSM meeting, Section on Statistics in Epidemiology* (pp. 2531–2535). Alexandria: American Statistical Association.
- Asparouhov, T., & Muthén, B. (2013). Computing the Strictly Positive Satorra-Bentler Chi-Square Test in Mplus. Mplus Web Notes: No. 12, <https://www.statmodel.com/examples/webnotes/SB5.pdf>
- Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Charlotte: Information Age Publishing.
- Asparouhov, T., & Muthén, B. (2019). Latent Variable Centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1), 119–142.
- Auh, S., Menguc, B., & Jung, Y. S. (2014). Unpacking the relationship between empowering leadership and service-oriented citizenship behaviors: A multilevel approach. *Journal of the Academy of Marketing Science*, 42(5), 558–579.
- Bass, B. M., & Bass, R. (2009). *The Bass handbook of leadership: Theory, research, and managerial applications*. Free Press, New York: Simon and Schuster.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. et al. (2016). Package ‘lme4’. <https://cran.r-project.org/web/packages/lme4/lme4.pdf>. Accessed 24 Sept 2020.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107(2), 238–246.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino: Multivariate Software.
- Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, 9(1), 30.
- Bijmolt, T. H., & Pieters, R. G. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters*, 12(2), 157–169.
- Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, 1(4), 355–373.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.
- Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 401–445). San Francisco: Jossey-Bass.
- Bliese, P. D. (2016). Multilevel Modeling in R (2.6) – A brief introduction to R, the multilevel package and the nlme package. https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf. Accessed 24 Sept 2020.

- Boichuk, J. P., Bolander, W., Hall, Z. R., Ahearne, M., Zahn, W. J., & Nieves, M. (2014). Learned helplessness among newly hired salespeople and the influence of leadership. *Journal of Marketing*, 78(1), 95–111.
- Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, 17(3), 303–316.
- Bosker, R. J., Snijders, T. A. B., & Guldemon, H. (2003). *PINT (Power in Two-level Designs)*. Groningen: University of Groningen.
- Brady, M. K., Voorhees, C. M., & Brusco, M. J. (2012). Service sweetheating: Its antecedents and customer consequences. *Journal of Marketing*, 76(2), 81–98.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258.
- Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods*, 8(4), 375–409.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- De Jong, A., De Ruyter, K., & Lemmink, J. (2004). Antecedents and consequences of the service climate in boundary-spanning self-managing service teams. *Journal of Marketing*, 68(2), 18–35.
- Deadrick, D. L., Bennett, N., & Russell, C. J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management*, 23(6), 745–757.
- Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(3), 327–351.
- Donavan, D. T., Brown, T. J., & Mowen, J. C. (2004). Internal benefits of service-worker customer orientation: Job satisfaction, commitment, and organizational citizenship behaviors. *Journal of Marketing*, 68(1), 128–146.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. New York: Routledge.
- Edeling, A., & Fischer, M. (2016). Marketing's impact on firm value: Generalizations from a meta-analysis. *Journal of Marketing Research*, 53(4), 515–534.
- Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice* (No. 96). Hoboken: Sage.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121.
- Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R*. Boca Raton: CRC Press.
- Fu, F. Q., Richards, K. A., Hughes, D. E., & Jones, E. (2010). Motivating salespeople to sell new products: The relative influence of attitudes, subjective norms, and self-efficacy. *Journal of Marketing*, 74(6), 61–76.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton: Taylor and Francis.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3–8.
- Goff, B. G., Boles, J. S., Bellenger, D. N., & Stojack, C. (1997). The influence of salesperson selling behaviors on customer satisfaction with products. *Journal of Retailing*, 73(2), 171–183.
- Groening, C., Mittal, V., & “Anthea” Zhang, Y. (2016). Cross-validation of customer and employee signals and firm valuation. *Journal of Marketing Research*, 53(1), 61–76.
- Gruca, T. S., & Rego, L. L. (2005). Customer satisfaction, cash flow, and shareholder value. *Journal of Marketing*, 69(3), 115–130.
- Hamaker, E. L., & Klugkist, I. (2011). Bayesian estimation of multilevel models. In *Handbook of advanced multilevel analysis* (pp. 137–162). New York: Routledge.
- Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, 25(3), 365.

- Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus* (3rd ed.). New York: Routledge.
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (2012). *Multilevel modeling of categorical outcomes using IBM SPSS*. New York: Routledge.
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (2014). *Multilevel and longitudinal modeling with IBM SPSS* (2nd ed.). New York: Routledge.
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York: Wiley.
- Hitt, M. A., Beamish, P. W., Jackson, S. E., & Mathieu, J. E. (2007). Building theoretical and empirical bridges across levels: Multilevel research in management. *Academy of Management Journal*, 50(6), 1385–1399.
- Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24(5), 623–641.
- Hohenberg, S., & Homburg, C. (2016). Motivating sales reps for innovation selling in different cultures. *Journal of Marketing*, 80(2), 101–120.
- Homburg, C., Wieseke, J., & Bornemann, T. (2009a). Implementing the marketing concept at the employee-customer interface: The role of customer knowledge. *Journal of Marketing*, 73(4), 64–81.
- Homburg, C., Wieseke, J., & Hoyer, W. D. (2009b). Social identity and the service-profit chain. *Journal of Marketing*, 73(2), 38–54.
- Homburg, C., Müller, M., & Klarmann, M. (2011). When should the customer really be king? On the optimum level of salesperson customer orientation in sales encounters. *Journal of Marketing*, 75(2), 55–74.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.
- Hox, J. (2011). Panel Modeling: Random Coefficients and Covariance Structures. *Handbook of advanced multilevel analysis*, 137–162.
- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Thousand Oaks: Routledge.
- Hughes, D. E., & Ahearne, M. (2010). Energizing the reseller's sales force: The power of brand identification. *Journal of Marketing*, 74(4), 81–96.
- Hunter, G. K., & Panagopoulos, N. G. (2015). Commitment to technological change, sales force intelligence norms, and salesperson key outcomes. *Industrial Marketing Management*, 50, 162–179.
- James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69(1), 85.
- James, L. R., Demaree, R. G., & Wolf, G. (1993). rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, 78(2), 306.
- Johnson, J. S., Friend, S. B., & Horn, B. J. (2014). Levels of analysis and sources of data in sales research: A multilevel-multisource review. *Journal of Personal Selling & Sales Management*, 34(1), 70–86.
- Josephson, B. W., Johnson, J. L., & Mariadoss, B. J. (2016). Strategic marketing ambidexterity: Antecedents and financial consequences. *Journal of the Academy of Marketing Science*, 44(4), 539–554.
- Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, 19(2), 195–229.
- Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In: K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). Jossey-Bass.
- Krasnikov, A., & Jayachandran, S. (2008). The relative impact of marketing, research-and-development, and operations capabilities on firm performance. *Journal of Marketing*, 72(4), 1–11.
- Kreft, I. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Unpublished Report, California State University, Los Angeles.
- Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Los Angeles: Sage.

- Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1–21.
- Lai, M. H., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The “design effect smaller than two” rule. *The Journal of Experimental Education, 83*(3), 423–438.
- Lam, S. K., Ahearne, M., Mullins, R., Hayati, B., & Schillewaert, N. (2013). Exploring the dynamics of antecedents to consumer–brand identification with a new brand. *Journal of the Academy of Marketing Science, 41*(2), 234–252.
- Larivière, B., Keiningham, T. L., Aksoy, L., Yalçın, A., Morgeson, F. V., III, & Mithas, S. (2016). Modeling heterogeneity in the satisfaction, loyalty intention, and shareholder value linkage: A cross-industry analysis at the customer and firm levels. *Journal of Marketing Research, 53*(1), 91–109.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852.
- LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding rWG, rWG, rWG (J), and rWG (J). *Organizational Research Methods, 8*(1), 128–138.
- Liao, H., & Chuang, A. (2007). Transforming service employees and climate: A multilevel, multisource examination of transformational leadership in building long-term service relationships. *Journal of Applied Psychology, 92*(4), 1006–1019.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks: SAGE.
- Longford, N. T. (1989). To center or not to center. *Multilevel Modelling Newsletter, 1*(3), 7.
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.
- Mathieu, J., Ahearne, M., & Taylor, S. R. (2007). A longitudinal cross-level model of leader and salesperson influences on sales force technology use and performance. *Journal of Applied Psychology, 92*(2), 528.
- Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*(5), 951–966.
- Matsueda, R. L. (2014). Key advances of in the history of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 17–42). New York: Guilford Press.
- Maxham, J. G., III, Netemeyer, R. G., & Lichtenstein, D. R. (2008). The retail value chain: Linking employee perceptions to employee performance, customer evaluations, and store performance. *Marketing Science, 27*(2), 147–167.
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259–284.
- Mikolon, S., Kolberg, A., Haumann, T., & Wieseke, J. (2015). The complex role of complexity: How service providers can mitigate negative effects of perceived service complexity when selling professional services. *Journal of Service Research, 18*(4), 513–528.
- Mikolon, S., Alavi, S., & Reynders, A. (2020). The Catch-22 of countering a moral occupational stigma in employee-customer interactions. *The Academy of Management Journal*. <https://doi.org/10.5465/amj.2018.1487>.
- Misangyi, V. F., Elms, H., Greckhamer, T., & Lepine, J. A. (2006). A new perspective on a fundamental debate: a multilevel approach to industry, corporate, and business unit effects. *Strategic Management Journal, 27*(6), 571–590.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29*(1), 81–117.
- Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). New York: Taylor and Francis.
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus User's Guide* (7th ed.). Los Angeles: Muthén & Muthén.
- Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316.

- Netemeyer, R. G., Heilman, C. M., & Maxham, J. G., III. (2012). Identification with the retail organization and customer-perceived employee similarity: Effects on customer spending. *Journal of Applied Psychology*, *97*(5), 1049–1058.
- Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, *30*(1), 66–85.
- Palardy, G. J. (2011). Review of HLM 7. *Social Science Computer Review*, *29*(4), 515–520.
- Palmatier, R. W. (2008). Interfirm relational drivers of customer value. *Journal of Marketing*, *72*(4), 76–89.
- Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2016). nlme: Linear and nonlinear mixed effects models. R package version 3.1–128. <http://CRAN.R-project.org/package=nlme>. Accessed 24 Sept 2020.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233.
- Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, *18*(2), 161–182.
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station: STATA Press.
- Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2016a). *MLwiN Version 2.36*. Bristol: Centre for Multilevel Modelling, University of Bristol.
- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2016b). *A user's guide to MLwiN*. Bristol: Centre for Multilevel Modelling, University of Bristol.
- Raudenbush, S. W. (1989). Centering predictors in multilevel analysis: Choices and consequences. *Multilevel Modelling Newsletter*, *1*(2), 10–12.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & Du Toit, M. (2019a). *HLM 8: Linear and nonlinear modeling*. Lincolnwood: Scientific Software International.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & Du Toit, M. (2019b). *HLM 8*. Lincolnwood: Scientific Software International.
- Roschk, H., & Hosseinpour, M. (2020). Pleasant ambient scents: A meta-analysis of customer responses and situational contingencies. *Journal of Marketing*, *84*(1), 125–145.
- Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583–601.
- Satorra, A., & Bentler, P. M. (1999). A Scaled Difference Chi-square Test Statistic for Moment Structure Analysis. Working Paper. <https://econ-papers.upf.edu/en/onepaper.php?id=412>.
- Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248.
- Schneider, B., Salvaggio, A. N., & Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology*, *87*(2), 220.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *23*(4), 323–355.
- Snijders, T. A. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570–1573). Chichester: Wiley.
- Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics*, *18*(3), 237–259.
- Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis*. London: Sage.
- Steenkamp, J. B. E., Van Heerde, H. J., & Geyskens, I. (2010). What makes consumers willing to pay a price premium for national brands over private labels? *Journal of Marketing Research*, *47*(6), 1011–1024.
- Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. In *Annual meeting of the Psychometric Society* (pp. 424–453). Iowa City.

- Stoel, R. D., & Garre, F. G. (2011). Growth curve analysis using multilevel regression and structural equation modeling. In *Handbook of advanced multilevel analysis* (pp. 97–111). New York: Routledge.
- Troy, L. C., Hirunyawipada, T., & Paswan, A. K. (2008). Cross-functional integration and new product success: An empirical investigation of the findings. *Journal of Marketing*, 72(6), 132–146.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1), 1–10.
- Van der Borgh, M., de Jong, A., & Nijssen, E. J. (2019). Balancing frontliners' customer-and coworker-directed behaviors when serving business customers. *Journal of Service Research*, 22(3), 323–344.
- Walsh, G., Shiu, E., & Hassan, L. M. (2014). Cross-national advertising and behavioral intentions: A multilevel analysis. *Journal of International Marketing*, 22(1), 77–98.
- Wieseke, J., Lee, N., Broderick, A. J., Dawson, J. F., & Van Dick, R. (2008). Multilevel analysis in marketing research: Differentiating analytical outcomes. *Journal of Marketing Theory and Practice*, 16(4), 321–340.
- Wieseke, J., Ahearne, M., Lam, S. K., & Van Dick, R. (2009). The role of leaders in internal marketing. *Journal of Marketing*, 73(2), 123–145.
- Wieseke, J., Kraus, F., Ahearne, M., & Mikolon, S. (2012). Multiple identification foci and their countervailing effects on salespeople's negative headquarters stereotypes. *Journal of Marketing*, 76(3), 1–20.
- Wieseke, J., Alavi, S., & Habel, J. (2014). Willing to pay more, eager to pay less: The role of customer loyalty in price negotiations. *Journal of Marketing*, 78(6), 17–37.
- Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015). Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices. *Organizational Research Methods*, 18(4), 704–737.
- Wu, Y. W. B., & Wooldridge, P. J. (2005). The impact of centering first-level predictors on individual and contextual effects in multilevel data analysis. *Nursing Research*, 54(3), 212–216.



Panel Data Analysis: A Non-technical Introduction for Marketing Researchers

Arnd Vomberg and Simone Wies

Contents

Introduction: Relevance of Panel Data for Marketing Research	412
Process for Panel Data Analysis	414
Define the Research Question	415
Collect Panel Data	415
Prepare Panel Data	416
Explore Panel Data	419
Analyze Panel Data Models	426
Interpret and Present Results	444
Additional Methods in Panel Data Analysis	444
Robust Inference	444
Combining the Fixed Effects and Random Effects Estimators	445
Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach	450
Summary of the Discussed Estimators and Their Underlying Assumptions	452
Modeling a Price-Response-Function in Differences	455
Advanced Topics in Panel Data Analysis	457
Dynamic Panel Data Estimation	457
Random Slope Models: A Multilevel Model Approach to Panel Data	461
Addressing Measurement Error with Structural Equation Modeling Based on Panel Data	462
Conclusion	464
Cross-References	465
References	465

A. Vomberg (✉)

Marketing Department, University of Groningen, Groningen, The Netherlands
e-mail: A.E.Vomberg@rug.nl

S. Wies

Goethe University Frankfurt, Frankfurt, Germany
e-mail: wies@econ.uni-frankfurt.de

Abstract

The analysis of panel data is now part of the standard repertoire of marketers and marketing researchers. Compared to the analysis of cross-sectional data, panel data allow marketers to alleviate endogeneity concerns when linking an independent variable (e.g., price) to an outcome variable (e.g., sales volume). The more accurate estimates that result from panel data analysis help improve marketers' decision-making in focal areas such as price setting and marketing budget allocation. Besides, panel data allow marketers to track customer behavior changes and distinguish real loyalty effects (i.e., same customer repeatedly buys a brand) from spurious effects (i.e., the same number of, but each time different set of, customers buys a brand). This chapter provides a nontechnical introduction to panel data analysis. Marketers will learn how to manage and analyze panel datasets in Stata. They will learn about the focal panel data estimators (pooled OLS, fixed effects, and random effects estimator), their underlying assumptions, advantages, and pitfalls. Besides, we introduce the between effects estimator, the combined approach, the Hausman-Taylor approach, and the first differences estimator as further techniques to analyze panel data. Finally, readers will receive an introduction to advanced topics such as dynamic panel models, panel data multilevel modeling, and using panel data to address measurement errors.

Keywords

Cluster-robust standard errors · Dynamic panel data models · Endogeneity · Fixed effects estimator · Hausman test · Hausman-Taylor method · Measurement error · Omitted variable bias · Panel data analysis · Pooled OLS · Random effects estimator · Serial correlation

Introduction: Relevance of Panel Data for Marketing Research

The analysis of panel data is now part of the standard repertoire of marketers and marketing researchers. Panel data, sometimes referred to as longitudinal data, contain observations about different cross-sectional units, also called clusters, across time. Hence, like cross-sectional data, panel data contain observations across a collection of clusters, and like time-series data, panel data contain observations about these clusters repeatedly collected over time. Examples of panel data include the following:

- Retail scanner data: Retailers track sales volume for their products over time.
- Online transaction data: Online retailers collect information about their customers over time.
- Market research institute data: Organizations collect consumer survey data for brands and products over time. For instance, the American Customer Satisfaction Index (ACSI) tracks the evolution of customer satisfaction for different

companies over time (e.g., Fornell et al. 1996). Similarly, the Young&Rubicum Brand Asset Valuator monitors consumers' brand sentiment for brands over time (e.g., chapter ► [“Assessing the Financial Impact of Brand Equity with Short Time-Series Data”](#) by Mizik and Pavlov in this Handbook).

Typically, the collection of panel data requires huge resources in time and money from investigators. Yet, as compared to cross-sectional data, panel data offer substantial advantages that can easily compensate for their data collection efforts.

First, panel data allow to study changes at the individual level and to disentangle real loyalty effects from so-called spurious carryover effects. As an example, aggregated brand sales data might resemble a stable pattern, indicating high loyalty. However, tracking changes at the individual level provides insights into whether the aggregated effect results from a loyalty effect (i.e., the same consumers purchase regularly) or an attraction effect (i.e., the company attracts new consumers but existing consumers do not repurchase).

Second, panel data allow addressing a potential omitted variable bias, a serious endogeneity threat in observational data. Broadly, endogeneity refers to a situation in which an independent variable is correlated with the error term, violating the basic exogeneity assumption of OLS and causing all coefficient estimates of the model to be biased and inconsistent (both properties that lead to misleading hypothesis tests). For instance, the investigator might be interested to know whether a company's distribution intensity contributes to its financial performance. If she omitted company factors such as brand strength, which likely drives both financial performance and access to distribution channels, regression results might overstate the impact of distribution intensity. In panel data analysis, we can control for general company effects, thereby helping to rule out many threats from omitted variables (see also chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#) by Ebbes et al. in this Handbook).

Third and relatedly, in contrast to typical applications with cross-sectional data, panel data allow to include time lags between dependent and independent variables. Thereby, panel data open up opportunities for novel research questions and, again, increase the researchers' ability to rule out endogeneity concerns (chapters ► [“Experiments in Market Research”](#) by Bornemann and Hattula, and ► [“Field Experiments”](#) by Valli et al. in this Handbook further discuss criteria for evaluating causality).

This chapter aims to introduce a general approach to analyzing panel data from an applied perspective, focusing on panel data management and analysis. Figure 1 outlines the process along which we structure our chapter. In the section [“Process for Panel Data Analysis,”](#) we discuss the panel data research process step-by-step. Based on a real-life research example, we first define our research objective (section [“Define the Research Question”](#)); specifically, we are interested in estimating a price-response-function for a company's newly launched headphone. Then, we discuss the collected dataset, which can be downloaded from the Handbook's Data Appendix (section [“Collect Panel Data”](#)), and explain how researchers can prepare the data for the analyses (section [“Prepare Panel Data”](#)). Afterward, readers learn how to explore the unique structure in panel data (section [“Explore Panel Data”](#)), and

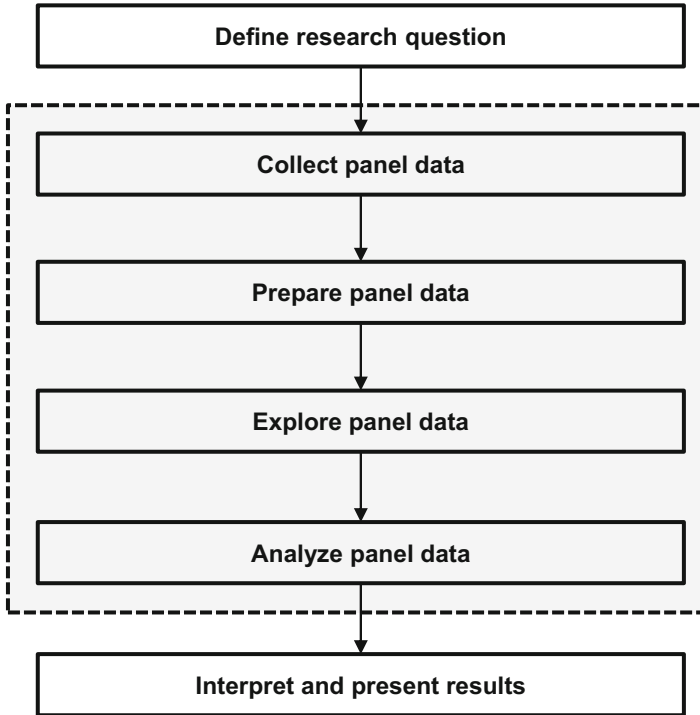


Fig. 1 Panel data research process

we discuss the focal panel data estimators (section “[Analyze Panel Data Models](#)”). We conclude this section with a short interpretation of the results (section “[Interpret and Present Results](#)”). In the section “[Additional Methods in Panel Data Analysis](#),” we provide more in-depth background information on panel data estimators. In the section “[Advanced Topics in Panel Data Analysis](#),” we explore more advanced topics in panel data analysis, including dynamic panel models and random coefficient models. Note that we keep a strictly intuitive and applied approach throughout the text but point the reader to the many excellent resources available that provide a more formal treatment of panel data estimation.

Process for Panel Data Analysis

Any research process should start with a clearly defined research question and end with a meaningful discussion of the results. However, a research process that involves panel data considerably differs from a process based on cross-sectional data alone regarding the data management and analysis part. Given that data management and analysis is more intricate for panel data, we devote special attention to describing how to prepare, explore, and analyze panel data.

In the spirit of an applied approach to the panel data research process, we guide the discussion in this chapter along a real-life dataset, using the statistical software Stata. Although most statistical software packages are well equipped to analyze panel data, Stata is particularly suited for such analysis given its convenient and efficient command structure for panel data.

Define the Research Question

The most important issue relates to the relationships researchers want to investigate. Developing an original and relevant research question is not trivial, yet, as mentioned earlier, panel data can offer exciting opportunities for examining more complex phenomena.

In this chapter, we want to aid a consumer electronics company on its pricing strategy for a newly launched headphone. We are interested in understanding how price-setting influences sales volume. This question is relevant yet challenging for the company as the headphone represents an expansion into a new category, with which the company only has limited experience. Economic theory suggests relying on so-called price-response-functions for price management. Price-response-functions link price to sales volume and represent a prerequisite for price optimization.

Collect Panel Data

Data for estimating price-response-functions can come from various sources, such as customer surveys (chapter ► [“Crafting Survey Research: A Systematic Process for Conducting Survey Research”](#) by Vomberg and Klarman in this Handbook), experiments (chapters ► [“Choice-Based Conjoint Analysis”](#) by Eggers et al. and ► [“Willingness to Pay”](#) by Klingemann et al. in this Handbook), or market-level data. In this chapter, we focus on market-level data. The real-life case we follow throughout the chapter is based on a dataset of a medium-sized European consumer electronics company. The company created the dataset to estimate a price-response-function for its most recently launched headphone.

An excerpt of the corresponding dataset available for analysis contains weekly sales data (“sales”) of the headphone model across $n = 19$ of the company’s franchised stores (cross-sectional component, labeled as “storeid”) across a time period of $T = 82$ weeks (time-series component, labeled as “week”), ranging from weeks 37 to 118. The company has recorded the headphone’s retail price (“price”) for each store and week. Besides, the company has collected information about whether the store featured major promotion activities for the headphone in a given week (dummy-coded “promo”: 1 “promotional activity in the focal week” 0 “no promotional activity in the focal week”). Moreover, the company has information on two general store descriptors: whether the store design includes multiple floors (dummy-coded “floor”: 1 “multiple floors” 0 “single floor”) and whether the store

is in a premium location with high traffic (dummy-coded “location”: 1 “premium location” 0 “nonpremium location”). The data is initially split across three different data files, mimicking the often disordered situation researchers and companies face when they seek to analyze data.

We begin by reviewing the steps necessary to combine the three files to arrive at the final dataset that we will use in the chapter’s remainder. Access to the different datasets is available via the Data Appendix. We provide the three individual raw datasets (“Sales_Original.dta,” “Sales_Additional.dta,” and “Explanatory.dta”), as well as the final dataset (“Dataset_Final.dta”). “Sales_Original.dta” contains the original headphone sales data the firm collected across its sampled stores. “Sales_Additional.dta” includes observations of an additional store for which data was shared separately. “Explanatory_Long.dta” features the retail price variable as well as the three other explanatory variables (promotional activities, “promo,” store floor design, “floor,” and store premium location, “location”). The Data Appendix also includes the script (the so-called “do-file” in Stata language) that summarizes the programming commands we use and complements the text with more detailed comments regarding the syntax.

Prepare Panel Data

Our overview begins with looking at three data management challenges that deserve our attention when working with panel datasets. Specifically, we discuss how to transform, combine, and prepare these datasets, so they are ready to use for our statistical analysis.

Transforming the Data Structure: Converting Wide Format to Long Format

In general, there are two ways of organizing panel data: wide format and long format. In wide format, a cluster’s repeated measurements are stored in a single row, and each measurement appears in a separate column. Hence, the dataset has as many rows as it has clusters and as many columns as time periods \times number of variables. Many datasets, especially when coming from commercial data providers, store panel data in wide format. Our original sales dataset (“Sales_Original.dta”) also comes in wide format, with each row containing sales data for an individual store across all weeks. Using the `list` command, we can see a subset of the first five stores and their respective sales volume in the first ten weeks in the dataset in wide format (Fig. 2).

```
. list storeid sales37-sales46 in 1/5, clean
```

	storeid	sales37	sales38	sales39	sales40	sales41	sales42	sales43	sales44	sales45	sales46
1.	1	309	339	291	208	224	194	181	179	168	312
2.	2	411	277	262	198	205	192	177	167	151	321
3.	3	161	162	161	164	173	166	156	165	144	166
4.	4	235	207	203	195	183	164	197	204	156	189
5.	5	139	139	144	144	141	143	140	63	62	61

Fig. 2 Dataset in wide format

Fig. 3 Converting data from wide to long format

```

. reshape long sales, i(storeid) j(week)

Data                wide  ->  long
-----
Number of obs.      18   ->  1476
Number of variables  83   ->   3
j variable (82 values)
xij variables:
    sales37 sales38 ... sales118 ->  sales
    
```

Fig. 4 Dataset in long format

```

. list storeid sales in 1/5, clean

      storeid  sales
    1.         1    309
    2.         1    339
    3.         1    291
    4.         1    208
    5.         1    224
    
```

In contrast, in long format, each measurement occupies one row with as many rows as clusters \times time periods (for balanced data: $n \times T$). The number of columns equals the number of variables. We can convert the data from wide to long format by using the Stata `reshape long` command (Fig. 3).

The command interprets the sales variables’ suffixes (e.g., `sales37`, `sales38`) as denoting the grouping that needs to be expanded to long-form. As visible in the output, we expanded the dataset from 18 observations (18 stores) to 1,476 observations (1,476 store-week pairs). In addition, Stata created a week identifier, called “week,” and collapsed the sales variables (“`sales37`” – “`sales118`”) to a new variable, called “sales.” Listing the first five observations of the new long-format dataset (using the `list` command) reveals the following data structure (Fig. 4).

Both datasets have exactly the same information. For panel estimation, however, most statistical software packages, Stata included, require the panel data to be organized in long format, with each observation being a distinct cluster-time pair.

Combining Panel Datasets: Appending and Merging Datasets

Data is often scattered across several data files, and our data example is no exception. It is necessary to combine the individual files into a new file containing all the variables and all the observations needed for analysis. There are different operations through which one can combine panel datasets, and we elaborate on the three most popular operations. First, one can add new *observations* to a given dataset, thereby *appending* data vertically by expanding the number of rows. This operation is typically used when one receives additional data for the same variables and seeks to add these new observations to the first dataset. In our data example, assume the marketing manager responsible for the study was able to add another store to the sample (see “Sales_ Additional.dta”), containing 87 observations. The investigator now needs to append this store’s data to the transformed long-format dataset. The `append` command is simple to use and adds the rows from the second dataset to the end of the first dataset

Fig. 5 Merging datasets

```

.      merge 1:1 storeid week using "Explanatory.dta"

Result      # of obs.
-----
not matched          0
matched             1,563  (_merge==3)

```

(append using “Sales_Additional.dta”). This expands the first dataset to 1,563 observations.

Second, one can add new *variables* to a given dataset, thereby *merging* datasets horizontally by expanding the number of columns. This operation is needed when variables are stored across different data files. In our data example, we want to add a set of variables that help explain sales levels. These explanatory variables include the retail price, promotional activities, premium location, and floor design and are stored in “Explanatory.dta.” Merging panel datasets requires that both datasets have variables in common on which the merging is based. If matching variables are found, merging two datasets is straightforward using the `merge` command. In our example, “store id” and “week” are the matching variables. We first sort on these matching variables and then perform the merge for a one-to-one (1 : 1) matching (Fig. 5). Note that alternative matching procedures involve one-to-many (1 : m) or many-to-one (m : 1) matching, depending on how the respective datasets are structured. For instance, if we would like to match a dataset that includes zip codes for the stores, we could use the many-to-one matching, knowing that we have multiple stores situated in the same zip code area.

The `merge` command creates a new variable called “_merge.” This variable takes a value of 1 if the observation is only contained in the first dataset, a value of 2 if it is included only in the second dataset, and a value of 3 if the observation is present in both datasets. In our example, the observations of the matching variables perfectly match and are present in both datasets. As a result, the “_merge” variable takes on the value of 3 for all 1,563 observations. In contrast to the `append` command, the `merge` command does not add new observations to the dataset so the sample size remains unchanged. After we performed the operation, we can delete the “_merge” variable (`drop _merge`).

Finally, Stata offers a third type of combining datasets in which one can merge datasets horizontally but form pairwise combinations within-cluster, using the `joinby` command. This command is similar to merging datasets but creates all possible combinations of the observations across both datasets. While not applicable to our data example, it can be a useful command in other settings. For instance, imagine two datasets. The first contains a list of executive team members (i.e., CEO and CMO) across several companies. The second includes a list of awards the companies have received (e.g., Innovative Design Award, Best Place to Work Award). Both datasets contain a company identifier variable, which links the executives and awards that belong to the same company. Using the `joinby` command, we can easily combine the datasets and create a new dataset that includes all combinations of executives and awards, hence one row for each executive and award combination per company.

frame can be regular (e.g., data is collected every week, month, or year) or irregular (e.g., data is collected on specific occasions). Picking up on panel data's dual nature, it is common to classify panel datasets based on the relation between the number of clusters and the number of time periods observed. Panels in which a relatively large number of cross-sectional clusters is tracked over a rather short period of time are referred to as short or micropanels. For instance, retailers may track customer satisfaction scores for many consumers (cross-sectional component) but only for a few years (time component).

In contrast, panels with frequent measurements for a relatively small section of units vis-à-vis a larger number of time periods are referred to as long or macro-units. For instance, we might track the monthly consumer confidence index back to the 1960s (time component) for only a few EU countries (cross-sectional component). It is worth noting that most datasets in marketing are micropanels and that the more common types of panel estimators are derived under the assumption of relatively fewer time periods and a larger number of cross-sectional units (Wooldridge 2016). Please note that in our example, for illustrative purposes, we work with a reduced dataset that includes only a limited number of stores (cross-sectional units).

Focal Challenge of Panel Data Analysis: Nonindependent Observations

Irrespective of the type of panel dataset, we face one focal challenge in analyzing panel data: repeated measurements of a given cluster over time are nonindependent. This has severe implications for estimating our model, which we discuss next. To illustrate this point, we can think of panel data as a nested data structure. Using an extract from our example, we visualize this idea in Fig. 7. Here, we observe headphone sales for three of the stores at different points in time. The figure illustrates that each sales measurement (lower-level unit) is nested in a particular store measurement (higher-level unit).

The hierarchical nature highlights that individual measurements are likely not independent from each other; rather, sales measurements from one store have more in common (i.e., are more strongly correlated) than they have with sales from different stores. Intuitively, if sales in week 40 were independent of a store's prior sales, a good prediction would be the mean sales levels for week 39 and week 38 across all stores' observations. However, it seems that a much better prediction is to rely on the store's prior measurements. This point can be visualized by plotting the



Fig. 7 Illustration of panel structure

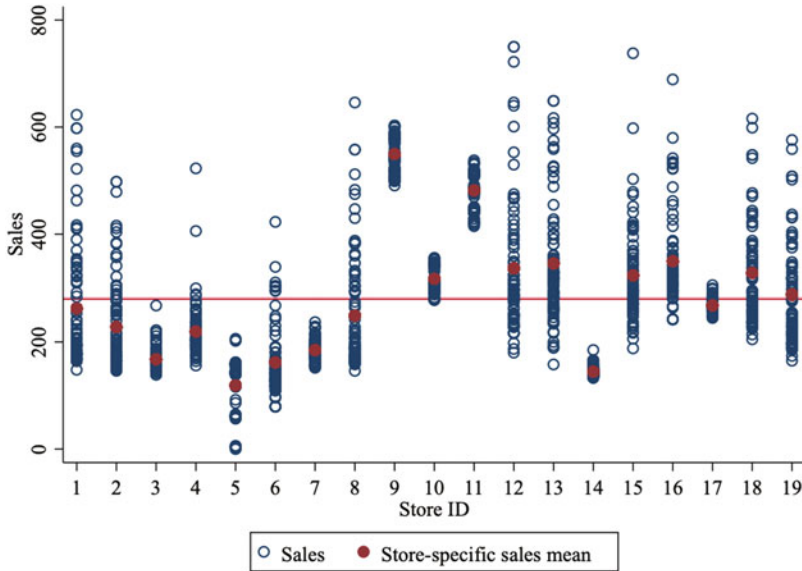


Fig. 8 Within and between dependence across stores

overall mean of sales, also called grand mean, across all measurements and the store-specific means of sales, as shown in Fig. 8. As can be seen in Fig. 8, there is substantial within dependence in the sales data (i.e., measurements within stores are similar).

As a more formal treatment of within dependence, we can use Stata’s time-series lag operators (`L.`) and the `correlate` command (`corr sales L1. sales`) and inspect the serial correlation in sales (see also section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)” for formal tests of serial correlation). Results suggest that headphone sales correlate substantially ($\rho = 0.67$) over time within a store; hence, there is a high serial correlation present in the sales data, and observations are nonindependent from each other. This dependency, in turn, violates a key assumption of traditional OLS estimation and foreshadows why standard OLS estimation is not feasible for panel data analysis (as detailed in section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)”). A thought experiment based on Fig. 7 further underpins this point: if sales measurements within a store are very similar, we do not observe nine observations (three measurements for three stores) but effectively only three. Put differently, it reduces the effective size of the sample we can use in estimating the model, which in turn has implications for calculating the standard errors we use for making inferences. We will revisit the challenge of nonindependent observations in the remainder of the chapter. In doing so, we will shift to the more common terminology of within and between variance instead of within and between dependency. Within variance describes the variation within a cluster across time. Between variance describes the variation between clusters.

Dependent Variable: Between and Within Variance

Given nonindependent observations being the focal challenge in panel data, it is crucial to understand the degree of such dependency. The `xtsum` command provides valuable information on the relative importance of within and between variance in the dependent variable. Figure 9 demonstrates the result for sales ($Sales_{it}$) in our data example.

Besides providing some general descriptive statistics (including mean and range of the variable, as well as the number of observations, N , number of clusters, n , and number of time periods, t , in the sample), the output distinguishes between an overall, a between, and a within variance in the sales data. The within component focuses on particular clusters (i.e., stores) and describes the data within these clusters (i.e., sales levels from across all periods for an individual store). The between component takes the average per cluster (i.e., average sales level for a particular store) and describes the data based on these averages. The overall component treats the observations as entirely independent and calculates the respective measures without considering their panel nesting.

To support the discussion that follows, we detail how the overall, between, and within variance is calculated. Let

$$\overline{\text{Sales}} = \frac{\sum_{i=1}^n \sum_{t=1}^T \text{Sales}_{it}}{n \times T} \tag{1}$$

be the overall mean of the dependent variable across all observations, also called the grand mean, and

$$\overline{\text{Sales}}_i = \frac{\sum_{t=1}^T \text{Sales}_{it}}{T} \tag{2}$$

be the corresponding within mean for each cluster i (i.e., store); we can easily compute the respective variances.

The overall variance is calculated in the same way as in cross-sectional data without considering any panel structure nesting:

$$\hat{\sigma}_{\text{Sales; Overall}}^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T \left(\text{Sales}_{it} - \overline{\text{Sales}} \right)^2}{n \times T - 1}. \tag{3}$$

```

.       xtsum sales

```

Variable	Mean	Std. Dev.	Min	Max	Observations
sales overall	280.1906	131.1669	0	750	N = 1558
sales between		110.3184	118.9024	549.5976	n = 19
sales within		75.28426	92.88575	695.0443	T = 82

Fig. 9 Decomposing the dependent variable sales volume

To calculate the within variance, we employ the following formula:

$$\hat{\sigma}_{\text{Sales}; \text{Within}}^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (\text{Sales}_{it} - \overline{\text{Sales}}_i)^2}{n \times (T - 1)}. \quad (4)$$

When calculating the within variance, some statistical programs, Stata included, add the grand mean back to the within mean in the numerator $(\text{Sales}_{it} - \overline{\text{Sales}}_i + \overline{\text{Sales}})$ to make results comparable across overall, within, and between variance.

The between variance is given as:

$$\hat{\sigma}_{\text{Sales}; \text{Between}}^2 = \frac{\sum_{i=1}^n (\overline{\text{Sales}}_i - \overline{\text{Sales}})^2}{n - 1} \quad (5)$$

For our sample of 1,558 observations, we estimate a between variance of 110.31 and a within variance of 75.28, indicating that in our sample, the variance between clusters is larger than the variance over time within clusters.

Independent Variables: Time-Constant and Time-Varying Variables

While the dependent variable is, by definition, time-varying in the context of panel data, the independent variables can be time-constant or time-varying. In the following, we will refer to time-constant variables as Z . Those variables constitute cluster characteristics that are stable over the observation period. For instance, in modeling firm sales, McAlister et al. (2016) include the firm's type of business strategy to achieve a competitive advantage as a time-constant independent variable that does not change over the observation period. In our data example, a store's location ($Location_i$) and a store's floor design ($Floor_i$) are both time-constant variables that we include as independent variables in explaining sales.

Other variables, however, may change over time and are therefore called time-varying variables. We will refer to these variables as X in the course of this chapter. Ataman et al. (2010), for example, study brand sales levels and include brand distribution breadth as a time-varying independent variable. Our data example includes retail price ($Price_{it}$) and promotional activities ($Promo_{it}$) as independent variables, which vary over time. We can use Stata's `xtsum` command to confirm the type of variation in the independent variables. Figure 10 shows that $Location_i$ and $Floor_i$ are time-constant as they exhibit zero within variation. $Price_{it}$ and $Promo_{it}$, however, are time-varying and display variation within and between stores.

Since three of the four explanatory variables are binary, a further helpful command is `xttab`, which decomposes categorical variables into within and between variation. Figure 11 shows that, overall, roughly 11% of the store-week sales observations result from stores with a multiple floor store design. The between column repeats the breakdown but does so in terms of stores rather than store-weeks. Given we have a balanced sample with the same number of stores and weeks,

```

.      xtsum location floor price promo

```

Variable		Mean	Std. Dev.	Min	Max	Observations
location	overall	.6842105	.4649788	0	1	N = 1558
	between		.4775669	0	1	n = 19
	within		0	.6842105	.6842105	T = 82
floor	overall	.1052632	.3069907	0	1	N = 1558
	between		.3153018	0	1	n = 19
	within		0	.1052632	.1052632	T = 82
price	overall	55.38318	12.596	29	85	N = 1558
	between		12.40116	38.42683	78.84146	n = 19
	within		3.587608	17.31001	64.3466	T = 82
promo	overall	.2727856	.4455345	0	1	N = 1558
	between		.1266187	.097561	.5487805	n = 19
	within		.4281387	-.2759949	1.175225	T = 82

Fig. 10 Decomposing the independent variables

```

.      xttab floor

```

floor	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	1394	89.47	17	89.47	100.00
1	164	10.53	2	10.53	100.00
Total	1558	100.00	19	100.00	100.00

(n = 19)

Fig. 11 Decomposing a time-constant categorical variable

and $Floor_i$ varies only between stores, we report the same proportions of observations as in the overall column. Finally, the within percent tells us the fraction of times a store reported a given value of the $Floor_i$ variable. For instance, conditional on a store ever having a $Floor_i$ value of 0 (first line), 100% of its observations have a $Floor_i$ value of 0. These numbers indicate the stability of the variable within a cluster. By definition, a time-constant variable will have a tabulation within-percent of 100, while time-varying variables will have a tabulation within-percent below 100 (Fig. 12).

The between-percent informs us about the percentage of stores that have ever reported a given value of the variable, in this case, initiate promotional activities during the observation period. Since all sampled stores had periods in which they engaged in promotional activities ($Promo_{it}$ value of 1) and periods in which they did not engage in promotional activities ($Promo_{it}$ value of 0), the between percentage adds up to 200.

Additional information on within and between variation for categorical variables can also be retrieved through the `xttrans` command, which provides transition probabilities within a cluster from one period to the next. Transition probabilities

```

.          xttab promo

```

promo	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	1133	72.72	19	100.00	72.72
1	425	27.28	19	100.00	27.28
Total	1558	100.00	38	200.00	50.00

(n = 19)

Fig. 12 Decomposing a time-varying categorical variable

Fig. 13 Transition probabilities for time-varying categorical variable

```

.          xttrans promo

```

Promotion	Promotion		Total
	0	1	
0	77.72	22.28	100.00
1	59.95	40.05	100.00
Total	72.90	27.10	100.00

Fig. 14 Transition probabilities for a time-constant categorical variable

```

.          xttrans floor

```

Multiple floor store	Multiple floor store		Total
	0	1	
0	100.00	0.00	100.00
1	0.00	100.00	100.00
Total	89.47	10.53	100.00

describe the probabilities of changing the state from one value of a variable to another value.

Figure 13 reports the results for the $Promo_{it}$ variable. As visible from the output, about 40% of the store-week observations that featured promotional activities in one period continue to feature promotional activities in the next period. For those store-week observations that did not feature promotional activities in one period, 22% start doing so in the next period. For time-constant variables, such as $Floor_i$, the diagonal entries will always be 100%, and the off-diagonal entries will always be 0% (Fig. 14).

Finally, it is worth noting that whether the independent variable is time -varying or time -constant has implications for the type of variance we can explain in the dependent variable. Specifically, time-constant variables can only account for between variation in the dependent variable. In contrast, time-varying independent variables can explain both between and within variation in the dependent variable, depending to which extent they vary within and between clusters. For instance, Vomberg et al. (2015) observe that market share tends to be relatively stable over time, while it differs substantially between companies. Hence, this time-varying variable will mainly explain between variation in the dependent variable.

The type of time dependence in the independent variables also helps identify the serial correlation sources in the dependent variable. First, variables that do not change over time (Z) will cause the dependent variable to have similar values across periods. For instance, assume we would like to explain a firm's brand equity levels by its branding strategy (multibrands vs. mono-brand). Since the type of branding strategy a firm pursues is likely to be constant over time (Rao et al. 2004), brand equity levels will be close to each other across periods. In statistical language, this cause of serial correlation is referred to as spurious state dependence. Two measurements (of the dependent variable) may be correlated because they are associated with a further variable (that characterizes the cluster).

Second, time-varying variables (X) might also be a source of serial correlation if they change only slowly over time. For instance, customer satisfaction is likely to drive brand equity. Yet, while customer satisfaction levels can change over time, they might only change to a small amount or just every other year. If customer satisfaction levels influence brand equity levels, the serial correlation among the time-varying customer satisfaction levels leads to statistical dependencies in the brand equity levels. This is another representation of spurious state dependence.

Third, the dependent variable's current value might directly influence the dependent variable's level in the next period. For instance, brand familiarity levels in the current year are likely to impact brand familiarity levels in the next year directly, given multiplier and spillover effects through word-of-mouth, like Lovett et al. (2013) show. This argument provides a substantive and direct cause of serial correlation, referred to as true state dependence.

From this discussion, it follows that if we were able to control for all relevant time-constant (Z) and time-varying (X) independent variables and included a lagged dependent variable (y_{t-1}), measurements of the dependent variable should be independent over time and serial correlation would be zero.

Analyze Panel Data Models

After collecting, preparing, and exploring the panel data structure, we are now ready to turn to the actual analysis of panel data. In this section, we will discuss three estimators that investigators can use to estimate the price-response-function: the (1) pooled OLS (section "[Pooled OLS Estimator: Ignoring the Panel Structure](#)"), (2) fixed effects (section "[Fixed Effects Estimator](#)"), and (3) random effects estimator (section "[Random Effects Estimator](#)"). We will discuss the latter two estimators in greater detail since those estimators explicitly account for the panel structure (in contrast to the pooled OLS estimator ignoring the panel structure). Besides, "most panel data models are estimated under either the fixed or random effects assumption" (Verbeek 2017, p. 384). We, therefore, call the fixed effects and random estimators proper panel estimators. To recall, here and in the remainder of the chapter, "panel structure" refers to nonindependent observations grouped by clusters over time. We will also outline the relationships between the estimators

(section “[Relationship Between Pooled OLS, Fixed Effects, and Random Effects Estimators](#)”), discuss why they may deliver divergent results (section “[What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?](#)”), and provide guidance on which estimator should be selected (section “[Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#)”). In section “[Additional Methods in Panel Data Analysis](#),” we will highlight additional methods useful in analyzing panel data. We will point to more advanced panel data analysis strategies in section “[Advanced Topics in Panel Data Analysis](#).”

Pooled OLS Estimator: Ignoring the Panel Structure

We will start our discussion by applying the cross-sectional OLS approach to panel data (chapter ► “[Regression Analysis](#)” by Skiera et al. provide a detailed discussion of OLS regression analysis in this Handbook). This technique is referred to as pooled OLS (POLS). The name reflects that POLS pools all observations across the individual cross-sections and time periods without accounting for the panel structure.

As a first step, we need to decide which variables to include in our regression model. Naturally, we include the impact of the price ($Price_{it}$), as this is our focal variable of interest. Since we rely on observational market-level data (instead of randomized experimental data), we also want to account for relevant control variables that explain variation in the dependent variable and reduce the threat of omitted variables that could bias our results. That is, by including these control variables, we can account for factors that might affect both sales volume and price.

We include an indicator for promotion periods ($Promo_{it}$) which by definition influences price and is also likely to boost sales. We include a variable that distinguishes premium locations ($Location_i$) because location advantages might allow managers to realize higher sales volume and to set higher prices. Finally, we account for multiple floor store design ($Floor_i$). Stores with several floors likely contain a larger sales area and a broader assortment. Consequently, customers face a more appealing shopping experience, leading to higher sales volume (e.g., customers may stay longer in the store and purchase more products) but possibly also higher willingness to pay. Moreover, we account for seasonal differences across weeks by including dummy variables for all but one week; we select the first week in our sample as the reference category. Controlling for the effect of weeks (also called week-fixed effects) picks up market-wide developments that equally impact all stores, for instance, stimulated by online buzz about the company or increased media attention in a focal week.

As a second step, we need to decide on the functional form of the price-response-function. For the sake of simplicity, we will focus on a linear model. However, in general, investigators could also estimate price-response-functions in different forms, such as a multiplicative price-response-function. We formulate the following regression equation, which also contains an error term (ξ , pronounced xi):

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \delta \text{WEEK} + \xi_{it} \quad (6)$$

The variable subscript notation confirms that Price_{it} and Promo_{it} are time-varying variables, varying by store i and week t , and that Location_i and Floor_i are time-constant variables that only differ by store i . The bolded expression for weeks indicates that we include a vector of week dummy variables.

POLS relies on the typical cross-sectional OLS assumptions. Thereby, one of the focal assumptions is that the error term does not display serial correlation. As outlined in section “[Independent Variables: Time-Constant and Time-Varying Variables](#),” unless we account for *all* time-varying and time-constant variables that impact the dependent variable, the error term is likely correlated between two measurements of the same cluster.

We will demonstrate that serial correlation persists in our model, even after controlling for price, promotion activities, premium location, multiple floor design, and seasonal week effects. We first predict the regression residuals from Eq. 6 (see Eq. 7):

$$\widehat{\xi}_{it} = \text{Sales}_{it} - (\beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \delta \text{WEEK}) \quad (7)$$

As a measure of serial correlation, we next derive pairwise correlations between different time periods. In Table 1, we compare the serial correlation in the residuals (`predict xi_hat, resid`) of the POLS model (Panel a) and in the raw sales data (Panel b) for 5 selected weeks from our dataset (weeks 37–41) to provide a general intuition. Besides eyeballing across the selected correlations, we can undertake more formal tests of serial correlation, either through the `correlate` command (e.g., `corr xi_hat L1.xi_hat`), as we did in section “[Focal Challenge of Panel Data Analysis: Nonindependent Observations](#),” or through the `regress` command (e.g., `reg xi_hat L1.xi_hat, beta`), as Wooldridge (2010) recommends. Also, Stata offers the user-written `xtserial` command as an alternative way to estimate serial correlation (available from SSC `findit xtserial`). From Table 1, it is easy to see that the serial correlation in the POLS residuals is smaller than the serial correlation in the raw sales data. The inclusion of the independent variables explains the reduction in serial correlation. However, it is also apparent that substantial serial correlation persists even after including the independent variables ($\text{corr}(\xi_t, \xi_{t+1}) = 0.61; p < 0.01$). Variables not included in the model explain the remaining serial correlation.

Serial correlation represents a common problem in POLS estimation, rendering standard errors calculated under the typical OLS assumptions misleading for panel data applications (Verbeek 2017). To partly account for serial correlation, investigators can rely on cluster-robust standard errors. Cluster-robust standard errors are computed differently than common OLS standard errors and account for the fact that the error structure differs across clusters (section “[Robust Inference](#)” details this point). The Stata suffix `cluster(clustvar)` following the

Table 1 Pairwise correlations of POLS residuals and raw data between weeks

Panel a)	POLS residual					Panel b)	Raw data				
	ξ_{37}	ξ_{38}	ξ_{39}	ξ_{40}	ξ_{41}		Sales ₃₇	Sales ₃₈	Sales ₃₉	Sales ₄₀	Sales ₄₁
ξ_{37}	1.00					Sales ₃₇	1.00				
ξ_{38}	0.77	1.00				Sales ₃₈	0.79	1.00			
ξ_{39}	0.55	0.79	1.00			Sales ₃₉	0.87	0.86	1.00		
ξ_{40}	0.56	0.89	0.87	1.00		Sales ₄₀	0.71	0.90	0.89	1.00	
ξ_{41}	0.48	0.69	0.96	0.83	1.00	Sales ₄₁	0.74	0.77	0.92	0.92	
Overall corr _{Sample} (ξ_t, ξ_{t+1}) = 0.61						Overall corr _{Sample} (Sales _t , Sales _{t+1}) = 0.67					

regression command indicates the use of cluster-robust standard errors. Thereby, `clustvar` represents the cross-sectional unit around which we cluster the standard errors. Since we want to cluster at the store-level, we use “storeid” as our clustering variable. We estimate the regression model with the following Stata syntax:

```
reg sales price promo location floor i.week, cluster(storeid)
```

In Table 2, we report the POLS estimates. The only difference between the two models is how the standard errors are calculated. Model 1 presents the POLS results with standard errors that follow from the common OLS assumptions. Model 2 relies on cluster-robust standard errors. As in the case of cross-sectional OLS, investigators can evaluate the overall model fit via R^2 values, the overall significance of the model with an F-statistic, and the statistical significance of individual regression coefficients with t-tests.

In our case, the regression results reveal that, in line with economic theory, the price has a negative relationship with sales volume ($\beta_1 = -1.48$). However, the effect of price on sales is only statistically significant in Model 1 ($p < 0.01$). Accounting for serial correlation with cluster-robust standard errors increases the estimated standard errors (Model 2), and in our example, the effect of price on sales turns insignificant. Note that clustering the standard errors does not affect the estimated regression coefficients.

To conclude, applying POLS with cluster-robust standard errors to panel data considers the panel structure to some extent but treats “it a nuisance, not as a phenomenon we are interested in” (Rabe-Hesketh and Skrondal 2012, p. 105). As a result, the challenge of serial correlation remains, and the usefulness of the estimation results is limited. In the following, we will focus on two estimators that more explicitly leverage the panel structure: the fixed effects estimator (section “Fixed Effects Estimator”) and the random effects estimator (section “Random Effects Estimator”).

Table 2 POLS estimates of the price-response-function

	POLS (OLS standard errors)		POLS (cluster-robust standard errors)	
	<i>Model 1</i>		<i>Model 2</i>	
	B	SE	B	SE
Constant (β_0)	283.75***	(28.64)	283.75***	(89.09)
Price (β_1)	-1.48***	(0.23)	-1.48 ^{n.s.}	(1.57)
Promo (β_2)	47.37***	(7.07)	47.37***	(12.74)
Location (β_3)	132.36***	(6.89)	132.36***	(41.28)
Floor (β_4)	-3.41 ^{n.s.}	(10.50)	-3.41 ^{n.s.}	(31.64)
Week-fixed effects	Included		Included	
R^2	0.35		0.35	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant

Modeling the Panel Structure

To introduce the idea of explicitly modeling the panel structure, we depict the development of headphone sales volume over time for five selected stores in Fig. 15. To derive Fig. 15, we fit a set of auxiliary regressions. Specifically, for each store, we run a regression of sales volume on a time period variable, for now excluding the independent variables we included in the POLS model. We see that some stores have systematically higher or lower sales volume than other stores. These systematic differences result from unobserved factors, captured in the error term ξ_{it} . We can classify these unobserved factors into two categories: time-varying and time-constant unobserved factors. In contrast to the POLS estimator, panel data estimators use this notion and split the error term into two parts: $\xi_{it} = u_i + e_{it}$. Thereby, u_i refers to unobserved predictors of the dependent variables that pertain to the cluster (i.e. store-level). Consequently, they are time-constant. The term e_{it} refers to unobserved predictors of the dependent variable that are time varying. We call ξ_{it} the composite error term, u_i the cluster-specific component, and e_{it} the idiosyncratic error term. Eq. 8 displays the rewritten price-response function:

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \delta \text{WEEK} + u_i + e_{it} \tag{8}$$

Figure 16 further illustrates the idea of two error components. In a nutshell, the inclusion of the cluster-specific term u_i extends POLS to handling panel data. The

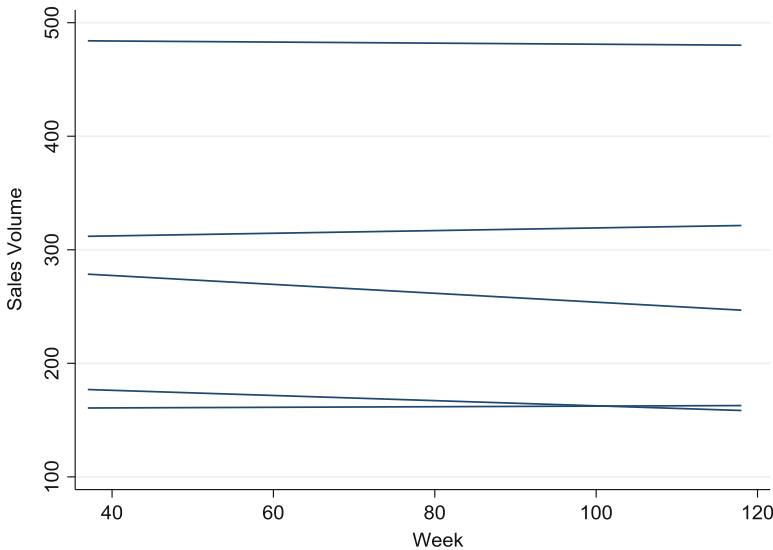


Fig. 15 Differences in sales volume development over time (Illustrated for fitted regressions of five selected stores)

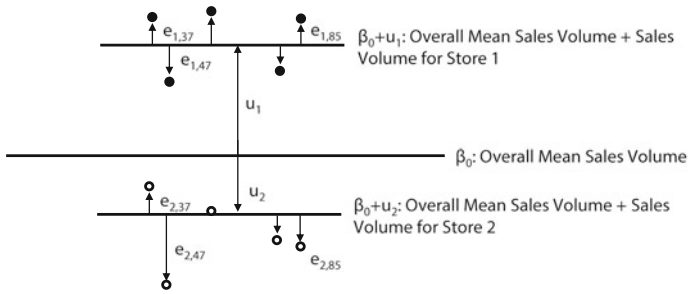


Fig. 16 Decomposition of the composite error term ($\xi_{it} = u_i + e_{it}$)

cluster-specific component captures that the mean sales volume per store differs from the overall mean sales volume across all stores. In Fig. 16, Store 1 displays systematically higher sales volume levels while the sales volume of Store 2 is systematically lower than the overall level of sales volume. While the store-specific component is constant over time, the idiosyncratic error term (e_{it}) indicates that measurements for a focal store (e.g., $e_{1,37}$) can deviate from this store’s mean sales volume (e.g., $\beta_0 + u_1$).

The cluster-specific component captures the effects of all unobserved store-level characteristics (e.g., manager ability). It is referred to as unobserved cluster-level heterogeneity. If the cluster-specific component is positive ($u_i > 0$), the mean composite error term (ξ_{it}) will be positive, leading to larger sales volume levels than predicted by the included independent variables. The reverse holds if the cluster-specific component is negative; in that case, sales levels will be lower than predicted by the included independent variables.

The idiosyncratic error term e_{it} is assumed to have zero population mean and exhibits no correlation both across measurement points and with the independent variables. Depending on the assumptions imposed on u_i , two estimators emerge:

- The fixed-effects estimator, which does not assume uncorrelated unobserved cluster-level heterogeneity: u_i can correlate with the independent variables. Therefore, the fixed effects approach allows consistent estimation even if the investigator omits focal store variables (e.g., manager ability) from the price-response-function. The fixed effects approach estimates the respective cluster-specific effects (u_i).
- The random effects estimator, which assumes uncorrelated cluster-level unobserved heterogeneity: u_i is uncorrelated with the independent variables. Therefore, the random effects approach requires that the investigator includes all focal store variables in the price-response-function. The random-effects approach treats u_i as an unobserved random variable with a particular distribution (e.g., normal distribution). Rather than calculating an estimate for every cluster, it estimates a single variance over the clusters.

Fixed Effects Estimator

The fixed effects estimator accounts for the panel structure by exploiting the within variation in the data. It takes each cluster as its control group and only relies on the cluster’s variation to estimate the model. Before we discuss the fixed effects estimator in more detail, we introduce a variant of it, called the least squares dummy variables (LSDV) regression, that shares the same logic but is implemented in a more straightforward way.

Least squares dummy variable regression. The name of the LSDV approach results from its reliance on a set of dummy variables. Specifically, the LSDV approach adds store dummies – referred to as store-fixed effects – for all but one store to our price-response-function (Eq. 9). The left-out store fixed effect serves as the reference category. The LSDV, thus, is the POLS model (section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)”) with an added vector of store-specific dummy variables.

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \delta \text{WEEK} + \gamma \text{STORE} + e_{it} \quad (9)$$

We do no longer incorporate u_i in the price-response-function since the store-fixed effects account for all observed and unobserved store-level effects in the data. As a consequence, however, we can no longer include the time-constant variables Location_i and Floor_i . These variables correlate perfectly linearly with the store-fixed effects, and thus their effects cannot be estimated. Time-constant variables that interact with time-varying variables (e.g., $\text{Price}_{it} \times \text{Location}_i$), however, could be included as the interaction does not perfectly correlate with the store-fixed effects.

We can estimate the LSDV model with an OLS regression with store-specific dummy variables (`i.storeid`) and, optionally, cluster-robust standard errors (`cluster(storeid)`). Stata will automatically only include $n-1$ store dummy variables. In section “Robust Inference,” we will elaborate on the rationale of using cluster-robust standard errors in addition to employing a panel estimator.

```
reg sales price promo location floor i.week i.storeid, cluster
(storeid)
```

Model 3 in Table 3 shows the LSDV regression results. Since the LSDV regression relies on the OLS estimator, investigators interpret results in the same manner as POLS estimates (section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)”). We want to highlight some important aspects:

1. The effect of price is slightly stronger in the LSDV regression than in the POLS model (LSDV: $\beta_1 = -1.85$, $p < 0.10$; POLS: $\beta_1 = -1.48$, *n.s.*) and statistically different from zero. Unobserved store-level variables in the POLS model may have suppressed the effect of price. The inclusion of store-specific effects in the LSDV model picks up these effects and explains this difference.
2. Reported R^2 values of the LSDV are typically large since the store-fixed effects explain all time-constant variation of sales volume between stores. LSDV’s

Table 3 LSDV and fixed effects estimates of the price-response-function

	LSDV		Fixed effects	
	Model 3		Model 4	
	B	SE	B	SE
Constant (β_0)	357.86***	(45.91)	394.66***	(56.13)
Price (β_1)	-1.85*	(1.00)	-1.85*	(1.00)
Promo (β_2)	45.21***	(8.63)	45.21***	(8.57)
Week-fixed effects	Included		Included	
Store-fixed effects	Included		Controlled	
R ²	0.75		0.24	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; *n.s.* = not significant; cluster-robust standard errors (SE) in parentheses

- R²value (Model 3: R² = 0.75) exceeds the one of the POLS model (Model 1: R² = 0.35).
- In line with our discussion, Stata will automatically drop time-constant variables from the model due to perfect collinearity. Even if investigators had included premium location and floor design to the price-response-function, these variables would not have been estimated.
 - The constant (β_0) represents the store-specific intercept of the left-out store dummy variable (i.e., storeid = 1) at the left-out week (i.e., week = 37).
 - Usually, investigators do not report estimates for the store dummy coefficients. They are included for statistical reasons but typically do not convey substantive insights. We can, however, test whether they are jointly significant, using a conventional F-test on the estimated coefficients (`testparm` command in Stata). If the test statistic is sufficiently high, we can reject the null hypothesis of zero store effects.

Within transformation. The LSDV approach estimates one regression parameter per store. In panel data applications with many clusters, the number of dummy variables will become large, reducing the degrees-of-freedom in the model and, therefore, the estimates' precision. An alternative way to estimate the model is to rely on time-demeaned data, which produces the same (or, at least, very close) estimates. The idea is first to calculate cluster-means per store for each variable. In the next step, investigators center each variable on their cluster-means by subtracting the cluster-mean of each variable from its observed values (for a more technical discussion, please refer to Cameron and Trivedi 2005, pp. 726–729; Greene 2003, pp. 194–196; Verbeek 2017, pp. 387–388). In the case of time-constant variables, the cluster-mean equals the observed variable.

$$\begin{aligned}
 (\text{Sales}_{it} - \overline{\text{Sales}}_i) &= (\beta_0 - \beta_0) + \beta_1(\text{Price}_{it} - \overline{\text{Price}}_i) + \beta_2(\text{Promo}_{it} - \overline{\text{Promo}}_i) \\
 &\quad + \beta_3(\text{Location}_i - \text{Location}_i) + \beta_4(\text{Floor}_i - \text{Floor}_i) + (u_i - u_i) + (e_{it} - \bar{e}_i) \\
 &= \beta_1(\text{Price}_{it} - \overline{\text{Price}}_i) + \beta_2(\text{Promo}_{it} - \overline{\text{Promo}}_i) + (e_{it} - \bar{e}_i)
 \end{aligned}
 \tag{10}$$

Like in the LSDV approach, time-demeaning removes the model's cluster-specific component (u_i), as visible in Eq. 10. For the same reason, all remaining time-constant variables also disappear (i.e., $Location_i$ and $Floor_i$). Since time-demeaning is referred to as within transformation, the fixed effects estimator is also called the within estimator.

To estimate the model, we could manually transform the data and run an OLS regression on the transformed data. However, such manual transformation would lead to calculating the wrong degrees-of-freedom. Importantly, the investigator would need to subtract the number of estimated cluster-means calculated as an intermediate step. Statistical software packages directly correct for this adjustment in the degrees-of-freedom, and hence we recommend relying on these pre-programmed commands. We use Stata's `xtreg` command with the `fe` option, and additionally clustering the standard errors.

```
xtreg sales price promo i.week, fe cluster(storeid)
```

Model 4 in Table 3 displays the results of the fixed effects estimation. The coefficient estimates of $Price_{it}$ and $Promo_{it}$ are identical between the two models. However, the reported R^2 values differ. The reason is that Stata uses a different denominator to compute the R^2 statistic between Model 3 and Model 4. Specifically, while Model 3 is based on the overall variation of sales volume, Model 4 relies only on the within variation of sales volume that enters the R^2 formula (Wooldridge 2016, pp. 437–438).

At this point, it is also worthwhile to review the three different types of R^2 values that are calculated after a fixed effects model. Statistical programs, such as Stata, conventionally report a within- R^2 , a between- R^2 , and an overall- R^2 . All three values provide insights into the model, but the within-value is typically of main interest after a fixed effects estimation. It indicates how much of the variation in the dependent variable within a store is captured by the model. The between- R^2 , correspondingly, describes how much of the variation in the dependent variable between stores is explained by the model. The overall- R^2 is the weighted average of the two.

Finally, the reported constant terms differ between Model 3 and Model 4. Eq. 10 suggests that time-demeaning removes the constant term. However, most statistical software packages do report a constant term by relying on a slightly different within-transformation. Specifically, Stata additionally subtracts each variable's overall mean, the grand mean, from the observed values (Eq. 11).

The estimated constant becomes the average of all store-specific effects: $\hat{\beta}_0 =$

$\sum_{i=1}^n \frac{\hat{u}_i}{n} = \bar{u}$. Thus, the interpretation of the constant in Model 4 differs from Model 3.

$$\begin{aligned} \left(\text{Sales}_{it} - \left(\overline{\text{Sales}_i} - \overline{\overline{\text{Sales}}} \right) \right) &= \beta_0 + \beta_1 \left(\text{Price}_{it} - \left(\overline{\text{Price}_i} - \overline{\overline{\text{Price}}} \right) \right) \\ &+ \beta_2 \left(\text{Promo}_{it} - \left(\overline{\text{Promo}_i} - \overline{\overline{\text{Promo}}} \right) \right) + (e_{it} - (\bar{e}_i - \bar{\bar{e}})) \end{aligned} \quad (11)$$

Random Effects Estimator

The random effects estimator represents another estimator that explicitly considers the panel structure. While the fixed effects estimator controls for the panel structure by removing the model's cluster-specific effects, the random effects estimator directly models the serial correlation stemming from the cluster-specific effects u_i ($\text{corr}(\xi_{it}, \xi_{it-1})$). Under the random effects assumptions, Eq. 12 expresses the serial correlation (Andreß et al. 2013, pp. 77–78 formally derive this equation) as:

$$\text{Corr}(\xi_{it}, \xi_{it-1}) = \frac{\sigma_{u_i}^2}{\sigma_{u_i}^2 + \sigma_{e_{it}}^2} \quad (12)$$

Equation 12 is referred to as the intraclass correlation coefficient (see Misangyi et al. 2006 for an application of the intraclass correlation-coefficient analysis on business segments and firm performance). It relates the cluster-specific variance to the overall variance (sum of variance between clusters and across time). If there were no serial correlation present in the data, Eq. 12 would produce a quantity of zero, and POLS would be an adequate estimator to use. However, while POLS is valid only in the particular case in which serial correlation is zero, the random-effects model is more general and explicitly models the degree of serial correlation.

Researchers can estimate the random effects model with the (feasible) generalized least squares estimator (FGLS). The FGLS estimator is a weighted least square estimator, which attributes more or less weight to a given observation depending on its variance structure (reflecting the relationship between $\sigma_{u_i}^2$ and $\sigma_{e_{it}}^2$). Like the fixed effects estimator, the FGLS estimator applies a quasi-demeaning procedure and subtracts the variable's cluster-mean from the variable's observed values. However, in contrast to the fixed effects estimator, the FGLS estimator only subtracts a fraction (θ , pronounced theta) between 0 and 1 of the cluster-mean from the respective value. Based on the quasi-demeaned data, the OLS estimator can then be applied. Since the FGLS estimator only subtracts a fraction, time-constant variables do not drop out of the model (see Eq. 13). For a more formal discussion of the underlying algebra, we refer to Cameron and Trivedi (2005, pp. 734–736), Greene (2003, pp. 200–205), Verbeek (2017, pp. 391–392), and Wooldridge (2010, Chap. 10).

$$\begin{aligned} (\text{Sales}_{it} - \theta \times \overline{\text{Sales}_i}) &= \beta_0 + \beta_1 (\text{Price}_{it} - \theta \times \overline{\text{Price}_i}) + \beta_2 (\text{Promo}_{it} - \theta \times \overline{\text{Promo}_i}) \\ &+ \beta_3 (\text{Location}_i - \theta \times \overline{\text{Location}_i}) + \beta_4 (\text{Floor}_i - \theta \times \overline{\text{Floor}_i}) \\ &+ (u_i - \theta \times \bar{u}_i) + (e_{it} - \theta \times \bar{e}_i) \end{aligned} \quad (13)$$

Equation 14 shows how to calculate θ . Essentially, θ comprises the relationship between the residual variance ($\sigma_{e_{it}}^2$) and cluster-specific variance ($\sigma_{u_i}^2$), which researchers can estimate with the FGLS estimator (Verbeek 2017, p. 392; Wooldridge 2016, p. 442).

$$\theta = 1 - \sqrt{\frac{\sigma_{e_{it}}^2}{\sigma_{e_{it}}^2 + T \times \sigma_{u_i}^2}} \tag{14}$$

We can obtain the random effects estimator using the Stata `xtreg, re` command, and, if we wish to, cluster the standard errors. To report the fraction of θ that Stata subtracts, we include (`theta`) as an additional option.

```
xtreg sales price promo location floor i.week, re cluster
(storeid) theta
```

Table 4 summarizes the results. Investigators can evaluate the overall model fit with the R^2 statistic and the overall model’s statistical significance with a Wald test (instead of an F-test that cannot account for serial correlation in the error term). As with the fixed effects estimator, Stata reports three types of R^2 . We inspect the overall- R^2 for the random effects model. Researchers can test the significance of individual regression coefficients with a z-statistic that draws on a normal distribution. Model 5 demonstrates that the effect of price on sales is negative and statistically significant ($\beta_1 = -1.82, p < 0.05$). In addition, promotional activities ($\beta_3 = 45.21, p < 0.01$) and stores in more attractive locations realize higher sales volumes ($\beta_3 = 132.79, p < 0.01$). However, stores with a multiple floor design ($\beta_4 = -1.48, n.s.$) do not associate with higher sales volumes.

The reported θ value (`theta`), which is used for the quasi-demeaning (Eq. 13), is 0.92 for our data. Stata will report only one value for θ if the panel is balanced. If the panel is unbalanced, Stata will report multiple θ values, depending on the number of weeks for which the store is observed.

We conclude with two additional comments:

1. The random effects approach treats the store-specific effects as unobservable random variables and not as model parameters (as the fixed effects model does). Still, investigators can obtain estimates for the store-specific intercepts

Table 4 Random effects estimates of the Price-response-function

	Random effects	
	<i>Model 5</i>	
	B	SE
Constant (β_0)	302.46***	(43.78)
Price (β_1)	-1.82**	(0.89)
Promo (β_2)	45.21***	(8.59)
Location (β_3)	132.79***	(41.76)
Floor (β_4)	-1.48 ^{n.s.}	(34.74)
Week-fixed effects	Included	
R^2	0.35	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses; store effects controlled for through random intercepts

(predict u_i , u ; Rabe-Hesketh and Skrondal 2012, p. 107, p. 161, discuss further options to get store-specific error terms).

- As an alternative to the FGLS estimator, investigators can also rely on a maximum likelihood estimator (mle option instead of re) to estimate the random effects model. In general, the two estimators will deliver equivalent results in large samples. We will rely on the maximum likelihood estimator when discussing panel data analysis from a multilevel modeling perspective in the section “[Random Slope Models: A Multilevel Model Approach to Panel Data.](#)”

Relationship between Pooled OLS, Fixed Effects, and Random Effects Estimators

Thus far, we relied on three different estimators (POLS, fixed effects, and random effects estimator) to estimate our price-response-function. Table 5 summarizes the results, introducing the two proper panel data estimators first (Models 6 and 7) and keeping the POLS estimator (Model 8) as a benchmark. We note that the three estimators lead to different results. The relationship among the estimators becomes apparent when we study the quasi-demeaning procedure (e.g., $\text{Sales}_{it} - \theta \times \overline{\text{Sales}_i}$) underlying the random effects estimator in more detail.

Specifically:

- For $\theta = 1$, the random effects estimator becomes the fixed effects estimator.
- For $\theta = 0$, the random effects estimator becomes the POLS estimator.

Table 5 Comparison of fixed effects, random effects, and POLS estimates of the price-response-function

	Fixed effects	Random effects	POLS
	<i>Model 6</i>	<i>Model 7</i>	<i>Model 8</i>
	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	302.46*** (43.78)	283.75*** (89.09)
Price (β_1)	-1.85* (1.00)	-1.82** (0.89)	-1.48 ^{n.s.} (1.57)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	47.37*** (12.74)
Location (β_3)	Omitted	132.79*** (41.76)	132.36*** (41.28)
Floor (β_4)	Omitted	-1.48 (34.74)	-3.41 ^{n.s.} (31.64)
Week-fixed effects	Included	Included	Included
Store-fixed effects	Controlled		
R ²	0.24	0.35	0.35

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses; in random effects model store effects controlled for through random intercepts

While in applied research, θ is unlikely to be exactly 0 or 1, the rearrangement of θ in Eq. 15 helps to understand which determinants affect θ .

$$\theta = 1 - \sqrt{\frac{1}{1 + T \times \left(\frac{\sigma_{u_i}^2}{\sigma_{\varepsilon_{it}}^2}\right)}} \quad (15)$$

First, with an increasing amount of measurement occasions T , the random effects estimator converges to the fixed effects estimator. In our example, we track sales volume over $T = 82$ weeks. The long observation period explains why the random effects estimator produces results close to those of the fixed effects estimator (Stata used $\theta = 0.92$, close to 1). The more measurement points we collect per cluster, the more the time effect dominates the store effect. Notably, the random effect estimator converges to the fixed effects estimator even if omitted time-constant variables correlate with the independent variables.

Second, the more the idiosyncratic error component dominates the store-specific component (i.e., $\frac{\sigma_{u_i}^2}{\sigma_{\varepsilon_{it}}^2}$ becomes small), the less important the store-specific effect (u_i) becomes. As a consequence, the more negligible difference it makes to explicitly model the panel structure, the more results from the random effects estimator resemble those of the POLS estimator. In the extreme case in which the store-specific variance is zero ($\sigma_u^2 = 0$), the POLS, random effects, and fixed effects estimators produce identical results. However, Wooldridge (2016, p. 443) notes that it is more common that $\frac{\sigma_u^2}{\sigma_{\varepsilon}^2}$ is large and that θ will be closer to 1.

Investigators can rely on the Lagrange multiplier test to formally test the null hypothesis of the store-specific effect's variance being zero ($H_0: \sigma_u^2 = 0$), equivalent to testing whether u_i is zero (Rabe-Hesketh and Skrondal 2012). The `xttest0` command in Stata implements the Lagrange multiplier test after the regression command.

```
xtreg sales, re cluster(storeid)
xttest0
```

In our example, the test statistic clearly rejects the null hypothesis ($\chi^2(\text{d.f.} = 1) = 28,030.42$; $p < 0.01$), and we conclude that the store-specific component is different from zero.

What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?

A natural follow-up question is as follows: Which estimator should be selected? To answer this question, we have to recall why the three estimators produce divergent estimates. Although all estimators rely on the same dataset, they exploit variation in the data in different ways. The fixed effects estimator purely uses within variation, whereas the POLS and the random effects estimators rely on both sources of variation.

Using both sources of variation, POLS and random effects estimators can thus exploit more variation (i.e., more information) than the fixed effects estimator. As such, POLS and the random effects estimators are more efficient (i.e., producing smaller standard errors) than the fixed effects estimator, with the random effects estimator being the most efficient (Verbeek 2017). Efficiency is a desirable property. The estimated coefficients from each sample draw deviate from the “true” parameters in the population due to sampling error (also referred to as standard error). As we typically only have one sample to use in our analysis, we are interested in the estimator that produces the smallest standard errors, i.e., the most efficient estimator.

However, as a downside, benefitting from those efficiency gains requires us to impose one additional assumption on the POLS and the random effects estimators: All relevant time-constant variables are included in the model. In other words, we assume that the model fully explains the more information that we use in the estimation. This assumption would be met in our example if we could safely assume that $Location_i$ and $Floor_i$ are the only relevant time-constant variables.

However, we might have overlooked other critical time-constant variables. For instance, managers across stores may differ in their levels of experience and capabilities. More experienced and more capable managers might better anticipate future market developments and set more reasonable prices. Since our observation period covers less than 2 years (weeks 37–118), we can assume that those experiences and capabilities are constant over time. Since we did not measure manager ability, it becomes part of the store-specific error term (u_i). This is problematic if manager ability also impacts realized sales volumes, in which case our price-response-function would suffer from endogeneity in the form of an omitted variables bias.

Figure 17 demonstrates two ways in which omitted manager ability could affect our estimated price-response-functions. In this figure, hallow (Store 1) and black-shaded (Store 2) circles represent observed price and sales volume combinations for two stores. Gray-shaded squares are the corresponding store-specific mean values.

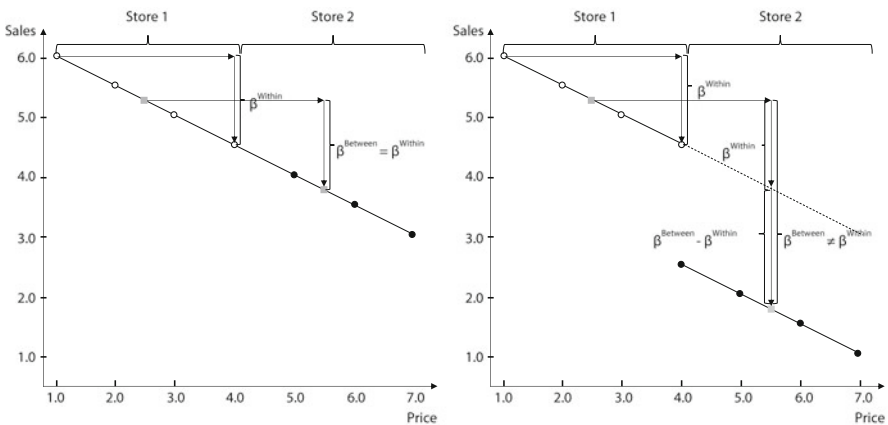


Fig. 17 Comparison of between and within effects

Based on these squares, we show the corresponding regression coefficients (β) predicted based on exploiting within (β^{Within}) versus between variation (β^{Between}). If using within versus between variation leads to the same results, then POLS and random effects estimators should be selected as these are the more efficient estimators. If leveraging within versus between variation leads to different results, then the fixed effect estimator should be chosen as estimators based on between variation (i.e., POLS and random effects) are likely to be biased and inconsistent.

Panel a visualizes a situation in which managers in Store 2 have set higher prices than managers in Store 1. However, the resulting sales volume in Store 2 is at the same level as it would be if Store 1 had set the same price (all observed price-sales-volume combinations lie on the same straight line). Consequently, we retrieve the same information when exploiting either the within or between variation. Thus, in Panel a, we obtain the same regression coefficient, irrespective of the type of variation we exploit ($\beta^{\text{Between}} = \beta^{\text{Within}}$). Such a situation is referred to as uncorrelated unobserved cluster-level heterogeneity, unobserved because we do not observe (or measure) manager ability, and uncorrelated because unobserved ability does not impact price (and realized sales volume). In such a situation, POLS and the random effects estimators are consistent and are more efficient than the fixed effects estimator.

Panel b in Fig. 17 demonstrates a situation in which managers in Store 2 have set higher prices than managers in Store 1, yet this time, the observed level of sales volume dropped below the level that Store 1 would have realized. The reason could be that the managers of Store 2 are less capable of inferring demand and hence price setting. These managers might also create less attractive store environments and make less appealing assortment decisions which lower sales volume. Such a situation is referred to as correlated unobserved cluster-level heterogeneity since, in this situation, the omitted capabilities correlate with the price (and the realized sales volume).

The fixed effects estimator is still consistent in such a situation because it only picks up the within variation. However, the between variation suffers from an omitted variable bias. In picking up the between variation, the random effects estimator assumes that the sales differences between the stores are driven by price only. However, sales are correlated with manager ability, which also impacts price setting (i.e., a situation of correlated unobserved cluster-level heterogeneity). In this case, the estimator is based on an invalid assumption. In other words, it uses more information (i.e., between variation in sales) but uses it under a wrong assumption (i.e., pretending that sales volume is only explained by price while sales volume is also explained by other factors that also correlate with price). As a consequence, the POLS and random effects estimators are not consistent anymore. In Panel b), the POLS and random effects estimators would overestimate the impact of price on sales volume ($|\beta^{\text{Between}}| > |\beta^{\text{Within}}|$).

In the following, we introduce a formal test that assists in choosing between the fixed effects and the random effects estimators. Note that we focus on the comparison between the fixed effects and the random effects estimators. If the random effects estimator is consistent, POLS is consistent, yet when choosing between

POLS and random effects, we recommend the more efficient random effects estimator. However, as we will see later (section “[Summary of the Discussed Estimators and Their Underlying Assumptions](#)”), the random effects estimator requires additional assumptions compared to POLS. For this reason, we also acknowledge that some econometricians suggest performing POLS despite potential efficiency gains from the random effects estimator (Angrist and Pischke 2008, p. 223).

Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator

The trade-off we discussed in the last section raises the natural follow-up question: How do we know when we can rely on the more efficient random effects estimator? Stated differently, how do we know whether the random effects estimator is inconsistent due to omitted time-constant variables (e.g., managerial ability)?

This question requires a formal statistical test. To evaluate whether the random effects estimator is appropriate, we can rely on a Hausman test (Hausman 1978). The basic idea of the Hausman test is that the fixed effects and random effects estimators are consistent under the assumption that the investigator omitted no relevant time-constant variables from the price-response-function. Thus, their estimates should not differ significantly from each other. However, under the assumption that the investigator omitted relevant time-constant variables, only the fixed effects estimator is consistent and the random and fixed effects model estimates should differ significantly (Cameron and Trivedi 2005, p. 717). The Hausman test formally evaluates the null hypothesis of equal fixed effects and random effect estimates ($H_0: \hat{\beta}^{FE} - \hat{\beta}^{RE} = 0$). Investigators can rely on the random effects estimator unless the Hausman test returns a significant test statistic.

For an individual coefficient, the Hausman test can be calculated with the following test statistic (Eq. 16) that follows a χ^2 distribution with 1 degree-of-freedom.

$$\chi^2 = \left(\frac{(\hat{\beta}^{FE} - \hat{\beta}^{RE}) - 0}{SE \begin{pmatrix} \hat{\beta}^{FE} & \hat{\beta}^{RE} \\ \hat{\beta} & -\hat{\beta} \end{pmatrix}} \right)^2 = \left(\frac{(\hat{\beta}^{FE} - \hat{\beta}^{RE}) - 0}{\sqrt{SE_{\hat{\beta}^{FE}}^2 - SE_{\hat{\beta}^{RE}}^2}} \right)^2 \tag{16}$$

If we included $Price_{it}$ as the only independent variable (e.g., xtreg sales price, re) in our price-response-function, we would obtain the random effects ($\hat{\beta}^{RE} = -1.82$; $SE_{\hat{\beta}^{RE}} = 0.53$) and fixed effects ($\hat{\beta}^{FE} = -1.84$; $SE_{\hat{\beta}^{FE}} = 0.52$) estimates that result in a nonsignificant test statistic ($\chi^2(\text{d.f.} = 1) = 0.01$; *n.s.*), favoring the random effects estimator. This result suggests that we have no reason to be concerned about correlated unobserved cluster-specific factors, such as manager ability.

For more common models that include more than one independent variable, we use a generalization of Eq. 16. Based on matrix algebra, we derive the general test statistic of the Hausman test (Eq. 17; Greene 2003, pp. 208–209) as:

$$\chi^2 = \left(\hat{\beta}^{FE} - \hat{\beta}^{RE} \right)' \times \left(\hat{\psi}^{FE} - \hat{\psi}^{RE} \right)^{-1} \times \left(\hat{\beta}^{FE} - \hat{\beta}^{RE} \right) \quad (17)$$

Thereby, $\hat{\beta}^{FE}$ and $\hat{\beta}^{RE}$ include all coefficient estimates of the fixed effects and the random effects model. The matrices $\hat{\psi}^{FE}$ and $\hat{\psi}^{RE}$ include the corresponding estimated variances and covariances of the estimates. The test-statistic follows an asymptotic χ^2 distribution with p degrees-of-freedom, with p being the number of coefficients tested.

The respective Stata syntax easily allows estimating the Hausman test statistic (`hausman`) after running the respective regressions. Note that we choose the `sigmamore` option to prevent Stata from considering the vector of week-fixed effects as tested coefficients and thereby inflate the reported degrees-of-freedom (degrees-of-freedom should be 2, for two tested coefficients, and not 83).

```
xtreg sales price promo i.week, fe
estimates store FE
xtreg sales price promo location floor i.week, re
estimates store RE
hausman FE RE, sigmamore
```

For our data, we obtain an insignificant test statistic ($\chi^2(\text{d.f.} = 2) = 0.09$; *n.s.*) and conclude that we can trust the random effects results. This result is not surprising since fixed and random effects estimates hardly differ in our example.

Finally, we end our discussion highlighting three important aspects:

1. The standard Hausman test is not valid when the investigator uses cluster-robust standard errors. In section “[Alternative Hausman Test](#),” we will introduce a fully robust Hausman test.
2. Equation 16 serves to demonstrate when the Hausman test is not likely to yield a significant result. First, the test statistic tends to be insignificant when the numerator is small. In this case, the estimates for the $Price_{it}$ coefficients do not differ significantly between the fixed effects and random effects estimators; that is, it does not matter which estimator we use. Second, the Hausman test also becomes insignificant when the denominator is large. If the fixed effects estimator displays a large standard error (e.g., the variation of prices over time is low), the Hausman test likely yields an insignificant result. Third, measurement error can provoke an attenuation bias in the fixed effects estimator, which describes an estimate’s bias converging to zero and underestimating the true value. As a result, fixed effects estimates could be smaller than the estimates from the random effects model, even if no relevant time-constant variables were omitted (sections “[Summary of the Discussed Estimators and Their Underlying Assumptions](#)” and “[Addressing Measurement Error with Structural Equation Modeling Based on Panel Data](#)”). Verbeek (2017, p. 395) notes in this regard:

Although the Hausman test is commonly used as a tool to decide between the random effects and the fixed effects estimators, it should be used with caution. Rejection should not automatically be interpreted as evidence that the fixed effects model is appropriate. Conversely, if the Hausman test does not reject it is not necessarily the case that the random-effects model should be preferred.

3. Even though the random effects estimator is biased and inconsistent if important time-constant variables are omitted, this bias is attenuated by the factor $(1-\theta)$ (see the quasi-demeaning in section “[Random Effects Estimator](#)”). As a consequence, the bias becomes increasingly less severe the closer θ approaches to 1.

Interpret and Present Results

As the last step, we economically interpret our findings. First, marketing managers are typically interested in the price-elasticity concept: the percentage change in sales volume when there is a 1% increase in price. Formally, the price-elasticity is defined as $\varepsilon = \frac{\partial \text{Sales}}{\partial \text{Price}} \times \frac{\text{Price}}{\text{Sales}}$ where *Price* is price and *Sales* is sales volume. Thereby, the first part is simply the derivative of our price-response-function ($\frac{\partial \text{Sales}}{\partial \text{Price}} = \beta_{\text{Price}}$). For *Price* and *Sales*, we use their respective sample means. Thus, for our example, the price elasticity is -0.36 ($\varepsilon = -1.82 \times \frac{55.38}{280.19}$), a relatively low value when compared to results from a meta-analysis reporting an average price elasticity of -2.62 (Bijmolt et al. 2005). Economically, this finding suggests that marketing managers can expect that a 1% price increase only lowers sales volume by 0.36%. Thus, consumers hardly react to price changes of the newly introduced headphone. Marketing managers, therefore, are likely tempted to increase the price of the headphone.

Based on our price-response-function, managers can also determine the revenue-optimal price. In our case, this price would be 79.95€ ($p^* = \frac{\beta_{\text{Constant}}}{2\beta_{\text{Price}}} = \frac{291}{2 \times 1.82}$), which would imply a price increase of 24.57€ compared to the current average price. Note that for this analysis, we used the intercept ($\beta_{\text{Constant}} = 291$) of a random effects model without week-fixed effects (xtreg sales price promo location floor, re); otherwise, the constant term would only be correct for the left-out week (Table 4: $\beta_{\text{Constant}} = 302.46$).

Additional Methods in Panel Data Analysis

Robust Inference

Obtaining correct standard errors of estimators is complicated for panel data since these data are likely to suffer from serial correlation, as mentioned earlier, as well as heteroskedasticity (Cameron and Trivedi 2005). Petersen (2009) compares the performance of different standard errors in panel datasets and, in general,

recommends the use of cluster-robust standard errors for micropanel datasets (large n and small T). In the spirit of Petersen's finding, Cameron and Trivedi (2005, p. 725) also recommend to "base inference on [cluster-] robust standard errors that do not require specifying a model for the error correlation." Marketing researchers commonly follow this recommendation (e.g., Bayer et al. 2020, Borah and Tellis 2014; Warren and Sorescu 2017).

Cluster-robust standard errors are calculated based on the observed distribution of residuals from the model. They are robust in the sense that they consider the clustered (or nested) data structure of panel data and assume that observations are independent between clusters but not necessarily within clusters. In other words, cluster-robust standard errors consider each store as a cluster with observations over time, "and arbitrary correlation—serial correlation—and changing variances are allowed within each cluster" (Wooldridge 2016, p. 433). Cluster-robust standard errors are beneficial because they also control for potential heteroskedasticity, a second challenge in estimating standard errors in panel data. Wooldridge (2010, Chap. 10), Greene (2003, pp. 211–213), and Cameron and Trivedi (2005, Chap. 21.2.3) formally derive cluster-robust standard errors in the context of panel data.

We have already discussed the importance of cluster-robust standard errors for POLS (section "[Pooled OLS Estimator: Ignoring the Panel Structure](#)") since serial correlation likely leads to an underestimation of common OLS standard errors. We have also requested cluster-robust standard errors for the fixed effects and random effects models when estimating the price-response-functions. This choice seems reasonable as, in general, accounting for fixed effects or random effects only lowers serial correlation but does not eliminate it. Thus, our general recommendation is to rely on cluster-robust standard errors for panel data analysis.

Combining the Fixed Effects and Random Effects Estimators

Our prior discussion suggested that investigators need to choose between the fixed effects and random effects estimators. However, the literature also offers a combined approach that seeks to leverage the advantages and alleviates the disadvantages of the two estimators. Specifically, it allows including time-constant variables and provides an alternative Hausman test valid for cluster-robust standard errors. First, before we discuss the combined approach, we will introduce an additional estimator – the between effects estimator – which is necessary to understand the combined approach (section "[Between Effects Estimator](#)"). Second, we will outline the combined approach (section "[Combined Approach](#)"). Third, we will outline the alternative (fully robust) Hausman test that researchers can perform (section "[Alternative Hausman Test](#)"). Fourth, we will outline why the combined approach allows the consistent estimation of time-varying variables (section "[Understanding How the Combined Approach Allows Consistent Estimation of Time-Varying Variables](#)").

Between Effects Estimator

Since marketing research rarely applies the between effects estimator (cf. Nath and Mahajan 2008 for an exception, Table 2, Model 3), we will only shortly overview the estimator. The between effects estimator relies exclusively on the between variation in the data and discards all time-series information. As such, the estimator computes the variables' average values per cluster (e.g., mean price: $\frac{1}{T} \sum_{t=1}^T \text{Price}_{it} = \overline{\text{Price}_i}$; effective sample size = $n = 19$ stores) and runs an OLS regression on these averages (Eq. 18). Thus, the between effects estimator exploits the information that the fixed effects estimator does not use. In contrast, the fixed effects estimator only exploits variance over time (within store variation; section “[What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?](#)”), for instance, in terms of price and sales volume as in our data example (e.g., $\text{Price}_{it} - \overline{\text{Price}_i}$; effective sample size = $n \times T = 19 \times 82 = 1,558$ observations).

$$\overline{\text{Sales}_i} = \beta_0 + \beta_1 \overline{\text{Price}_i} + \beta_2 \overline{\text{Promo}_i} + \beta_3 \overline{\text{Location}_i} + \beta_4 \overline{\text{Floor}_i} + u_i + \bar{\epsilon}_i \quad (18)$$

Equation 18 displays the price-response-function that we will estimate with the between effects estimator. In Stata, investigators can request the between effects estimator with the `be` option after `xtreg`. Since we rely on balanced data, week-fixed effects are the same for all companies and are automatically omitted.

```
xtreg sales price promo location floor, be
```

Model 12 in Table 6 displays the results of the between effects estimator and compares those results to the fixed effects (Model 9), the random effects (Model 10), and the POLS estimators (Model 11). The results of the between effects estimator are not the focus of our discussion. Instead, we use them to extend our discussion from section “[What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?](#)” and provide further insights into the random effects and POLS estimators.

We see that the random effects (Model 10: $\beta_1 = -1.82, p < 0.05$) and the POLS (Model 11: $\beta_1 = -1.48, n.s.$) estimates of price on sales volume lie in between the fixed effects (Model 9: $\beta_1 = -1.85, p < 0.10$) and between effects estimates (Model 12: $\beta_1 = -0.87, n.s.$). The random effects estimator is closer to the fixed estimator, and the POLS estimator is closer to the between estimator. This general pattern directly follows from our discussion in section “[What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?](#)”: The POLS and the random effects estimators represent weighted compromises between the fixed effects and the between effects estimators. The random effects estimator is the more efficient one.

In the following, we will demonstrate how we can leverage this idea on the between effects estimator to derive a combined model that allows consistent estimates of time-varying variables (as the fixed effects estimator does) and allows the inclusion of time-constant variables (as the random effects estimator does).

Table 6 Comparison of fixed effects, random effects, POLS, and between effects estimates of the price-response-function

	Fixed effects	Random effects	POLS	Between effects
	<i>Model 9</i>	<i>Model 10</i>	<i>Model 11</i>	<i>Model 12</i>
	B (SE)	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	302.46*** (43.78)	283.75*** (89.09)	208.34 ^{n.s.} (245.05)
Price (β_1)	-1.85* (1.00)	-1.82** (0.89)	-1.48 ^{n.s.} (1.57)	-0.87 ^{n.s.} (3.18)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	47.37*** (12.74)	119.00 ^{n.s.} (318.37)
Location (β_3)	Omitted	132.79*** (41.76)	132.36*** (41.28)	128.54 ^{n.s.} (57.72)
Floor (β_4)	Omitted	-1.48 ^{n.s.} (34.74)	-3.41 ^{n.s.} (31.64)	-3.76 ^{n.s.} (84.30)
Week-fixed effects	Included	Included	Included	Omitted
Store-fixed effects	Controlled			
R ²	0.24	0.35	0.35	0.26

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; *n.s.* = not significant; cluster-robust standard errors (SE) in parentheses; in random effects model store effects controlled for through random intercepts

Combined Approach

The combined approach relies on the random effects estimator but adds the cluster-means of the time-varying variables (i.e., store means for $Price_{it}$ and $Promo_{it}$) to the model, just as the between effects estimator does. Eqs. 19 and 20 demonstrate two options of how investigators can employ a combined approach.

$$Sales_{it} = \beta_0 + \beta_1 Price_{it} + \beta_2 Promo_{it} + \beta_3 Location_i + \beta_4 Floor_i + \beta_5 \overline{Price}_i + \beta_6 \overline{Promo}_i + u_i + e_{it} \tag{19}$$

$$Sales_{it} = \beta_0 + \beta_1 (Price_{it} - \overline{Price}_i) + \beta_2 (Promo_{it} - \overline{Promo}_i) + \beta_3 Location_i + \beta_4 Floor_i + \beta_5 \overline{Price}_i + \beta_6 \overline{Promo}_i + u_i + e_{it} \tag{20}$$

Including cluster-means of time-varying variables is equivalent to demeaning the raw data. However, given that we apply this procedure only to time-varying variables, we can still retain the time-constant variables $Location_i$ and $Floor_i$. If we had an unbalanced panel dataset, we would also need to include the time average of any time-varying variable. We provide the corresponding Stata syntax in the do-file and report results in Table 7.

In Table 7, Model 14 relies on Eq. 19 and Model 15 on Eq. 20. In both models, the estimated price coefficients (Model 2: $\beta_1 = -1.85, p < 0.10$; Model 3: $\beta_1 = -1.85, p < 0.10$) are identical to the corresponding fixed effects estimate (Model 1: $\beta_1 = -1.85, p < 0.10$). The interpretation of the cluster-mean variables of $Price_{it}$

and $Promo_{it}$, however, differs between Model 14 and Model 15. The cluster-means in Model 15 (e.g., $\beta_5 = -0.87, n.s.$) directly replicate the between effect estimator (Model 16: $\beta_1 = -0.87, n.s.$). Model 14 tests the difference between the fixed effects and the between effects results (e.g., $\beta_5^{Model\ 2} = \hat{\beta}_1^{Model\ 4} - \hat{\beta}_1^{Model\ 1} = -0.87 + 1.85 = 0.98$).

Alternative Hausman Test

The interpretation of Model 14’s cluster-means (Table 7) allows an alternative way to test the appropriateness of the random effects estimator. Both the between effects and the fixed effects estimators are consistent when the investigator did not omit any relevant time-constant independent variables. Significant cluster-means from Model 14 (i.e., significant differences between the fixed and between effect estimates) imply that relevant time-constant variables are omitted. Jointly testing both cluster-means from Model 14 via a Wald test (H_0 : All cluster-mean values are zero; H_0 : $\overline{Price}_i = \overline{Promo}_i = 0$), thus, represents an alternative to the Hausman test from section “Hausman Test: Selecting Between the Fixed Effects and the Random

Table 7 Combined approach estimates of the price-response-function

	Fixed effects	Combined approach (Eq. 19)	Combined approach (Eq. 20)	Between effects
	Model 13	Model 14	Model 15	Model 16
	B (SE)	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	232.82 ^{n.s.} (197.17)	232.82 ^{n.s.} (197.17)	208.34 ^{n.s.} (245.05)
Price (β_1)	-1.85* (1.00)	-1.85* (1.00)	-1.85* (1.00)	-0.87 ^{n.s.} (3.18)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	45.21*** (8.59)	119.00 ^{n.s.} (318.37)
Location (β_3)	Omitted	128.54*** (44.85)	128.54*** (44.85)	128.54 ^{n.s.} (57.72)
Floor (β_4)	Omitted	-3.76 ^{n.s.} (34.50)	-3.76 ^{n.s.} (34.50)	-3.76 ^{n.s.} (84.30)
Mean price (β_5)		.98 ^{n.s.} (2.76)	-0.87 ^{n.s.} (2.65)	
Mean promo (β_6)		73.79 ^{n.s.} (280.36)	119.00 ^{n.s.} (280.56)	
Week-fixed effects	Included	Included	Included	
Store-fixed effects	Controlled			
R ²	0.24	0.35	0.35	0.26

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses

Effects Estimator.” The test-statistic follows a χ^2 distribution with p degrees-of-freedom, with p being of the number of cluster-means tested. In line with our prior interpretation of the Hausman test, investigators rely on the random effects estimator (i.e., we can drop the cluster-means from Eq. 19) unless the Wald test returns a significant test statistic.

```
test price_mn promo_mn
```

Hausman (1978, p. 1263) and Hausman and Taylor (1981, p. 1382) showed that both tests are asymptotically equivalent (Wooldridge 2010, p. 332). Thus, to test whether omitted time-constant company variables correlate with the independent variables, we can either compare the fixed effects and random effects estimates (which we did in section “[Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach](#)”) or the fixed effects with the between effects estimates which we do in this section. Wooldridge (2010) recommends the latter (regression-based) version using cluster-robust standard errors. For our example, we cannot reject the null hypothesis of equal between effects and fixed effects ($\chi^2(\text{d.f.} = 2) = 0.13$; $p = 0.94$), replicating the substantive conclusion of section “[Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#).” Thus, we can rely on the random effects estimator.

Understanding How the Combined Approach Allows Consistent Estimation of Time-Varying Variables

The random effects estimator’s core challenge is the assumption that the independent variables are uncorrelated with the unobservable time-constant cluster effects. In the following, we build on Wooldridge (2016, pp. 445–446) to explain why including cluster-means leads to consistent estimates of time-varying independent variables. Let us assume that manager ability was indeed a driver of sales volume, yielding the specification in Eq. 21. For the sake of simplicity, we will assume that we do not need to account for promotion, premium location, and floor design.

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Manager_Ability}_i + u_i + e_{it} \quad (21)$$

Since we have not collected data on manager ability, we would falsely specify the price-response-function, as in Eq. 22:

$$\text{Sales}_{it} = \beta'_0 + \beta'_1 \text{Price}_{it} + u'_i + e'_{it} \quad (22)$$

The omitted manager ability becomes part of the store-specific error term ($u'_i = \beta_2 \times \text{Manager_Ability}_i + u_i$). Since manager ability is time-constant in our example, it can only correlate with any time-constant effect in Price_{it} (i.e., $\text{Manager_Ability}_i = \delta_0 + \delta_1 \times \overline{\text{Price}_i} + \phi_i$). If the cluster-mean of price displays a nonzero relationship with manager ability ($\delta_1 \neq 0$), Price_{it} becomes endogenous, i.e., price correlates with the store-specific error term ($\text{corr}(u_i, \text{Price}_{it}) \neq 0$). However, by including the cluster-mean of Price_{it} , the correlation of Price_{it} and the

store-specific error term u_i disappears ($\beta_0'' = (\beta_0' + \beta_2 \times \delta_0)$, $\beta_2'' = (\beta_2 \times \delta_1)$, and $u_i'' = (\beta_2 \times \phi_1 + u_i)$).

$$\begin{aligned} \text{Sales}_{it} &= \beta_0' + \beta_1' \text{Price}_{it} + (\beta_2 (\delta_0 + \delta_1 \overline{\text{Price}}_i + \phi_i) + u_i) + e_{it}' \\ &= \beta_0'' + \beta_1' \text{Price}_{it} + \beta_2'' \overline{\text{Price}}_i + u_i'' + e_{it}' \end{aligned} \quad (23)$$

The Price_{it} coefficient estimate (Eq. 23) will be the same as in the fixed effects model. The intuition behind this result is that while the final regression model (Eq. 23) still omits manager ability, it accounts for unobserved correlated ability effects by including the cluster-mean of price. The included cluster-mean of the price typically has no substantive interpretation but serves to control for an omitted variable bias.

Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach

The combined approach (section “[Combining the Fixed Effects and Random Effects Estimators](#)”) allows including time-constant variables and allows that the Price_{it} and Promo_{it} variables correlate with the store-specific error component u_i (like the fixed effects estimator). However, the combined model still requires that time-constant variables (e.g., Location_i and Floor_i) and the cluster-specific error component u_i are uncorrelated. If this assumption does not hold, the estimates of time-constant variables are not consistent.

Hausman and Taylor (1981) propose a method for obtaining consistent estimates of time-constant variables. Essentially, the Hausman-Taylor approach treats variables differently, depending on whether they are time-constant, time-varying, and correlated or uncorrelated with the cluster-specific component u_i . Specifically, the approach discriminates between time-varying endogenous, time-varying exogenous variables, as well as time-constant endogenous and time-constant exogenous variables.

For instance, we may have reason to believe that omitted time-constant variables could affect Price_{it} and Floor_i (i.e., these are endogenous variables). Knowing that higher-level company executives and not the individual store managers are involved in promotion timing and location decisions, we are not worried that omitted time-constant store variables could impact Promo_{it} and Location_i (i.e., these are exogenous variables). Please note that these rationales are only exemplary and require more careful theoretical and empirical justification.

The Hausman-Taylor approach’s idea is now to derive so-called panel-internal instrumental variables for the endogenous variables Price_{it} and Floor_i . Panel-internal instruments imply that we can use simple transformations of variables that are already included in the price-response-function as instruments. Thus, panel data offer the advantage that investigators do not have to collect external instrumental variables, which are often not readily available.

Specifically, in the Hausman-Taylor approach, demeaned variables (known from the fixed effects estimation) serve as instruments for time-varying endogenous variables (e.g., $(Price_{it} - \overline{Price}_i)$ serves as an instrument for $Price_{it}$). Cluster means of time-varying variables serve as instruments for time-constant endogenous variables (e.g., \overline{Promo}_i serves as an instrument for $Floor_i$). We want to confirm that both instrumental variables are exogenous ($corr(Price_{it} - \overline{Price}_i, u_i) = 0$ and $corr(\overline{Promo}_i, u_i) = 0$). Importantly, the Hausman-Taylor approach requires at least as many time-varying exogenous variables as time-constant endogenous variables for identification. If we had additional time-varying exogenous variables, we could evaluate the strength of the selected instrumental variables using the `xtoverid` command.

We can obtain estimates for the Hausman-Taylor approach with the `xthtaylor` command in which we specify the endogenous variables (both time-varying and time-constant) with the `endog` option; the other variables are considered exogenous. Model 19 in Table 8 displays the results. We see that the Hausman-Taylor approach results in different estimates for some of the variables, including $Floor_i$ and $Location_i$, compared to the combined approach.

```
xthtaylor sales price promo location floor w2-w82, endog(price
floor) vce(cluster storeid)
```

The Hausman-Taylor approach finds initial application in marketing research. For instance, Boulding and Christen (2003, 2008) employ this approach in the context of

Table 8 Hausman-Taylor approach estimates of the price-response-function

	Fixed effects <i>Model 17</i>	Combined approach (Eq. 19) <i>Model 18</i>	Hausman-Taylor <i>Model 19</i>
	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	232.82 n.s. (197.17)	295.32 ^{n.s.} (306.82)
Price (β_1)	-1.85* (1.00)	-1.85* (1.00)	-1.85* (1.00)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	45.21*** (8.58)
Location (β_3)	Omitted	128.54*** (44.85)	141.43 ^{n.s.} (338.88)
Floor (β_4)	Omitted	-3.76 n.s. (34.50)	24.49 ^{n.s.} (1008.37)
Mean price (β_5)		0.98 n.s. (2.76)	
Mean promo (β_6)		73.79 n.s. (280.36)	
Week-fixed effects	Included	Included	Included
Store-fixed effects	Controlled		
R ²	0.24	0.35	0.34

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses

new product introduction strategies. For further applications and their implied panel-internal instruments, please consult Butt et al. (2018), Germann et al. (2015), Ho-Dac et al. (2013), Rao et al. (2004), and Steenkamp and Geyskens (2014).

Summary of the Discussed Estimators and Their Underlying Assumptions

We have discussed a range of estimators to estimate the price-response-function. The estimators produce different results since they exploit the panel structure in different ways: The between effects estimator uses the between variation in the data while the fixed effects estimator exploits the within variation in the data. Finally, the POLS and the random effects estimators exploit both types of variation, with the latter being more efficient. Notably, if the random effects model's key assumption is met, all potential estimators are consistent, and the random effects estimator is the most efficient estimator. However, if the cluster-specific effects u_i are correlated with the independent variables, the random effects model's key assumption is not met. In this case, only the fixed effects estimator is consistent; the Hausman test (sections "[Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#)" and "[Alternative Hausman Test](#)") formally tests this assumption.

Given the divergent results, investigators should understand the different assumptions underlying the estimators (called model-identifying assumptions) and justify their choices in applied research. We review those assumptions in Table 9, which we classify into A, B, and C assumptions according to Von Auer (2013). The A-assumptions address the functional specification of the regression model. B-assumptions focus on the error term structure, and C-assumptions relate to individual variables of the model. The following sources provide a more in-depth discussion of these assumptions: Kennedy (2008), Greene (2003), Skiera et al. (chapter ► "[Regression Analysis](#)"), Verbeek (2017), Wooldridge (2010, 2016).

Regarding the A-assumptions, all estimators require that the investigator includes all relevant variables in the model (assumption A1). We call such an approach a rich data model. For the POLS and the random effects estimator, this implies including all relevant time-constant and time-varying control variables. The fixed effects estimator only requires that the investigator includes all relevant time-varying control variables in the model.

Assumption A2 relates to the relationship between independent and dependent variables, which we assume is linear. Skiera et al. (chapter ► "[Regression Analysis](#)") discuss different data-transformations appropriate for linearizing curvilinear relationships, and those transformations equally apply to the panel data context.

Assumption A3 involves the effect of independent variables on the dependent variables to be constant. If this is not the case, investigators could include interaction terms to perform a moderated regression analysis (e.g., Vomberg et al. 2015). By including interaction terms, investigators effectively model slope heterogeneity. While the inclusion of interaction terms requires that researchers measure the

Table 9 Assumptions of the different estimators

	POLS estimator	Fixed effects estimator	Random effects estimator
A1	No relevant time-varying and time-constant variables are missing	No relevant time-varying variables are missing	No relevant time-varying and time-constant variables are missing
A2	True relationship between independent and dependent variable is linear		
A3	Estimated parameters are constant over all observations		
B1	The expected value of the idiosyncratic error term is zero: $E(\xi_{it} \mathbf{X}, \mathbf{Z}) = 0$	The expected value of the idiosyncratic error term is zero: $E(e_{it} \mathbf{X}) = 0$	The expected value of the idiosyncratic error term is zero: $E(e_{it} \mathbf{X}, \mathbf{Z}, u_i) = 0$ The intercept captures the expected value of the unit-specific error term: $E(u_i \mathbf{X}, \mathbf{Z}) = \beta_0$
B2	Homoskedasticity: $\text{Var}(\xi_{it} \mathbf{X}, \mathbf{Z}) = \sigma^2$	Homoskedasticity: $\text{Var}(e_{it} \mathbf{X}) = \sigma_e^2$	Homoskedasticity: $\text{Var}(e_{it} \mathbf{X}, \mathbf{Z}, u_i) = \sigma_e^2$ $\text{Var}(u_i \mathbf{X}, \mathbf{Z}) = \sigma_u^2$
B3	No serial correlation: $\text{Cov}(\xi_{it}, \xi_{is} \mathbf{X}, \mathbf{Z}) = 0$	No serial correlation: $\text{Cov}(e_{it}, e_{is} \mathbf{X}) = 0$	No serial correlation: $\text{Cov}(e_{it}, e_{is} \mathbf{X}, \mathbf{Z}, u_i) = 0$
B4	Error terms are normally distributed: $\xi_{it} \sim N(0, \sigma^2)$	Idiosyncratic error terms are normally distributed: $e_{it} \sim N(0, \sigma_e^2)$	Idiosyncratic and cluster-specific error terms are normally distributed: $e_{it} \sim N(0, \sigma_e^2)$ $u_i \sim N(\text{constant}, \sigma_u^2)$
C1	Error terms are uncorrelated with independent variables: $\text{Cov}(\xi_{it}, \mathbf{X}) = 0$; $\text{Cov}(\xi_{it}, \mathbf{Z}) = 0$	Idiosyncratic error terms are uncorrelated with independent variables: $\text{Cov}(e_{it}, \mathbf{X} - \bar{\mathbf{X}}) = 0$	Idiosyncratic and cluster-specific error terms are uncorrelated with independent variables: $\text{Cov}(e_{it}, \mathbf{X}) = 0$; $\text{Cov}(e_{it}, \mathbf{Z}) = 0$ $\text{Cov}(u_i, \mathbf{X}) = 0$; $\text{Cov}(u_i, \mathbf{Z}) = 0$
C2	No multicollinearity		
C3	Measurement error-free		

Notes: Based on a stylized regression of $y_{it} = \beta_0 + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + \beta_j z_{ji} + \xi_{it}$ with cluster i in time period t . ξ_{it} is the composite error term, u_i is the cluster-specific error component, and e_{it} is the idiosyncratic error term. x_{it} denotes time-varying variables, and z_i indicates time-constant variables. \mathbf{X} represents the vector of time-varying variables x_{it} , and \mathbf{Z} denotes the vector of time-constant variables z_i

respective moderating variables, in section “[Random Slope Models: A Multilevel Model Approach to Panel Data](#),” we will show that panel data methods also model slope heterogeneity without measured moderating variables.

Regarding the B-assumptions, panel data is especially prone to violate assumptions about the error term distribution given the clusters’ dependency. In this regard, we emphasized the need to rely on cluster-robust standard errors if the investigator decides to employ POLS (section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)”). We also recommended their usage for the fixed effects and random effects estimators (section “[Robust Inference](#)”), as they help account for

serial correlation (violation of assumption B3) and heteroskedasticity (violation of assumption B2).

Regarding assumption C1, the POLS estimator assumes that the composite error term does not correlate with the independent variables (exogeneity). The fixed effects estimator requires that the idiosyncratic error term is uncorrelated with the independent variables. Note that because the cluster-specific effects omit all time-constant independent variables from the model, the exogeneity assumption applies to time-varying independent variables only. The exogeneity assumption is also made for the idiosyncratic error term for the random effects estimator, which is assumed to be uncorrelated with the independent variables. Assumption C2 requires that all variables display unique variation and, hence, are not perfectly correlated (no perfect multicollinearity). For the POLS and the random effects estimators, this is fulfilled if variables have unique variance over time, between clusters, or both. For the fixed effects estimator, this assumption requires that variables display unique variation over time. Standard errors of variables with little variation over time will become large and reduce statistical power. This assumption also explains why investigators cannot add time-constant variables in the fixed effects approach because they perfectly correlate with the cluster-specific effect.

In this regard, we emphasize three further comments:

1. Investigators should not decide between the fixed effects and the random effects approach based on whether they are interested in the effect of time-constant variables. If the effect of time-constant variables is of interest to investigators, they may employ the combined approach (section “[Combining the Fixed Effects and Random Effects Estimators](#)”) or the Hausman-Taylor approach (section “[Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach](#)”).
2. The fixed effects estimator’s inefficiency, which results from little within variation, might favor random effects estimation. For instance, Wolters et al. (2020) justify a random effects over a fixed effects specification by noting a lack of sufficient variation in their focal variable over time. Warren and Sorescu (2017) suggest a random effects approach since their unbalanced panel dataset contains several clusters with only one observation.
3. Moreover, we want to acknowledge recent calls for a more balanced view between bias and efficiency when deciding between the fixed effects and random effects estimator. For instance, Kummer and Schulte (2019) mention little within variation as a limitation to their fixed effects approach. Additionally, Andreß et al. (2013, p. 173) state the following:

Inefficiency of the [fixed effects] estimator is a particular problem if the within-unit variance is low and variables hardly change over time. Since you are never in the lucky situation of statistical theory, which assumes repeated sampling, your single sample may provide you with estimates quite different from the true population parameters. In that case, the fact that fixed effects are unbiased (i.e., correct on average) is no comfort for you. Hence, more research is needed that provides a more balanced view of both estimators that takes into account both unbiasedness and efficiency.

Finally, all estimators require that the variables are measured without error (assumption C3) as measurement error represents another form of endogeneity that can lead to violation of assumption C1 (chapter ► “Crafting Survey Research: A Systematic Process for Conducting Survey Research” by Vomberg and Klarmann). In section “Addressing Measurement Error with Structural Equation Modeling Based on Panel Data,” we will demonstrate how investigators can leverage panel data to reduce measurement error concerns. Notably, due to the within-transformation, the fixed effects estimator is particularly susceptible to attenuation bias from measurement error (Angrist and Pischke 2008; Griliches and Hausman 1986; Wooldridge 2016, p. 440).

Modeling a Price-Response-Function in Differences

The discussion so far has focused on modeling the price-response-function in levels. For the sake of completeness, we briefly discuss an additional estimator for the analysis of panel data: the first difference estimator. The first difference estimator takes the first difference of all variables (see Eq. 24) and then performs an OLS regression on the transformed variables (for a more technical discussion, please refer to Cameron and Trivedi 2005, pp. 729–731).

Equation 24 reveals that unobserved time-constant variables disappear after taking the first differences since they do not change over time. Like the fixed effects estimator, the first difference estimator thereby controls for an omitted variable bias stemming from unobserved time-constant variables. As a consequence, the first difference estimator does not allow to include time-constant variables.

Investigators can manually create the first differences of the variables. Alternatively, Stata automatically creates first differences when placing a difference operator (D.) in front of the respective variable. In line with our prior discussion, we also recommend that researchers should rely on cluster-robust standard errors for first difference models.

$$\begin{aligned}
 (\text{Sales}_{it} - \text{Sales}_{it-1}) &= (\beta_0 - \beta_0) + \beta_1(\text{Price}_{it} - \text{Price}_{it-1}) + \beta_2(\text{Promo}_{it} - \text{Promo}_{it-1}) \\
 &\quad + \beta_3(\text{Location}_i - \text{Location}_i) + \beta_4(\text{Floor}_i - \text{Floor}_i) + (u_i - u_i) \\
 &\quad + (e_{it} - \bar{e}_{it-1}) = \beta_1 \Delta \text{Price}_{it} + \beta_2 \Delta \text{Promo}_{it} + \Delta e_{it}
 \end{aligned}
 \tag{24}$$

```
reg D.sales D.price D.promo i.week, cluster(storeid)
```

Table 10 demonstrates the results of the analysis. Since the first difference estimator is obtained via OLS, investigators can evaluate model fit with standard measures such as the R^2 statistic and use t-tests to determine the significance of individual regression coefficients. We find a negative, though not significant,

Table 10 First difference estimates of the price-response-function

	First difference		Fixed effects	
	<i>Model 20</i>		<i>Model 21</i>	
	B	SE	B	SE
Constant (β_0)	14.79 ^{n.s.}	(20.66)	394.66***	(56.13)
Price (β_1)	-1.66 ^{n.s.}	(1.05)	-1.85*	(1.00)
Promo (β_2)	57.76***	(10.67)	45.21***	(8.57)
Location (β_3)	Omitted		Omitted	
Floor (β_4)	Omitted		Omitted	
Week-fixed effects	Included		Included	
Store-fixed effects	Controlled		Controlled	
R ²	0.23		0.24	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; *n.s.* = not significant; cluster-robust standard errors (SE) in parentheses

impact of price on sales volume (Model 20: $\beta_1 = -1.66$, *n.s.*). Comparing the first difference estimates with the fixed effects estimates reveals that, despite their similarity in removing time-constant cluster-specific effects, coefficient estimates differ. Only for the two time periods case (e.g., 2 weeks), first-differencing and the within transformation will result in identical results.

Since both estimators are unbiased and consistent, investigators can choose freely between the two estimators. Wooldridge (2016, p. 439) and Cameron and Trivedi (2005, p. 705) though note that under certain conditions (i.e., no serial correlation and a homoskedastic error term structure), the fixed effects estimator is more efficient than the first difference estimator. Relatedly, a drawback of the first difference estimator becomes apparent the more unbalanced the panel data is. In the case of balanced panels, only the first measurement occasion is dropped since investigators cannot calculate the first difference. However, in the case of unbalanced panel data – which represents the typical case in applied research – first differences can tremendously reduce the sample size. Picking up Verbeek’s (2017) critique, the investigator might want to carefully inspect the model specification when the fixed effects and first difference estimator yield substantially different results. Such differences likely point to misspecification issues that might violate the fixed effects estimator’s strict exogeneity assumption.

Finally, we want to point to a particular application area of the first difference estimator, common in the marketing literature (e.g., Gill et al. 2017; Manchanda et al. 2015): the difference-in-differences estimator. The difference-in-differences approach mimics an experimental design while using observational data. Its typical set up includes a binary independent variable that discriminates between a treatment and control group (Cameron and Trivedi 2005, Chap. 22; Verbeek 2017, p. 390; Wooldridge 2016, p. 410). Artz and Doering (chapter ► “Exploiting Data from Field Experiments”) discuss its application in more detail in this Handbook.

Advanced Topics in Panel Data Analysis

Dynamic Panel Data Estimation

Dynamic Panel Models Without Cluster-Specific Effects

We now consider dynamic panel models in which investigators include the lagged dependent variable as a time-varying variable. These models are also called lagged-response models, autoregressive models, or Markov models, and Eq. 25 shows their more general form.

$$y_{it} = \beta_0 + \lambda_1 y_{i,t-1} + \dots + \lambda_l y_{i,t-l} + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + \xi_{it} \quad (25)$$

The most popular dynamic model is the autoregressive lag-1 (AR1) model, where the current value of the dependent variable (y_{it}) is regressed on its one-period lagged value ($y_{i,t-1}$). Analogously, in an autoregressive lag-2 AR(2) model, the dependent variable is lagged by two periods ($y_{i,t-2}$). Srinivasan (chapter ► [“Modeling Marketing Dynamics Using Vector Autoregressive \(VAR\) Models”](#)) and Wang and Yildirim (chapter ► [“Applied Time-Series Analysis in Marketing”](#)) offer an in-depth discussion of AR processes in this Handbook, also covering related topics, including vector autoregressive (VAR) models. Note that we assume the independent variables in all AR model variants to be uncorrelated with the error term and the error term to be serially uncorrelated (see Table 9).

Dynamic panel models can be helpful in various situations. First, by including the lagged dependent variable as an independent variable, dynamic panel models help reduce an omitted variable bias. This notion extends our earlier discussion which has focused on accounting for time-constant cluster-specific effects. Including the lagged dependent variable as a time-varying control relaxes the assumption of omitted variables being only time constant and accounts for time-varying effects. In a classical study of explaining market share, Jacobson and Aaker (1985), for instance, use such an approach to model omitted factors such as customer loyalty and distribution systems, which might have influenced market share in the prior periods as well as in the current period. Since no fixed or random effects are included in such a model, it can be conveniently estimated by OLS.

Second, dynamic models are employed when the effect of the lagged dependent variable is itself of scientific interest. For instance, in a study of synergy effects in multimedia communications, Naik and Raman (2003) examine the degree of carry-over in sales levels.

Dynamic panel models have received quite some attention in the marketing literature. Germann et al. (2015), for instance, offer an extensive study on the presence of a CMO on firm performance using a lagged dependent variable in their model. Homburg et al. (2020) show that a dynamic approach is even feasible when only the dependent variable is constructed as a panel variable. In their study, the authors investigate the impact of multichannel sales system design (obtained from a cross-sectional survey) on firm performance (derived from secondary panel performance data). The authors include a lagged measurement of firm performance

as an independent variable to control for variables that equally impacted performance in different time periods.

Despite its advantages, adopting a dynamic panel structure comes with certain limitations that we now briefly review (see Rabe-Hesketh and Skrondal 2012 for a detailed discussion). First, estimating a dynamic panel model is only feasible when occasions are equally spaced in time. For instance, modeling a dynamic panel structure for data collected over several survey waves with different time intervals would not be a sensible task. It is quite a stretch to assume that the lagged dependent variable has the same effect on the current level of the dependent variable regardless of the time interval between them.

Second, the sample size is considerably reduced when a lagged dependent variable is included because lags are missing for each cluster's first observation (section "[Modeling a Price-Response-Function in Differences](#)"). In cases of gaps in the data, the problem of missingness becomes exacerbated because the missing measurement itself is discarded as well as the subsequent observation with its missing lagged measurement.

Third, while not necessarily a limitation, it is worth highlighting that the interpretation of the model's coefficients changes when including the lagged dependent variable. The coefficients now describe the independent variable's effect on the difference between the current and lagged dependent variable. Rearranging the AR (1) dynamic panel model, Eq. 26 shows this point clearly. If λ was equal to 1, the equation would model the change in the dependent variable instead of its level.

$$y_{it} - \lambda y_{i,t-1} = \beta_0 + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + \beta_j z_{ji} + \xi_{it} \quad (26)$$

Finally, when including the time-varying lagged dependent variable as a control to account for omitted variables, such a model makes the strong assumption that all within dependence is due to the lagged dependent variable. In other words, the investigator assumes that the omitted variables are fully accounted for by the lagged dependent variable, there are no cluster-specific effects remaining, and there is no serial correlation present in the error term. This is an admittedly strong assumption, which we can test for, and which we relax in the following section in discussing dynamic panel models that also include cluster-specific effects.

Dynamic Panel Models With Cluster-Specific Effects

We now provide a model that accounts for unobserved cluster-level heterogeneity and adds the lagged dependent variable as an independent variable. A useful feature of such a model is that it can distinguish between two competing explanations of within dependence: unobserved cluster-level heterogeneity (represented by the time-constant cluster effects) and state dependence (represented by the lagged dependent variable). For instance, referring to our data example, the within-store dependence of headphone sales might result from some stores employing very capable managers. Capabilities are likely to be constant over the observation period, in which case this type of dependence represents unobserved cluster heterogeneity. Within-store dependence might also result from high current sales attracting high future sales

through improved financial resources to attract new sales, a case of true state dependence. The corresponding model takes the following form:

$$\text{Sales}_{it} = \beta_0 + \lambda \text{Sales}_{i,t-1} + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \delta \text{WEEK} + u_i + e_{it} \quad (27)$$

It would be tempting to fit the model using the estimation techniques we introduced in the earlier sections. Unfortunately, the conditions for consistent estimation of Eq. 27 are much more demanding than those required for estimations relying on cluster-specific effects or lagged dependent variables alone (Angrist and Pischke 2008).

First, estimating Eq. 27 with the random effects estimator would lead to inconsistent estimates of the coefficients because the lagged dependent variable is per definition correlated with the cluster-specific effect u_i in the error term. As a result, we would run into a problem of correlated unobserved cluster-level heterogeneity and conclude that the random effects estimator is not feasible when estimating dynamic panel models.

Second, estimating Eq. 27 using the fixed effects estimator does not solve this problem either. The within transformation mechanically correlates the within-transformed lagged dependent variable ($\text{Sales}_{i,t-1} - \overline{\text{Sales}}_i$) with the within-transformed error term ($\varepsilon_{i,t-1} - \bar{\varepsilon}_i$) since $\text{Sales}_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$ and hence with $\bar{\varepsilon}_i$. Thus, a fixed effects approach is also not feasible when estimating dynamic panel models.

Finally, estimating Eq. 27 using a first difference approach will also produce inconsistent results, as Eq. 28 shows.

$$\begin{aligned} (\text{Sales}_{it} - \text{Sales}_{i,t-1}) &= \lambda (\text{Sales}_{i,t-1} - \text{Sales}_{i,t-2}) + \beta_1 (\text{Price}_{1it} - \text{Price}_{1i,t-1}) \\ &\quad + \beta_2 (\text{Promo}_{it} - \text{Promo}_{i,t-1}) + (u_i - u_i) + (e_{it} - e_{i,t-1}) \end{aligned} \quad (28)$$

The lagged difference of the dependent variable correlates with the difference of the error term ($\varepsilon_{i,t} - \varepsilon_{i,t-1}$) because $\text{Sales}_{i,t-1}$ is related to its error term $\varepsilon_{i,t-1}$. As such, the first difference approach also violates the assumption of exogeneity (section “[Summary of the Discussed Estimators and Their Underlying Assumptions](#)”).

At the same time, note that $\varepsilon_{i,t} - \varepsilon_{i,t-1}$ is not correlated with lagged differences of the dependent variable beyond the first lag ($\text{Sales}_{i,t-1}$), opening up the possibility of instrumenting the lagged difference of the dependent variable with higher lags. Under the assumption that the error term is serially uncorrelated, Anderson and Hsiao (1981, 1982) introduced such a panel-internal instrumental-variable approach. They suggest that investigators can either use the second lag of the dependent variable ($y_{i,t-2}$) or the lag of the first difference ($y_{i,t-2} - y_{i,t-3}$) as instrumental variables for the differenced dependent variable ($y_{i,t-1} - y_{i,t-2}$). As with all instrumental variable estimation, such an approach assumes that the

instrumental variables fulfill the relevance and validity assumptions (see chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#) by Ebbes et al. in this Handbook).

As an application in marketing, Mizik and Jacobson (2008) rely on the Anderson-Hsiao estimator to model changes in firm profitability. Investigators can obtain a version of this estimator via the `xtivreg` command in Stata.

More efficient IV estimators use additional lags of the dependent variable as instruments, an idea Arellano and Bond (1991) and Blundell and Bond (1998) developed in their generalized method of moments estimator. The Arellano-Bond approach identifies how many lags of the dependent variable are valid instruments and includes all of these lags as instruments (together with the first differences of the model’s exogenous variables). Wies et al. (2019), for instance, employ the Arellano-Bond approach when modeling the time series of advertising expenses in studying how firms manipulate their advertising efforts in response to receiving shareholder complaints. The Arellano-Bond approach can be implemented in Stata using the `xtabond`, `twostep` command.

The Blundell-Bond approach considers Eqs. 27 and 28 as a system of equations. It then uses the lagged differences of the dependent variable as instruments for the levels equation (Eq. 27) and the lagged levels of the dependent variable as instruments for the differences equation (Eq. 28). The Blundell-Bond estimator has improved precision properties, as reflected in lower standard errors. It is implemented in Stata using the `xtdpdsys` approach or the user-written `xtabond2` command (available from SSC `findit xtabond2`; Roodman 2009), and the syntax structure follows `xtabond`.

Stata’s `xtabond` command family provides a set of practical postestimation specification tests to validate the two critical assumptions underlying a panel-internal instrumental variable approach. Specifically, the `estat sargan` command offers a test of overidentifying restrictions, which is useful in confirming the assumption of instrument validity. The `estat abond` command tests whether the error is serially uncorrelated, another desired property we need to confirm when using either the Arellano-Bond or Blundell-Bond estimator. If the test rejects the latter assumption, we can resort to Stata’s `xtdpd` command. This command fits a dynamic panel model based on the Arellano-Bond or Blundell-Bond estimator at the cost of a more complicated syntax but, importantly, allows for low-order autocorrelation in the error term. For a more detailed exposition of handling dynamic panel models in Stata, we recommend Cameron and Trivedi (2005, 2009).

We close with a final recommendation on assessing the usefulness of our selected dynamic panel model when including cluster-specific effects. Instead of estimating variants of the instrumental-variable model (this section), it might be an insightful task to estimate a separate fixed effects (section [“Fixed Effects Estimator”](#)) and random effects model (section [“Random Effects Estimator”](#)) plus a separate lagged dependent variable model (section [“Dynamic Panel Models Without Cluster-Specific Effects”](#)). If the lagged dependent variable model is correct, but one estimates a fixed effects model, the estimated effect will be too large. If a fixed effects model is correct, but one estimates a lagged dependent variable model, the estimated effect will be too small. Therefore, one can think of the fixed effects and

the lagged dependent variable models as bounding the true causal effect of interest (Angrist and Pischke 2008, p. 246).

Random Slope Models: A Multilevel Model Approach to Panel Data

In section “[Dependent Variable: Between and Within Variance](#),” we introduced the focal challenge of nonindependent sales observations in panel data and illustrated this point along with a nested structure figure (Fig. 7). As visible in this figure, panel data can be considered a multilevel model (please refer to chapter ▶ “[Multilevel Modeling](#)” by Haumann et al. in this Handbook for a general introduction to multilevel modeling).

In multilevel terminology, the panel dataset represents a two-level data structure, i.e., sales volume measured over time (Level 1) nested in stores (Level 2). In contrast to typical multilevel data, such as sales reps (Level 1) nested in sales managers (Level 2), panel data have an inherent order at the lowest level.

Knowledge of multilevel models allows expanding the discussion on the random effects estimator. Specifically, multilevel models can consider random cluster-specific intercepts as well as random cluster-specific slope coefficients. While prior models (sections “[Analyze Panel Data Models](#)” and “[Additional Methods in Panel Data Analysis](#)”) assumed that all regression coefficients (besides the cluster-specific intercepts u_i) are the same across stores, in the following, we will allow divergent slope coefficients across clusters (Hox 2010; Raudenbush and Bryk 2002).

For instance, we can extend the random effects model and allow $Price_{it}$ to vary between stores. We indicate this additional variability by including the subscript i to the focal regression coefficient (Eq. 29). Eqs. 30 and 31 formally describe the random intercept and random slope, respectively. Eq. 32 shows that we chose not to include a random slope for $Promo_{it}$.

$$Sales_{it} = \beta_{0i} + \beta_{1i}Price_{it} + \beta_{2i}Promo_{it} + \delta WEEK + e_{it} \tag{29}$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01}Location_i + \gamma_{02}Floor_i + u_{0i} \tag{30}$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \tag{31}$$

$$\beta_{2i} = \gamma_{20} \tag{32}$$

Substituting Eqs. 30, 31, and 32 into Eq. 29 leads to the complete multilevel regression model (Eq. 33). Equation 33 resembles the random effects model (section “[Random Effects Estimator](#)”) and includes the random slope coefficient of the price ($u_{1i} \times Price_{it}$).

$$\begin{aligned} Sales_{it} &= (\gamma_{00} + \gamma_{01}Location_i + \gamma_{02}Floor_i + u_{0i}) + (\gamma_{10} + u_{1i})Price_{it} + \gamma_{20}Promo_{it} \\ &+ \delta WEEK + e_{it} = \gamma_{00} + \gamma_{01}Price_{it} + \gamma_{20}Promo_{it} + \gamma_{01}Location_i \\ &+ \gamma_{02}Floor_i + \delta WEEK + u_{0i} + u_{1i} \times Price_{it} + e_{it} \end{aligned} \tag{33}$$

Including random slopes represents an extension of the random effects model and requires a different estimator. While FGLS (`xtreg`, `re`) is typically used to estimate random effects models, multilevel models require a maximum likelihood estimator.

Stata users need to employ the `xtmixed` command, whose syntax slightly differs from the `xtreg` command, to estimate random slope models. Rabe-Hesketh and Skrondal (2012) offer an in-depth discussion on how to build multilevel models in Stata. As before, we recommend that investigators rely on robust standard errors (`vce(robust)` option).

```
xtmixed sales price promo location floor i.week || storeid:
price, vce(robust)
```

Marketing researchers commonly perform random slope applications of multilevel modeling on panel data. For instance, Anderson et al. (2004) rely on panel data in their multilevel model on customer satisfaction's impact on firm performance. Overall, the authors observe a positive effect of customer satisfaction. However, they also find that the effect of customer satisfaction on firm performance significantly varies between companies and industries. Gruca and Rego (2005), Sorescu and Spanjol (2008), Vomberg et al. (2015), and Wlömert and Papiés (2019) similarly employ multilevel modeling to panel data to obtain deeper insights on how effects vary across clusters.

Addressing Measurement Error with Structural Equation Modeling Based on Panel Data

Our discussion so far has focused on the benefits of panel data to address endogeneity concerns that may arise from an omitted variable bias. In this section, we discuss the implications of another essential source of endogeneity: measurement error, that is, “situations where one or more regressors cannot be measured exactly and are observed with an error” (chapter ► [“Dealing with Endogeneity: A Non-technical Guide for Marketing Researchers”](#) by Ebbes et al.).

All estimators discussed in the previous sections rely on the assumption that the variables are measured without error (section [“Summary of the Discussed Estimators and Their Underlying Assumptions”](#)). However, this assumption might be violated already in the context of rather objectively verifiable information such as reported price or sales volume. Measurement error may arise due to transmission errors into databases. The measurement error problem becomes even more concerning for more abstract constructs frequently investigated in marketing research (e.g., consumers' brand perceptions or customer satisfaction). In this regard, measurement theory suggests that observed variables (e.g., observed customer satisfaction scores) represent the net result of a true score and some random error (chapter ► [“Crafting Survey Research: A Systematic Process for Conducting Survey Research”](#) by Vomberg and Klarmann).

Figure 18 illustrates the measurement error problem (Andreß et al. 2013). Please assume that we measured customer satisfaction scores for two consecutive years (t and $t + 1$). Following standard conventions in the literature (e.g., chapter ► “Structural Equation Modeling”), we present the observed values of customer satisfaction in rectangular boxes. As predicted by measurement theory, those values are influenced by the true level of customer satisfaction (“Customer Satisfaction*”), which is specified as circled in Fig. 18 and by a measurement error (ϵ_t).

The amount of measurement error will impact the estimated relationship between customer satisfaction over time, which we could model with dynamic panel data models (section “Dynamic Panel Data Estimation”). For instance, true state dependence (serial correlation of customer satisfaction over time) might be $r = 0.82$ ($\text{corr}(\text{CustomerSatisfaction}^*_t; \text{CustomerSatisfaction}^*_{t+1})$). However, we measured customer satisfaction with error so that the true customer satisfaction level is not translated one-to-one into an observable customer satisfaction level. Statistically speaking, customer satisfaction has a factor loading of $\lambda = 0.70$ (if customer satisfaction was measured without error, the factor loading would be $\lambda = 1.00$). As a consequence, the observed correlation between customer satisfaction becomes smaller ($\text{corr}(\text{CS}_t; \text{CS}_{t+1}) = 0.70 \times 0.82 \times 0.70 = 0.40$) and true state dependence is underestimated.

In the context of cross-sectional data, problems of measurement error only concern the independent variables. The error term captures the measurement error of the dependent variable. However, in the context of panel data, measurement error concerns the dependent variable, too. For instance, as illustrated previously, measurement error will bias state dependence estimates toward zero in dynamic panel data models (section “Dynamic Panel Data Estimation”). Additionally, measurement error does not only impact models in which lagged dependent variables are of substantive interest. Demeaning (fixed effects estimator), quasi-demeaning (random effects estimator), or first-difference transformations are equally affected by measurement error. The fixed effects estimator is particularly susceptible to an

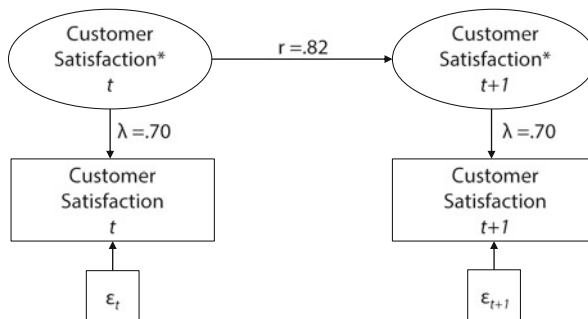


Fig. 18 Measurement error in dynamic panel models illustrated

attenuation bias from measurement error (Angrist and Pischke 2008; Griliches and Hausman 1986; Wooldridge 2016, p. 440).

Structural equation modeling represents a way to directly model measurement error (Baumgartner and Weijters present an introduction to ► “[Structural Equation Modeling](#)” chapter in this Handbook). In cross-sectional analyses, researchers use different indicators for the same construct to estimate a construct’s reliability and to directly model measurement error. In the panel data context, researchers can use the same indicators from different time periods to capture the underlying construct. For instance, the investigator may use customer satisfaction scores of 3 years (e.g., customer satisfaction measured in $t = 2019, 2020,$ and 2021) to estimate a latent customer satisfaction construct. This approach is identical to an evaluation of test-retest reliability and offers the advantage of directly accounting for measurement error. However, the downside of this approach is that fewer possibilities exist to estimate the model (e.g., employing fixed effects estimation is not feasible). Investigators can implement a structural equation model via Stata’s `sem` and `gsem` commands.

Luo and Bhattacharya (2006) offer an application example in the marketing context. The authors rely on panel data obtained from Fortune’s Most Admired Companies and use repeated corporate social responsibility measurements to capture the underlying latent corporate social responsibility construct. Cho and Pucik (2005) apply a similar approach when modeling how firm innovativeness and product quality relate to market value.

Conclusion

This chapter sought to provide a gentle nontechnical introduction to panel data analysis for marketing researchers. At the core of panel data analysis is the challenge of how best to account for the dependency of observations within and across clusters. We discussed the POLS estimator’s limitations and reviewed the two most popular panel estimators that explicitly model the panel structure: the fixed effects and random effects estimators. For completeness, we also discussed the between effects estimator and first difference estimator, as well as the combined approach and the Hausman-Taylor approach. Using a real-life example, we applied these estimators in the context of a price-response-function for headphone sales. We conducted the empirical analysis in Stata, a very user-friendly statistical software package for analyzing panel data. Despite using the same dataset to estimate the price-response-functions, we find results differ considerably depending on the selected estimator. These divergent results demonstrate the need to thoroughly understand the different model-identifying assumptions, benefits, and limitations of each estimator. As such, we hope our chapter contributes to turning readers into cognizant “regression engineers” (Germann et al. 2015) and offers researchers the necessary skill set to conduct meaningful analyses. Panel data provide exciting opportunities to investigate new research questions, and we hope that readers find this introduction helpful in developing their models.

Cross-References

- ▶ [Applied Time-Series Analysis in Marketing](#)
- ▶ [Assessing the Financial Impact of Brand Equity with Short Time-Series Data](#)
- ▶ [Choice-Based Conjoint Analysis](#)
- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Exploiting Data from Field Experiments](#)
- ▶ [Modeling Marketing Dynamics Using Vector Autoregressive \(VAR\) Models](#)
- ▶ [Multilevel Modeling](#)
- ▶ [Regression Analysis](#)
- ▶ [Structural Equation Modeling](#)
- ▶ [Willingness to Pay](#)

References

- Anderson, T. W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 76(375), 598–606.
- Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1), 47–82.
- Anderson, E. W., Fornell, C., & Mazvancheryl, S. K. (2004). Customer satisfaction and shareholder value. *Journal of Marketing*, 68(4), 172–185.
- Andreß, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied panel data analysis for economic and social surveys*. Berlin, Heidelberg: Springer-Verlag.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. New Jersey: Princeton University Press.
- Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2), 277–297.
- Ataman, M. B., Van Heerde, H. J., & Mela, C. F. (2010). The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, 47(5), 866–882.
- Bayer, E., Srinivasan, S., Riedl, E. J., & Skiera, B. (2020). The impact of online display advertising and paid search advertising relative to offline advertising on firm performance and firm value. *International Journal of Research in Marketing*, 37(4), 789–804.
- Bijmolt, T. H., Van Heerde, H. J., & Pieters, R. G. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42(2), 141–156.
- Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.
- Boulding, W., & Christen, M. (2003). Sustainable pioneering advantage? Profit implications of market entry order. *Marketing Science*, 22(3), 371–392.
- Boulding, W., & Christen, M. (2008). Disentangling pioneering cost advantages and disadvantages. *Marketing Science*, 27(4), 699–716.
- Borah, A., & Tellis, G. J. (2014). Make, buy, or ally? Choice of and payoff from announcements of alternate strategies for innovations. *Marketing Science*, 33(1), 114–133.
- Butt, M. N., Antia, K. D., Murtha, B. R., & Kashyap, V. (2018). Clustering, knowledge sharing, and intrabrand competition: A multiyear analysis of an evolving franchise system. *Journal of Marketing*, 82(1), 74–92.
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.
- Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata*. Texas: Stata Press.

- Cho, H. J., & Pucik, V. (2005). Relationship between innovativeness, quality, growth, profitability, and market value. *Strategic Management Journal*, 26(6), 555–575.
- Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: Nature, purpose, and findings. *Journal of Marketing*, 60(4), 7–18.
- Germann, F., Ebbes, P., & Grewal, R. (2015). The chief marketing officer matters! *Journal of Marketing*, 79(3), 1–22.
- Gill, M., Sridhar, S., & Grewal, R. (2017). Return on engagement initiatives: A study of a business-to-business mobile app. *Journal of Marketing*, 81(4), 45–66.
- Greene, W. H. (2003). *Econometric analysis* (6th ed.). United Kingdom: Pearson Education.
- Griliches, Z., & Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics*, 31(1), 93–118.
- Gruca, T. S., & Rego, L. L. (2005). Customer satisfaction, cash flow, and shareholder value. *Journal of Marketing*, 69(3), 115–130.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251–1271.
- Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica: Journal of the Econometric Society*, 1377–1398.
- Ho-Dac, N. N., Carson, S. J., & Moore, W. L. (2013). The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? *Journal of Marketing*, 77(6), 37–53.
- Homburg, C., Vomberg, A., & Muehlhaeuser, S. (2020). Design and governance of multichannel sales systems: Financial performance consequences in business-to-business markets. *Journal of Marketing Research*, 57(6), 1113–1134.
- Hox, J. J. (2010). *Multilevel analysis – techniques and applications*, 2nd edn. New York: Routledge.
- Jacobson, R., & Aaker, D. A. (1985). Is market share all that it's cracked up to be? *Journal of Marketing*, 49(4), 11–22.
- Kennedy, P. (2008). *A Guide to econometrics* (6th ed.). New Jersey: Wiley.
- Kummer, M., & Schulte, P. (2019). When private information settles the bill: Money and privacy in Google's market for smartphone applications. *Management Science*, 65(8), 3470–3494.
- Lovett, M. J., Peres, R., & Shachar, R. (2013). On brands and word of mouth. *Journal of Marketing Research*, 50(4), 427–444.
- Luo, X., & Bhattacharya, C. B. (2006). Corporate social responsibility, customer satisfaction, and market value. *Journal of Marketing*, 70(4), 1–18.
- Manchanda, P., Packard, G., & Pattabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science*, 34(3), 367–387.
- McAlister, L., Srinivasan, R., Jindal, N., & Cannella, A. A. (2016). Advertising effectiveness: The moderating effect of firm strategy. *Journal of Marketing Research*, 53(2), 207–224.
- Misangyi, V. F., Elms, H., Greckhamer, T., & Lepine, J. A. (2006). A new perspective on a fundamental debate: A multilevel approach to industry, corporate, and business unit effects. *Strategic Management Journal*, 27(6), 571–590.
- Mizik, N., & Jacobson, R. (2008). The financial value impact of perceptual brand attributes. *Journal of Marketing Research*, 45(1), 15–32.
- Naik, P. A., & Raman, K. (2003). Understanding the impact of synergy in multimedia communications. *Journal of Marketing Research*, 40(4), 375–388.
- Nath, P., & Mahajan, V. (2008). Chief marketing officers: A study of their presence in firms' top management teams. *Journal of Marketing*, 72(1), 65–81.
- Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, 22(1), 435–480.
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata, volumes I and II: Multilevel and longitudinal modeling using stata*. Texas: Stata Press.
- Rao, V. R., Agarwal, M. K., & Dahlhoff, D. (2004). How is manifest branding strategy related to the intangible value of a corporation? *Journal of Marketing*, 68(4), 126–141.

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models – Applications and data analysis methods*. California: Sage.
- Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal*, 9(1), 86–136.
- Sorescu, A. B., & Spanjol, J. (2008). Innovation's effect on firm value and risk: Insights from consumer packaged goods. *Journal of Marketing*, 72(2), 114–132.
- Steenkamp, J. B. E., & Geyskens, I. (2014). Manufacturer and retailer strategies to impact store brand share: Global integration, local adaptation, and worldwide learning. *Marketing Science*, 33(1), 6–26.
- Verbeek, M. (2017). *A Guide to modern economics* (5th ed.). New Jersey: Wiley.
- Vomberg, A., Homburg, C., & Bornemann, T. (2015). Talented people and strong brands: The contribution of human capital and brand equity to firm value. *Strategic Management Journal*, 36(13), 2122–2131.
- Von Auer, L. (2013). *Ökonometrie – eine einföhrung*, 4th edition, Berlin Heidelberg: Springer.
- Warren, N. L., & Sorescu, A. (2017). Interpreting the stock returns to new product announcements: How the past shapes investors' expectations of the future. *Journal of Marketing Research*, 54(5), 799–815.
- Wies, S., Hoffmann, A. O. I., Aspara, J., & Pennings, J. M. (2019). Can advertising investments counter the negative impact of shareholder complaints on firm value? *Journal of Marketing*, 83(4), 58–80.
- Wlömert, N., & Papiés, D. (2019). International heterogeneity in the associations of new business models and broadband internet with music revenue and piracy. *International Journal of Research in Marketing*, 36(3), 400–419.
- Wolters, H. M., Schulze, C., & Gedenk, K. (2020). Referral reward size and new customer profitability. *Marketing Science*, 39(6), 1166–1180.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Massachusetts: MIT Press.
- Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Massachusetts: Cengage Learning.



Applied Time-Series Analysis in Marketing

Wanxin Wang and Gokhan Yildirim

Contents

Introduction	470
Univariate Time-Series Treatments and Diagnostics	471
Autoregressive (AR) and Moving Average (MA) Process	471
Testing for Evolution Versus Stationarity	472
ARIMA Models	474
Single Equation Time-Series Models with Exogenous Variables	476
Multiple Time-Series Models: Dynamic Systems	479
Granger Causality Tests	481
Cointegration Test	483
Vector Autoregressive and Vector Error–Correction Model	483
Order of Lags in VAR Models	484
Generalized Impulse Response Functions	485
Generalized Forecast Error Variance Decomposition	487
Volatility Models	489
Conclusion	491
Cross-References	491
Appendix	492
Software Application	492
Data Visualizations	493
ARIMA Modeling	494
VAR Model Steps	502
References	512

Abstract

Time-series models constitute a core component of marketing research and are applied to solve a wide spectrum of marketing problems. This chapter covers traditional and modern time-series models with applications in extant marketing

W. Wang · G. Yildirim (✉)

Imperial College Business School, Imperial College London, London, UK

e-mail: wanxin.wang13@imperial.ac.uk; g.yildirim@imperial.ac.uk

© Springer Nature Switzerland AG 2022

C. Homburg et al. (eds), *Handbook of Market Research*,

https://doi.org/10.1007/978-3-319-57413-4_37

469

research. We first introduce basic concepts and diagnostics including stationarity test (the augmented Dicky-Fuller test of unit roots), and autocorrelation plots via autocorrelation function (ACF) and partial autocorrelation function (PACF). We then discuss single-equation time-series models such as autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models with and without exogenous variables. Multiple-equation dynamic systems including vector autoregressive (VAR) models together with generalized impulse response functions (GIRFs) and generalized forecast error variance decomposition (GFEVD) are then discussed in detail. Other relevant models such as generalized autoregressive conditional heteroskedasticity (GARCH) models are covered. Finally, a case study accompanied by data and R codes is provided to demonstrate detailed estimation steps of key models covered in this chapter.

Keywords

Time-series models · Marketing · ARIMA · VAR · GIRF · GFEVD · GARCH

Introduction

I have seen the future and it is very much like the present, only longer. – Kehlog Albran, *The Profit*

Firms collect data on the past to understand the present and to forecast the future. Performance measures such as sales, market share, and revenues are often path dependent, meaning that their past values can inform the present. This is great news for marketing analysts because they are able to forecast the future using historical time-stamped data series. The key challenge is how they can accurately capture the dynamics and variation patterns of data using proper statistical techniques. Further, firms take various marketing actions (e.g., TV and radio advertising, online display and search engine advertising, and social media campaigning, etc.) to help boost performance. More importantly, these marketing actions are usually designated to exert influence not only instantaneously in the current period, but persistently into the future. For example, when assessing advertising effectiveness, analysts typically add up past advertising efforts (while considering a certain level of decay) and evaluate their cumulative impact on sales. Given an omnichannel marketing scheme, firms also need to examine cross-effects, or interactions, between different marketing instruments to see whether there are any synergies or cannibalizations. Further, the relationship between marketing efforts and performance is often bilateral; feedback loops exist so that performance from previous periods helps decide marketing strategies in the current period.

Analytical tools that capture dynamics of performance measures and that between marketing and performance are time-series models. Over the decades, time-series models have been evolving from univariate to multivariate, and then to dynamic

systems consisting of multiple time series. Univariate models (e.g., autoregressive moving average or ARMA) assume that the contemporaneous value of a series is only influenced by its past values. Multivariate models extend and usually outperform univariate models by including current and lagged values of other factors (e.g., price, marketing, distributions, and regulatory change) that should also exert influence on performance. Dynamic systems like vector autoregressive (VAR) models are widely adopted to tackle more complex relationships between series. These models can be used to capture dual causality between performance and other predicting variables (e.g., do past sales affect current marketing decisions?) and between predicting variables (e.g., does online marketing complements or substitutes offline marketing?).

Time-series analyses can also inform researchers of more specific patterns of the dataset that they are working with. For example, when forecasting sales of a firm, previous marketing research can inform analysts what variables should be incorporated into the model. However, previous findings are not able to tell the exact order of lags (e.g., how many weeks of past sales, price, and marketing investments should we consider?) or the direction of causality (e.g., does consumer social media sentiment directly or indirectly impact sales?). We need to rely on a set of time-series diagnostic tests to find answers to these questions.

More impressively, time-series models, especially the modern ones (e.g., dynamic systems) can enable researchers to uncover some powerful linkage between factors that are previously overlooked (Srinivasan et al. 2016). For example, researchers have bridged firm offline marketing with consumer online activities and sales through VAR modeling (Srinivasan et al. 2016); others have found that growth or decrease in the volume of consumer social media posts can affect the stock market valuation of firms (van Diejen et al. 2019).

This chapter proceeds as follows. We will start from the basics, including treatment and diagnostics of univariate time-series models. Then we will talk about traditional time-series models, for instance, autoregressive integrated moving average (ARIMA) models. We will then discuss modern time-series models like vector autoregressive (VAR) models. Additionally, we also cover generalized autoregressive conditional heteroscedasticity (GARCH) models that deal with time-series volatility. Finally, in the [Appendix](#), we present a case study where we apply methods introduced in this chapter using R to solve marketing challenges for a firm.

Univariate Time-Series Treatments and Diagnostics

Autoregressive (AR) and Moving Average (MA) Process

Let us start with the simplest model where we describe certain performance metric of a brand in each time period t with its first lag. Specifically, let s_t denote the sales of a brand in week t , and s_t is determined by sales in the previous week s_{t-1} . We call this a first-order autoregressive, or $AR(1)$ process:

$$s_t = c + \varphi s_{t-1} + \varepsilon_t \quad (1)$$

where c is a constant term and φ is the parameter that captures the effect of sales in the previous week. ε_t is white noise with mean zero and variance σ_ε^2 .

Another form of time series, i.e., moving average (MA) process, assumes that sales at the current period is affected by a past shock (e.g., a natural disaster) other than its own past values. A first-order moving average or a $MA(1)$ process of sales at time t is written as:

$$s_t = c + \varepsilon_t + \theta \varepsilon_{t-1} \quad (2)$$

where ε_{t-1} is the error term or the shock from the last period. θ captures the impact of such past shock (e.g., an earthquake that happened one week ago) on current sales.

A $MA(1)$ process differs from an $AR(1)$ process in that instead of assuming the past shock as coming from past sales s_{t-1} , it assumes that such shock comes from the random component of s_{t-1} , namely ε_{t-1} .

$AR(1)$ and $MA(1)$ process can be generalized to $AR(p)$ and $MA(q)$ process, respectively, where p and q refer to the highest order of lagged value and error term, respectively.

What if we want to model weekly sales while taking the impact of both past sales and past random shocks into consideration? We can combine an $AR(p)$ and an $MA(q)$ processes to have an autoregressive moving average, or $ARMA(p, q)$ process. For instance, an $ARMA(1, 1)$ process is written as:

$$s_t = c + \varphi s_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t \quad (3)$$

Figure 1 shows an example of an $ARMA(1, 1)$ process, with $\varphi = -0.70$ and $\theta = 0.99$. $ARMA$ models have been quite commonly adopted in marketing research during 1970–1990s and proved suitable to capture dynamics in various contexts (e.g., Dekimpe and Hanssens 1995).

Testing for Evolution Versus Stationarity

Marketing models only make sense when the time series being analyzed is mean stationary or trend stationary, meaning that it always converges back to a fixed mean or a fixed mean plus any trend detected. Otherwise, the series is said to be non-stationary or evolving. Why do we need to emphasize on series stationarity? Consider an evolving series whose value is constantly increasing. In such case, sample statistics such as mean and variance are not really descriptive of the data pattern, since they are not stable and keep getting larger as we include more data points. Therefore, reporting the mean and variance of an evolving series is not informative or helpful for decision-making. Further, we will not be able to generate reliable results if we use evolving variables to predict sales. This is why evolving series need to be transformed to stationarity.

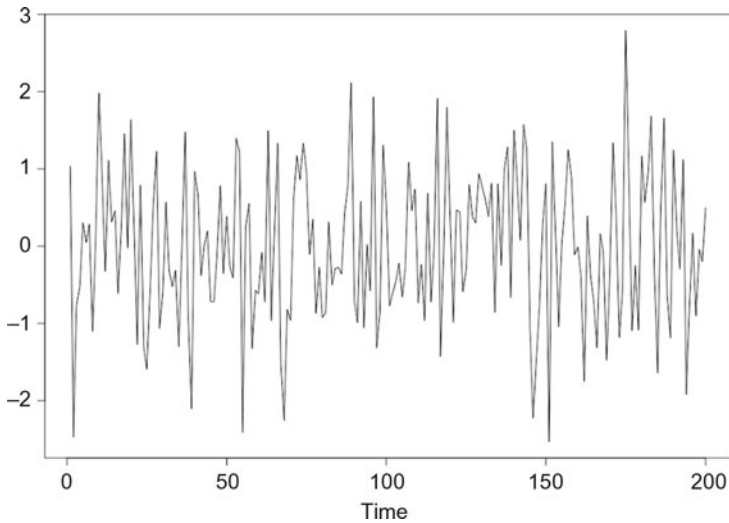


Fig. 1 An Example of an ARMA (1,1) process

Given an $ARMA(1, 1)$ process in Eq. (3), how can we determine if it is stationary or not? The answer lies in the φ term:

- If $|\varphi| < 1$, we call this series *stationary*, with a *time-independent* mean and variance, meaning that the mean $E(s_t)$ and variance $\sigma^2(s_t)$ are the same for all t .
- If $|\varphi| = 1$, then the effect of past sales s_{t-1} is said to have a permanent effect on current sales s_t . In this case, sales will not be reverting to a certain level, but instead evolving.
- If $|\varphi| > 1$, then the effect of past sales through φ will exert an even stronger influence as time goes. Such a data pattern should be rarely observed in the context of marketing.

Note that we only need to evaluate the AR part (through φ) of an ARMA process for stationarity. This is because ε_t is independently and normally distributed with a zero mean, indicating that an MA process is stationary regardless of the value of θ .

More formally, we can test for stationarity of an ARMA process through *unit root test* (For the stationarity test for time series data with a permanent step-change, please refer to unit root test with structural breaks (e.g., Deleersnyder et al. 2002)). A time series is evolving if it has a unit root, and stationary otherwise. The most widely adopted method is the augmented Dickey-Fuller (ADF) test (Kwiatkowski et al. 1992). Let us use an $AR(2)$ version of the weekly sales time series introduced in Eq. (1). Note that we only focus on the AR part of ARMA process since the MA part is always stationary.

Recall that brand sales in week t can be written as an $AR(2)$ process:

$$s_t = c + \varphi_1 s_{t-1} + \varphi_2 s_{t-2} + \varepsilon_t \quad (4)$$

If we define $\Delta s_t = s_t - s_{t-1}$, we can reformulate the process in Eq. (4) into:

$$s_t = c + (\varphi_1 + \varphi_2) s_{t-1} - \varphi_2 (s_{t-1} - s_{t-2}) + \varepsilon_t \quad (5)$$

And further:

$$\Delta s_t = c + \eta s_{t-1} + \lambda \Delta s_{t-1} + \varepsilon_t \quad (6)$$

where $\eta = \varphi_1 + \varphi_2 - 1$, and $\lambda = -\varphi_2$. The $AR(2)$ process in Eq. (5) has a *unit root* if $\varphi_1 + \varphi_2 = 1$. This is equivalent to testing whether $\eta = 0$ in Eq. (6).

In practice, researchers can first-difference the sales series, and then estimate a linear regression model where the first-differenced sales at time t (i.e., Δs_t) is the dependent variable, and lagged values of Δs_t 's and s_{t-1} are independent variables. The null hypothesis that $\eta = 0$ will be rejected if the regression coefficient of s_{t-1} (i.e., η) is statistically significant, indicating series stationarity. Alternatively, we can use statistical software such as R, Stata, and EViews to quickly generate test results. Readers may refer to “[Appendix](#)” of this chapter for detailed guidance on how to use R to perform ADF test.

ARIMA Models

What if the time series that we work on is found nonstationary with ADF test indicating a unit root? A proper technique to deal with this situation is to transform the series to stationarity by *differencing*. For example, if we find the weekly sales series in Eq. (3) is evolving, we can first-difference it to a series z_t , where $z_t = s_t - s_{t-1}$. By subtracting the first lag from the current value, the series may become *difference stationary*. The ARMA process with a differencing operation is called an ARIMA (integrated ARMA) process. ARIMA model is often adopted by practitioners and researchers for prediction purpose, e.g., demand forecasting. For example, given the sales, price, and marketing activities in the past 24 months, what is the predicted sales for the next 12 months?

An $ARIMA(1, 1, 1)$ model, which is a combination of $AR(1)$, $MA(1)$, and a first-order differencing operation, can be written as:

$$z_t = c + \varphi_1 z_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (7)$$

where $z_t = s_t - s_{t-1}$. This model can be easily extended to a generalized case where we have an $ARIMA(p, d, q)$ process, with p and q representing the highest order of lags in AR and MA component, respectively, and d the order of differencing.

ACF and PACF Analysis

Given a stationary time series (e.g., an ARIMA process), how can we determine the exact order of lags to include (i.e., the value of p and q)? Researchers typically rely on the autocorrelation function (ACF) and partial autocorrelation function (PACF) to determine the MA and AR part of the process, respectively.

Again, let s_t be the value of sales at week t . The ACF series is derived from calculating the correlations between s_t and s_{t-k} (i.e., sales k weeks ago) for every k . To calculate these correlations, we first derive the unconditional mean and variance of s_t for an $AR(1)$ process:

$$E(s_t) = E(c + \phi s_{t-1} + \varepsilon_t) = c + E(\phi s_{t-1}) + E(\varepsilon_t) = c + \phi E(s_{t-1}) \quad (8)$$

Hence

$$E(s_t) = \frac{c}{1 - \phi} \quad (9)$$

$$\begin{aligned} \sigma^2(s_t) &= \text{var}(c + \phi s_{t-1} + \varepsilon_t) = 0 + \text{var}(\phi s_{t-1}) + \text{var}(\varepsilon_t) \\ &= \phi^2 \text{var}(s_{t-1}) + \sigma_\varepsilon^2 \end{aligned} \quad (10)$$

$$\sigma^2(s_t) = \frac{\sigma_\varepsilon^2}{1 - \phi^2} \quad (11)$$

where σ_ε^2 is the constant variance of disturbance term ε_t .

Hence the correlation between two data points that are k periods apart is:

$$\rho_k = \phi^k \quad (12)$$

From Eq. (12), when $|\phi| < 1$, ρ_k will converge or oscillate towards zero as k gets larger.

Partial autocorrelation function (PACF) of k^{th} order refers to the correlation between two data points in a time series that are k lags apart, holding all other $(k - 1)$ intermediate observations constant. For example, let us denote correlation between sales in period m and n as $\omega_{m, n}$. Given that a stationary time series has constant autocorrelation, we have $\omega_{t, t+1} = \omega_{t+1, t+2} = \rho_1$. Then the PACF between s_t and s_{t+2} while holding s_{t+1} constant can be written as:

$$\omega_{t,t+1,t+2} = \frac{\omega_{t,t+2} - \omega_{t,t+1}\omega_{t+1,t+2}}{\sqrt{(1 - \omega_{t,t+1}^2)(1 - \omega_{t+1,t+2}^2)}} = \frac{(\rho_2 - \rho_1^2)}{(1 - \rho_1^2)} \quad (13)$$

Graphically, if we plot PACF against k , we will see a spike of PACF equal to ACF at $k = 1$, and zeros afterwards. For example, Fig. 2 below shows weekly sales of a mature consumer good brand (brand A) for one year. The time series of sales can be described as an $AR(1)$ process with $\phi = 0.5$. Panel (a) and (b) in Fig. 3 show the ACF

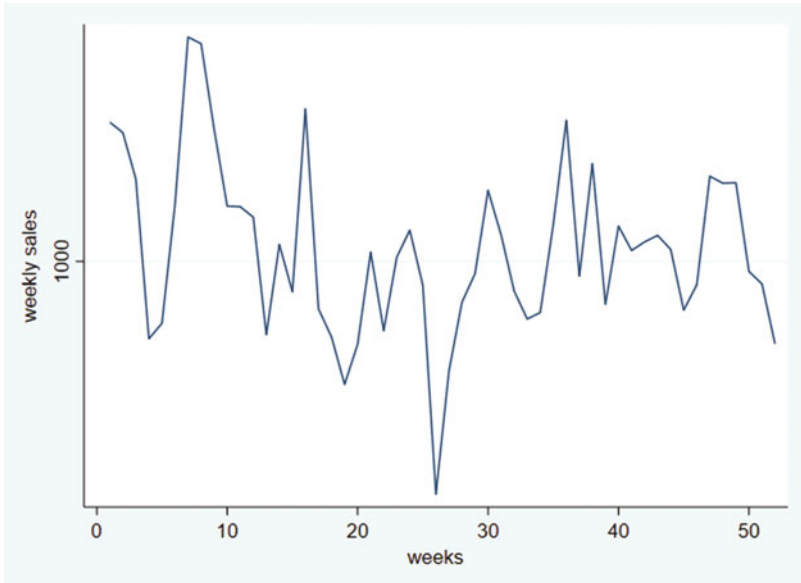


Fig. 2 Weekly sales (in thousands) of brand A

and PACF of the sales series, respectively. In panel (b), the PACF plot has a significant drop (or cutoff) from lag 1 to lag 2, indicating that the series is an $AR(1)$ process.

In terms of the order of AR and MA process, we can refer to the rules as follows:

- The lag at which the PACF cuts off is the indicated maximum order of AR lags.
- The lag at which the ACF cuts off is the indicated maximum order of MA lags.

Single Equation Time-Series Models with Exogenous Variables

In the previous sections, we have been studying univariate time series. More specifically, we have considered sales to be only affected by its past values and past random shocks as described in ARIMA processes. While these models are able to capture sales dynamics, the reality is that firms spend a lot of effort on many other activities to improve their sales performance. Univariate models are hence limited since they fail to incorporate other factors that also make a substantial difference such as various marketing activities. Meanwhile, marketing managers are keen to justify their marketing expenditure by evaluating the effect of marketing on sales. For example, what will be the change in sales if we spend 10% more on marketing?

This section links the role of marketing with firm performance explicitly by introducing multivariate time-series models, or single equation time-series

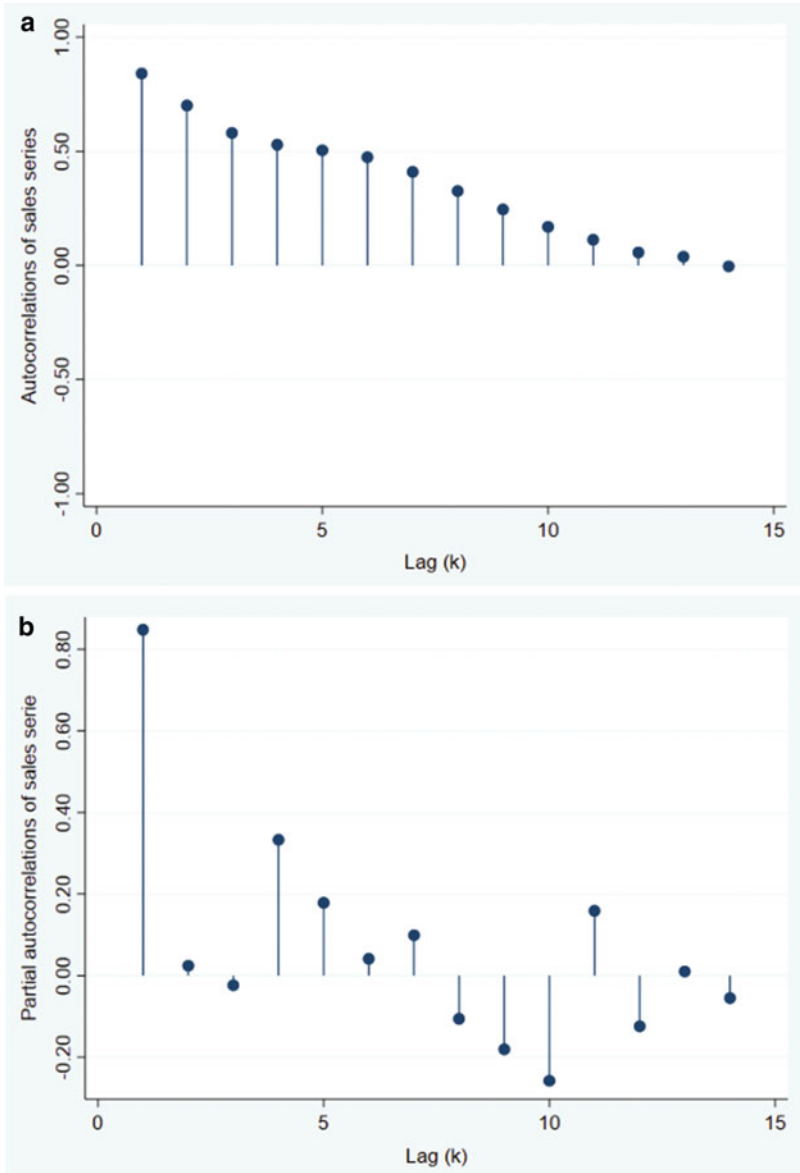


Fig. 3 ACF and PACF plot of weekly sales of brand A

models with exogenous variables. We emphasize “single equation” here to distinguish this type of model from dynamic systems consisting of multiple time series. These models are sometimes known as ARIMA-X, where X stands for “exogenous.”

Assume that besides sales itself, only one more variable, price, has impact on sales and that price itself is also subject to time-series patterns. We can incorporate the current and lag price into our model to predict sales through a *transfer function* $v_k(B)$, where $v_k(B) = v_0 + v_1B + v_2B^2 + \dots + v_kB^k$. B is the backshift operator, where $B^k y_t = y_{t-k}$, and k is the highest order of lags. A dynamic regression model of sales s_t on price variable is:

$$s_t = c + v_k(B)p_t + \varepsilon_t \quad (14)$$

where $v_k(B)p_t = v_0p_t + v_1p_{t-1} + \dots + v_kp_{t-k}$.

The transfer function $v_k(B)$ is also called the *impulse response function* (IRF) with the coefficients v 's called *impulse response weights*. Further, referring to Eq. (14), if we find that $v_0 = 0$ and that all other v -coefficients are nonzero, then price in the current period t does not have an impact on sales in the same period, while prices in the past periods do. The model is said to have a “wear-in time” (or “dead time”) of one, i.e., a change made to the price will start to exert influence only from the next period onwards. Wear-in time reflects the speed at which changes in a firm’s marketing mix impact sales performance. More formally, “wear-in time” of a model is measured by the number of *consecutive* v -coefficients with zero value, starting from v_0 .

Finally, similar to the idea of structural break in time-series data, we could also accommodate certain discrete events that shock the series significantly into our sales models. For example, exogenous shocks such as regulatory change or introduction of a new product introduced by a rival brand may lead to significant rise or drop in sales data. Here we distinguish two types of effects that a shock can have: a *pulse effect* or a *step effect* (Pauwels 2017).

A pulse effect is a temporary effect that decays or disappears gradually. In contrast, a step effect is supposed to have permanent effect once it occurs. To analyze the impact of shocks requires *intervention analysis*, which extends the transfer function approach described above.

Figure 4 shows example of pulse effect at time t' on a stationary process (panel a) and nonstationary process (panel b), respectively. In the case depicted in panel (a), the intervention could be a price promotion at time t' of a mature consumer good product with stationary demand from consumers. The promotion results in a temporary spike at sales of time t' , after which sales revert to its stationary mean. The corresponding transfer function is v_0x_t at $t = t'$ and zero elsewhere. Panel (b) shows a situation of pulse intervention, where the intervention function is the same, but sales is nonstationary (e.g., a new brand with high market potential). Here the pulse intervention results in a temporary drop in sales at $t = t'$. Sales return to the level that is determined by its nonstationary character afterwards.

Figure 5 shows two examples of step interventions. Here the change to sales after time t' is long lasting, meaning that it could be permanent (panel a) or semi-permanent (panel b). For example, we can think of the series in panel (a) as sales of a brand that successfully introduced a major market innovation at time t' . Sales jumped to a higher level immediately at $t = t'$ and stay at the new level for all $t > t'$. The transfer function,

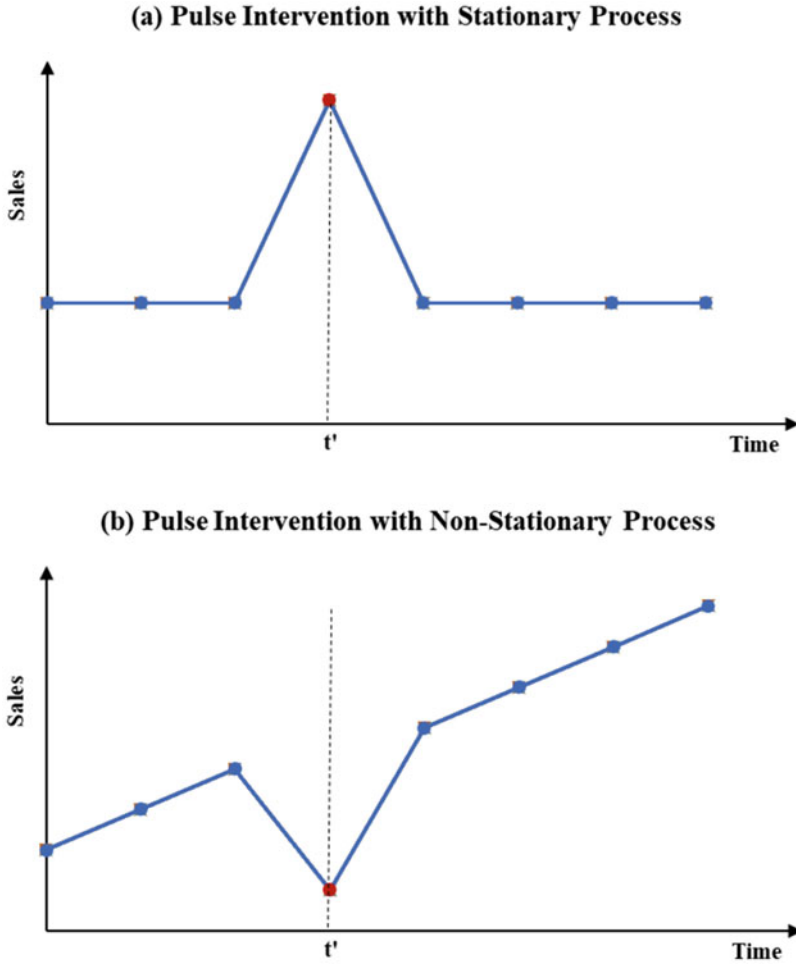


Fig. 4 Examples of pulse interventions

in this case, would be one where v_0x_t , and $x_t = 0$ for $t < t'$ and $x_t = 1$ for $t \geq t'$. Panel (b) shows the case where a step intervention lasts for several periods (three periods in this case) but not “forever,” i.e., semi-permanent. The transfer function would be one where $x_t = 0$ for $t < t'$, and $x_t = 1$ for $t = t', t' + 1, t' + 2$.

Multiple Time-Series Models: Dynamic Systems

Multivariate models are often preferred over univariate models because they can capture not only the effect of past performance, but also that of other model covariates (e.g., price, marketing efforts, and competitors’ offerings). However,

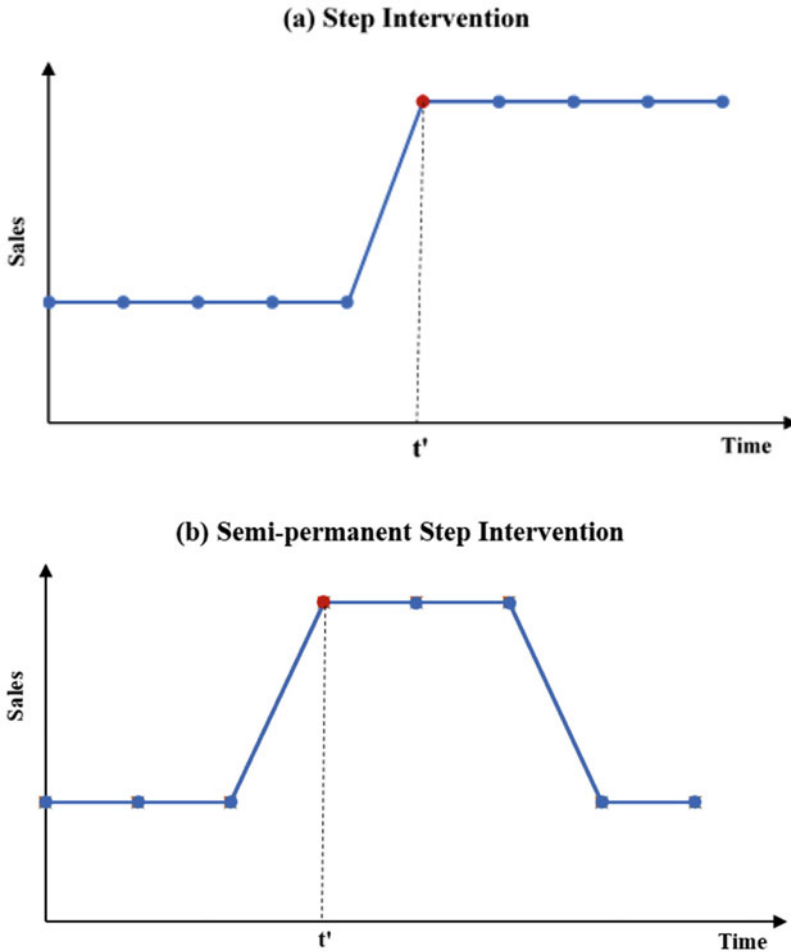


Fig. 5 Examples of step interventions

these models are still subject to certain limitations in dealing with joint endogeneity between variables (e.g., dual causality and feedback loops, etc.).

For example, we can quantify the effect of a firm's marketing activities on sales by fitting a linear regression model. The assumption here is that marketing impacts sales, not the other way around. However, in practice, to maximize overall return on marketing investment, firms also adjust their marketing strategies based on sales performance in the previous periods (please refer to more detailed discussions and applications in chapters ► [“Measuring Sales Promotion Effectiveness”](#) and ► [“Return on Media Models”](#) in this book). Additionally, firms need to be aware of potential synergies or cannibalizations between different marketing actions and refine their marketing portfolios from time to time. For instance, one can expect a certain level of complementarity

(positive “spill-over”) between online display advertising and offline in-store promotion. On the other hand, firms may consider alternating TV and radio ads instead of having them simultaneously to avoid consumer fatigue (i.e., negative “spillover”).

These problems described above can be taken care of via estimating a dynamic system of multiple time series, or a vector autoregressive (VAR) model. Models as such are more accurate in both model fitting and forecasting, compared with traditional time-series models (Lütkepohl 2005).

Our structure for this section is adapted from the “persistence modeling framework” that was first developed by Dekimpe and Hanssens (1995). You can find most of the recent marketing research papers following the procedure described in this framework (e.g., De Haan et al. 2016; Srinivasan et al. 2016).

This section will cover topics listed below:

1. *Granger causality tests*, which focus on understanding the direction of causality between model variables
2. *Unit root and cointegration tests*, which focus on understanding whether model variables are stationary over time or evolving and on whether the evolving variables (if any) are tied in certain long-term equilibrium, respectively.
3. *Dynamic system modeling*, which is typically done via vector autoregressive (VAR) model or vector error correction (VEC) model, depending on results obtained from 2.
4. *Policy simulation analysis*, which focuses on evaluating short-term and long-term impact of marketing on performance via impulse response function (IRF) analysis.
5. *Drivers of performance*, which answers the question of “what is the relative importance of each performance driver’s past in explaining performance variance?” via generalized FEVD (GFEVD)

Granger Causality Tests

Marketing decisions of firms can be informed by sales performance from previous periods and activities of rival brands. Different types of marketing actions can affect each other (e.g., complementary versus substitutive) as well. These issues are called marketing endogeneity, which, if not tested and treated, can lead to misinterpretation of situation and wrong understanding of the effectiveness of marketing (for further detailed discussions, we refer our readers to the ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#) chapter in this book). Granger causality tests (Granger 1969; Hanssens and Pauwels 2016) are needed to examine the existence and direction of causality between model variables. Results from Granger causality tests determine the appropriate functional form of time-series models: for example, a multiple-equation system should be adopted if feedback loops between variables are detected.

The idea of Granger causality tests is that a variable x is considered Granger causing another variable y if the lag values of variable x improve performance of a

model where y is predicted based only on its own past. The most common method to conduct Granger causality tests is by estimating the following regression model of variable y on its own past and lags of variable x :

$$y_t = \alpha + \sum_{i=1}^m \beta_i y_{t-i} + \sum_{j=1}^n \gamma_j x_{t-j} + \varepsilon_t \quad (15)$$

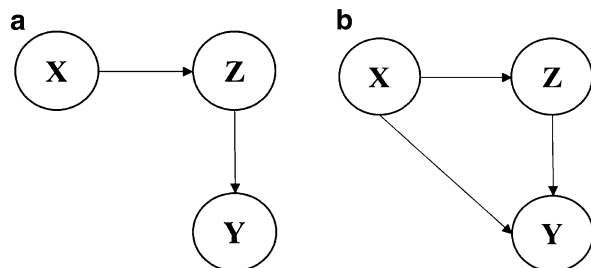
where m and n refer to the maximum lag order for y and x , respectively. β_i and γ_j are the regression coefficients of lag value of y and x , respectively. Once we obtain the regression results, variable x is said to Granger-cause y if any of the γ_j coefficients are statistically significant.

Note that what we described above is pairwise Granger causality test, where only two variables are tested. However, the causality from x to y might be an indirect one that is mediated by another variable z that lies in between. Figure 6 provides a graphical illustration of what we mean here: by conducting only pairwise Granger causality tests repeatedly between variable X and Y , Y and Z , and X and Z , we are not able to distinguish the situation in panel (a) and (b).

We can extend pairwise Granger causality test to the case of n ($n > 2$) variables (“conditional Granger causality test”). To do this, estimate an autoregressive model with n variables. A variable x is said to Granger-cause y if incorporating lagged values of x improves prediction accuracy of y on its own past values and all other ($n - 2$) variables.

There are some applications of Granger causality in the marketing literature that generated some interesting insights (Ilhan et al. 2018). For example, in understanding the relationships between the firm offline marketing mix (distribution, price, and TV advertising), consumer online activities (paid search clicks, website visitations, and Facebook likes), and sales over time, Srinivasan et al. (2016) tested for causalities between variables prior to formal model estimation. The authors found that sales are Granger-caused by all offline and online metrics with varying sensitivities (e.g., the elasticity of paid search clicks on sales is about 5.2 times higher than that of Facebook likes). In particular, TV advertising is not Granger-caused by any of the other marketing mix elements and consumer activities, and hence should play a relatively less influential role in the dynamic system and in impacting sales.

Fig. 6 Pairwise and conditional Granger causality



Cointegration Test

Once we acquire insights about causalities between model variables, we can proceed unit root test and cointegration test to determine the exact form (e.g., level versus first-differenced) in which our variables are included. We have introduced unit root test in section “[ARIMA Models](#),” hence we will focus on cointegration test in this section.

Cointegration test aims to find if two variables of interest are tied in certain long-run equilibrium. For example, one might find that the variations of marketing and sales in the current week seem to be uncorrelated, yet they actually co-move with each other closely in the long run, since it usually takes time for the effect of marketing investments to accumulate and to be reflected on changes in sales. Put it in simple language, two time series, y_t and x_t , are cointegrated if both are non-stationary (with unit root) and if there exists a certain linear combination of the two series that is stationary. It is important to test for cointegration to inform our model choice, since, for example, if marketing expense and sales are evolving and cointegrated, knowing the value of one would enable us to predict that of the other. In this case, a VAR-in-difference model, which essentially deals with change or growth rate instead of the specific values of variables, is not ideal since we will lose valuable information and prediction power.

The cointegrating equation quantifies equilibrium between variables y and x as:

$$y_t = \alpha + \beta \times x_t + \varepsilon_t \quad (16)$$

where ε_t is the equilibrium error that is supposed to be stationary.

There are several ways to test for cointegration; for example, the procedure developed by Engle and Granger (1987) first estimates Eq. (16) via ordinary least squares method, and next test stationarity versus evolution of the error term ε_t . Johansen’s full information maximum likelihood (FIML) is a more popular way for cointegration test (Johansen 1995; Srinivasan et al. 2010). It is a multivariate generalization of the Dicky-Fully unit root test and allows for structural breaks in the relationship among variables.

Vector Autoregressive and Vector Error–Correction Model

The vector autoregressive (VAR) model is an extension of the univariate autoregressive model. It is typically used when we are not only interested in the effect of marketing on performance, but also the feedback of performance on marketing and the effects of marketing activities on each other.

When our data is stationary and without cointegration, we can estimate a VAR model. We focus on reduced-form VAR (obtained from structural VAR) where all explanatory variables are lag values that are predetermined at current time t . This is the form of VAR model that we usually take in analyzing time-series data. For more details on other forms of VAR models and their applications, see the chapter

► “Modeling Marketing Dynamics Using Vector Autoregressive (VAR) Models” in this book.

Specifically, a reduced VAR model can be written as follows:

$$y_t = \alpha + B_1 y_{t-1} + B_2 y_{t-2} + \dots + B_p y_{t-p} + e_t \quad (17)$$

where y_t is an $n \times 1$ vector of n endogenous variables, α is the vector of constant terms including a deterministic time trend and seasonality terms, and B_i , $i = 1, 2, \dots, p$ is an $n \times n$ coefficient matrix of a given lag of order i . e_t is the error term that is contemporaneously correlated. p is the maximum order of lag that is determined via certain statistical criteria that we will explain later.

A VAR system can be efficiently estimated through ordinary least squares (OLS) equation by equation without imposing causal ordering. This is especially important in the context of marketing research where a relatively large number of parameters are modeled together. For example, Pauwels et al. (2016) studied interactions among marketing, eWOM topic, and online and offline store traffic of a retailer by estimating a VAR model with ten endogenous variables. Colicev et al. (2018) estimated 11 endogenous variables in their study that tries to link consumer mindset metrics with firm social media impact and shareholder values.

One can reduce the number of coefficients that need estimation by including some variables as exogenous (instead of endogenous). Such treatment needs to be supported by relevant marketing theories and by appropriate tests. For example, a decision rule developed by De Haan et al. (2016) is to treat variables that do not have a Granger-causal relationship to any other variables as exogenous.

In terms of actual estimation practice, when model variables are stationary, we estimate a VAR in *levels* (e.g., the volume of sales and level of prices). When some variables are found evolving but *not* cointegrated, we can estimate a VAR in *differences*. For example, if we find both sales and online advertising investments are evolving and not cointegrated, our VAR-in-difference model explains the effect of change (growth) in advertising spending on the change (growth) of sales, or the elasticity of ad spending on sales.

When variables are evolving *and* cointegrated, the vector error correction (VEC) model is an extension of VAR to make sure that both the *levels* and *differences* of those cointegrated variables are taken into consideration (see Kireyev et al. 2016). Other extensions of VAR model include panel-VAR (PVAR) model, where cross-sectional heterogeneities are added to the standard VAR. This is as if we incorporate a dependent-variable-specific fixed effect into each equation of VAR. An application of PVAR in marketing can be found in Colicev et al. (2019), where the authors included industry-specific heterogeneity.

Order of Lags in VAR Models

To determine the order of lags in VAR model, we need to trade-off between having a better model fitting and suffering from model complexity. There are several ways to

determine the “best” lag order p of a VAR model. Information criteria (There are several softwares such as EViews, STATA, and R that can help us automatically pick the best order of lags based on information criteria.) are commonly adopted in the vast literature. For example, Akaike Information Criteria (AIC) evaluates model predictive accuracy while imposing a punishment for adding more lags (Akaike 1973). For a VAR model with lag order p , the AIC is formulated as:

$$\text{AIC} = -2LL + 2K \quad (18)$$

where LL refers to the log likelihood function of the model and K refers the number of predictors in the model. K increases with the number of lags incorporated in the VAR system, i.e., the higher the order of lags, the larger the punishment. For example, a VAR system with 3 endogenous variables and lag order of 1 has 9 parameters to be estimated and 18 parameters to be estimated if lag order is 2. We should select the value of p that gives the lowest AIC.

Another criterion is Bayesian Information Criterion (BIC) (Schwarz 1978):

$$\text{BIC}(p) = -2LL + K \ln(T) \quad (19)$$

where T is the sample size. Compared with the AIC, the BIC has a stronger punishment for increasing lag order p , because $\ln(T)$ is greater than two (meaning that $T > 7.39$) in most of the time-series datasets.

Finally, the Hannan-Quinn (HQ) criterion (Hannan and Quinn 1979) also has a stronger punishment on adding lags than the AIC does:

$$\text{HQ}(p) = -2LL + 2K \ln(\ln(T)) \quad (20)$$

Specifically, the BIC and the HQ tend to give consistent results as sample size T approaches infinity.

Generalized Impulse Response Functions

Result interpretation is relatively straightforward for some models like multiple linear regressions, where the effect of an explanatory variable on the outcome variable is simply quantified by the corresponding coefficient (if statistically significant). However, interpreting VAR model results directly from model outputs is not easy due to multicollinearity issues and feedback loops between variables, to name a few. In a word, given an intertwined dynamic system where variables are interrelated in various ways, we need a technique that can get us the “net” effect of each variable. Further, we are not only interested in the short-term effect, but also the long-term impact of model variables that can help us plan for the future.

Impulse–response functions (IRFs) can help us by simulating the overtime impact of a change to a variable on the whole dynamic system (Bronnenberg et al. 2000; Pauwels et al. 2016).

To derive an IRF, we start by substituting each lag of each endogenous variable in the reduced-form VAR model in Eq. (14) using the same equation ($y_t = \alpha + B_1 y_{t-1} + B_2 y_{t-2} + \dots + B_p y_{t-p} + e_t = \alpha + B_1(\alpha + B_1 y_{t-2} + B_2 y_{t-3} + \dots + B_p y_{t-p-1} + e_{t-1}) + B_2(\dots) + B_p(\dots) + e_t$, and so on). We can then express the right-hand side of Eq. (17) as a function of only contemporaneous and lagged values of the error terms. We call such expression the vector moving average (VMA) representation:

$$y_t = \rho + e_t + A e_{t-1} + A^2 e_{t-2} + \dots + A^t e_0 \quad (21)$$

Interpreting Eq. (21), each endogenous variable is explained by a weighted average of current and past errors, or “shocks” both to itself and to the other endogenous variables.

An IRF tracks the impact of a shock to each variable in the system during the shock (period 0) and each period afterwards (period 1, 2, etc.). Most of the times the effects of shocks will die out (i.e., converge back to its steady-state, or pre-shock level), and this usually happens to established and mature brands (e.g., effect of advertising investment on sales). However, in some rare cases we may observe that a shock has a permanent impact. Such pattern is more likely to be observed among young and innovative brands. We usually name the effect of a shock in period zero the *contemporaneous* or *immediate effect*, and the cumulative effect of a shock from period one onwards as the *long-term effect*.

A critical limitation of IRF is that it requires a causal ordering for the immediate effects. For example, when trying to model a dynamic system consisting of weekly brand sales, price, online and offline marketing expenditure, and consumer traffic, we are expected to clearly understand the causal sequence between these variables. However, it is typically unclear, for example, whether online marketing expenditure should precede or follow offline marketing expenditure in leading to sales. Generalized IRFs (GIRFs) are useful when theories or knowledge do not inform us with such ordering.

Figure 7 shows an example of impulse response of a shock in firm’s advertising effort in week 0 on sales performance. The horizontal axis in Fig. 7 represents weeks 0 to 15, and the vertical axis is the coefficient of IRF analysis. An immediate incremental effect of around 5 in week 0 is the highest across all weeks. The cumulative or long-run effect is measured by the shaded area under the curve in Fig. 7, which is approximately 8.4. Finally, brand sales seem to revert to its steady state, with incremental impact staying at zero from week 11 onwards. The permanent effect of the shock to advertising is hence zero. It is very important for researchers to interpret the incremental effect of a shock in IRF or GIRF analysis results. The incremental effect turning zero does not mean zero sales, but instead no *additional* sales.

In the recent marketing literature, Pauwels et al. (2016) applied GIRF to examine the short- and long-term elasticities of different electronic word-of-mouth (eWOM) on online and offline store traffic. They found that the long-term elasticity of brand-related eWOM is twice as high as that of advertising-related eWOM in driving up offline store traffic. While offline and online traffic is approximately equally affected by purchase-related eWOM in the short run, yet its impact on the former in the long run is 16 times higher than that on the latter.

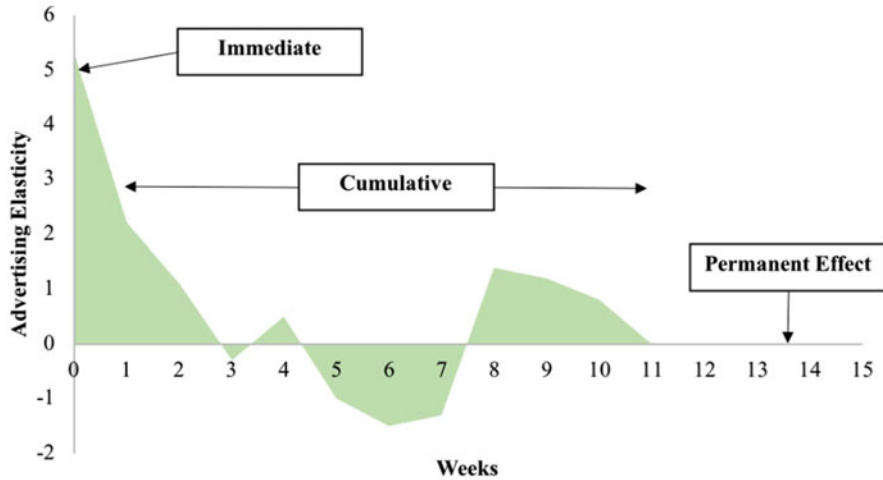


Fig. 7 Example of an impulse response of advertising on sales

Firms can also rely on (G)IRF to improve their marketing resource allocation and improve ROI. For example, Pauwels (2004) investigate how firms should allocate efforts between new product introduction (NPI) and sales promotion to maximize firm value. IRF result shows that sales promotion has a higher short-run elasticity of firm value, whereas its long-term impact turns negative (from 0.12 to -0.78). In contrast, NPI turns out to have persistent and positive impact on firm value (0.02 in the short run and 1.14 in the long run). Inspired by such result, firms should allocate more resource on new product introduction, even though sales promotion can bring them immediate sales boost.

Generalized Forecast Error Variance Decomposition

Analogous to a “dynamic R^2 ,” generalized forecast error variation decomposition (GFEVD) shows the relative importance of each VAR variable in contributing to the variation in the performance variable. For example, we can use GFEVD to determine what is the contribution of online display ad, among in-store promotion, social media marketing, and search engine optimization, to the variation in brand weekly sales. Compared with FEVD, GFEVD is more widely adopted by marketing literature in that it does not impose causal ordering between variables.

The GFEVD is given by:

$$\theta_{ij} = \frac{\sum_{l=0}^n (\psi_{ij}(l))^2}{\sum_{j=1}^m \sum_{l=0}^n (\psi_{ij}(l))^2}, j = 1, 2, \dots, m \tag{22}$$

where $\psi_{ij}(l)$ is the value of a GIRF following a one standard error shock to variable j on variable i at period l . GFEVD allows an initial shock to affect all other endogenous variables instantaneously (i.e., the coefficient for period zero of other variables can be nonzero).

Judging from Eq. (22), θ_{ij} is a percentage term, and that all the θ_{ij} 's always sum up to 100%. It is typical to find that most of the variance of a variable is explained by its own past, which is referred to as “inertia” (e.g., price inertia, see Nijs et al. 2007). Panel(a) in Fig. 8 shows an example of analysis on contribution to variation in sales of firm’s past sales, online advertising, and offline advertising effort. The contribution of past sales (i.e., inertia) contributes the most (70%), while online advertising ranks the second (20%) and offline advertising the last (10%). In contexts where inertia is of little interest, researchers can take it out and have a better visualization of the relative importance of *other* variables. Panel (b) in Fig. 8 shows GFEVD results without inertia. The relative contribution of two advertising channels remains the same (2:1).

Continuing the example of resource allocation in firm value maximization that we raised in 3.5, the authors contrasted contribution of sales promotion and NPI to firm value FEVD (see Fig. 2 in Pauwels et al. 2004, pp. 151). The gap between contribution of NPI and sales promotion gets wider as time goes, with the former turning eight times greater than the latter in two quarters’ time.

Other applications of FEVD and GFEVD include the work of Srinivasan et al. (2016), where the authors examined the contribution to sales growth of traditional marketing mix variables and online customer activity metrics. Without considering sales inertia, the authors found that distribution is subject to 60% of the volume variance, while only 2% for online paid search.

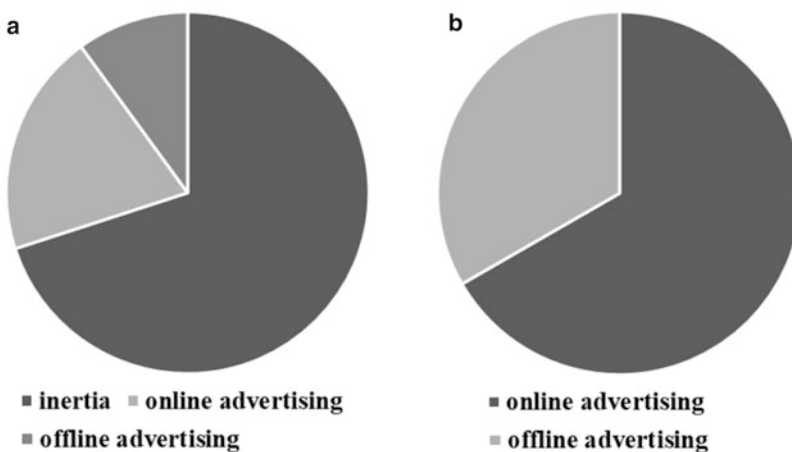


Fig. 8 Example of GFEVD of sales with and without inertia

Volatility Models

Time-series data can be described from many angles, among which the mean and the volatility (or variance) are the two most typical metrics. While the mean (or expected value) of a time series captures the average level of the data, variance captures the level of turbulence, namely the level of fluctuation of the series around its mean (Franses and van Dijk 1996, 2000). In this section, we introduce a model that deals with the assumption of *constant conditional variance*.

Let us again use weekly sales as an example. The models that we have discussed so far (e.g., ARIMA) estimate the expected weekly sales conditional on different marketing variables (hence “conditional mean”). These models assume that the volatility of weekly sales stays constant over time. However, in the real world, there are plenty of examples where such an assumption does not hold; instead, the level of volatilities across periods could be related. For example, if sales have been volatile during the past two weeks, then it is very reasonable to expect that sales in the coming week are going to be more volatile than usual as well. Additionally, an exogenous event could also shock sales by significantly increasing volatility for a certain period of time.

Figure 9 shows an example of a brand’s weekly sales (in thousands) over a time span of 160 weeks. The vertical dashed line refers to the time (week 73) when the brand launched a new product. One might first notice that the average sales increased greatly after the new product introduction (NPI) as expected (i.e., from 50,163 to 148,255). Moreover, the volatility of brand sales (measured by standard deviation) also rose greatly from the pre- to post-NPI period (i.e., from 179,209.1 to 47,904.47). When trying to model weekly sales, it is important to at least incorporate a step change in both mean and volatility after week 73.

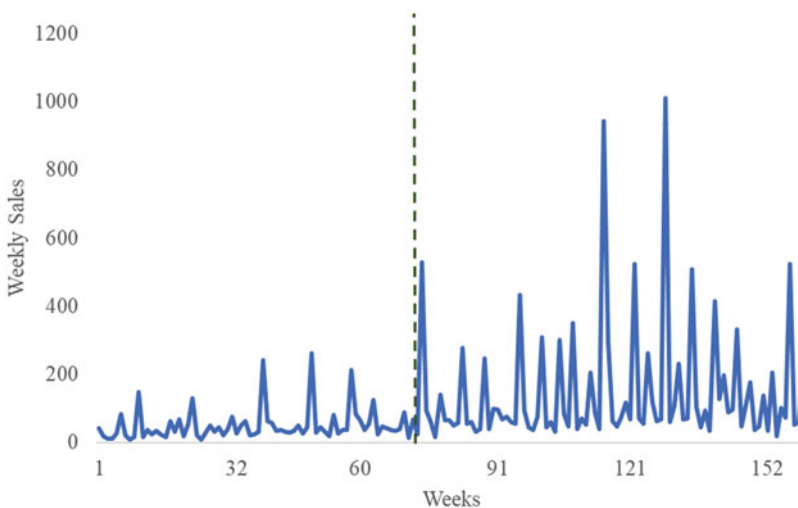


Fig. 9 Weekly sales with time-varying volatilities

From the practical point of view, firms care about performance volatility, since a highly uncertain future is hard to plan for. With high sales volatility, refining the marketing mix based on historical ROMI is not meaningful. Further at the retailer level, a high level of sales volatility of a brand or a product category leads to difficulties in inventory management (Esteban-Bravo et al. 2017). Finally, firms with volatile performance are usually deemed to have higher idiosyncratic risk, which can be harmful for firm valuation and stock market performance (Fischer et al. 2016). Hence, models that describe, explain, and predict volatilities are of great value and importance.

The focus of this section, generalized autoregressive conditional heteroscedasticity (GARCH) model (Other variations of GARCH volatility models that are extensively used in the literature include VEC-GARCH models (Bollerslev et al. 1988), constant conditional correlation (CCC)-GARCH model (Yildirim et al. 2020), and BEKK model (Esteban-Bravo et al. 2017)), incorporates dynamics in data volatility by recognizing a time-varying conditional variance (Engle et al. 1987; Fischer et al. 2016). A GARCH model first determines the conditional mean of a series and then the volatility. It estimates conditional variances of data series in an explicit way, similar with how conditional mean is estimated by ARIMA model.

Let us start by introducing the ARCH specification and then generalizing it to GARCH. Consider again sales data of a brand that is described by an AR(p) process:

$$s_t = \varphi_0 + \varphi_1 s_{t-1} + \dots + \varphi_p s_{t-p} + \varepsilon_t \tag{23}$$

with $\sum_{i=1}^p \varphi_i < 1$.

Let us assume that the squares of the error term in Eq. (23) can be captured by an AR(q) process:

$$\varepsilon_t^2 = \eta_0 + \eta_1 \varepsilon_{t-1}^2 + \dots + \eta_k \varepsilon_{t-q}^2 + \nu_t \tag{24}$$

where ν_t is white-noise variables with $E(\nu_t) = 0$ and $E(\nu_t, \nu_{t+m}) = 0$ for nonzero m 's and $E(\nu_t, \nu_{t+m}) = \sigma^2$ for $m = 0$. The representation of the white-noise process in Eq. (24) is called an ARCH(q) process. Note that the conditional variance of ε_t^2 varies by time, whereas the unconditional variance is constant and is given by $\sigma^2 = E(\varepsilon_t^2) = \frac{\eta_0}{1 - \eta_0 - \dots - \eta_q}$.

When fitting a model, a linear representation of ARCH process in Eq. (24) is not always efficient; a more common representation of ε_t is written as:

$$\varepsilon_t = \sqrt{h_t} z_t \tag{25}$$

where z_t is an i.i.d process with zero mean and unity variance, and

$$h_t = \eta_0 + \eta_1 \varepsilon_{t-1}^2 + \dots + \eta_k \varepsilon_{t-q}^2 \tag{26}$$

The generalized ARCH, or GARCH, model represents h_t as a function of its own past values and past values of ε_t^2 :

$$h_t = a + b_1 h_{t-1} + \dots + b_m h_{t-m} + \eta_1 \varepsilon_{t-1}^2 + \dots + \eta_k \varepsilon_{t-q}^2 \quad (27)$$

If a process ε_t is generated by a process described in Eq. (27), then ε_t is a *GARCH* (m, q) process.

Beyond the univariate specifications that we discussed above, we can use GARCH model, combined with VAR model, to estimate multiple endogenous variables. For example, the work of Esteban-Bravo et al. (2017) recognizes that not only sales volatility but also covolatilities (i.e., conditional covariance) between sales and marketing actions are time varying. Using a VAR-BEKK model, the authors generated fresh insights for managers to deal with performance volatility that is often overlooked by prior research. van Diejen et al. (2019) examined the interaction between volatility in volume of firm-related user-generated content (UGC) and volatility in firm stock return. The authors estimated a multivariate GARCH model and found significant cross-effect between UGC growth and stock returns. Further, the authors discovered new product launch events as a driver of UGC growth volatility, though the exact direction of impact (i.e., an increase or a decrease in volatility) is determined by the specific UGC content.

Conclusion

Time-series models are great tools for researchers and practitioners to tackle marketing problems. These models, especially modern dynamic systems, are also quite powerful in generating new insights by bridging dynamics between factors that are previously overlooked. This chapter introduces traditional and modern time-series analytics such as ARIMA and VAR models. We also discuss model applications in marketing such as evaluating return on marketing investments, measuring elasticities of marketing activities, and refining allocation of marketing resources, to name a few. As the field evolves, researchers are adopting a broader range of models to explore marketing challenges. For instance, recent research has emerged using Markov chain models to solve sales attribution problems. We hence expect further methodological advancements in time-series modeling in marketing and highlight the importance of reviewing this domain of research from time to time.

Cross-References

- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Measuring Sales Promotion Effectiveness](#)
- ▶ [Modeling Marketing Dynamics Using Vector Autoregressive \(VAR\) Models](#)
- ▶ [Return on Media Models](#)

Appendix

Software Application

The purpose of this software application section is to show our readers how analysts and modelers tackle real-world marketing problems using time-series models like ARIMA and VAR that are covered in this chapter. We hereby introduce R (You might want to know more about this powerful analytical tool via <https://www.r-project.org/>), an open-source and free software for statistical computing and data analysis. It is widely adopted by academics and industry practitioners as a powerful analytical tool to facilitate them when dealing with data-rich challenges.

Let's start by walking into the following scenario:

ABC company operates in the kitchen appliance industry in an emerging market. The company so far has focused all its marketing efforts on offline flyer advertising and online Google AdWords. However, recent performance reports showed that ABC's sales have not been reaching the management's expectations. The CMO, in preparation for a meeting with the CEO and CFO, is keen to know if and to what ABC's sales will look like in the next quarter. Further, he/she wonders to what extent marketing expenditures are effective in driving up sales. The CMO is also curious if there's any potential for ABC to optimize its current marketing budget allocation.

As the director of the marketing analytics department, you are presented with ABC's historical weekly sales and marketing expenditure on flyer advertising and Google AdWords advertising over a time span of 122 weeks (i.e., 122 observations). Having met with the CMO, you make a summary of the questions to be answered and your action plans as follows:

(a) ***What would be the forecast of demand for the next 12 weeks?***

We are going to predict future sales using two approaches: ARIMA and multiple linear regression (MLR) and compare their estimation and prediction results.

(b) ***What drives sales in the long run? What is the contribution of each marketing action to sales (i.e., return on marketing investment)?***

We are going to estimate a VAR model and perform FEVD to evaluate the relative importance of AdWords, flyers, and sales inertia played in determining long-run sales.

(c) ***To what extent do AdWords and flyers impact sales in the short versus long run?***

We are going to perform IRF analysis to evaluate the short- and long-run elasticities of each marketing action.

(d) ***How should ABC allocate marketing budget between AdWords and flyers to get the best result?***

We are going to use long-run elasticities of AdWords and flyer marketing to determine the optimal resource allocation scheme for ABC.

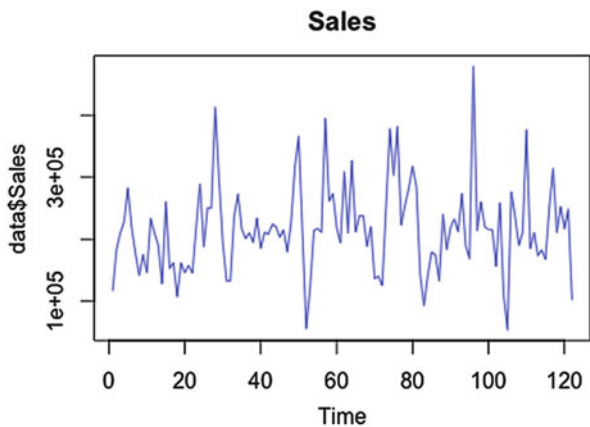
Data Visualizations

Here is a brief preview of the first 10 rows of our dataset:

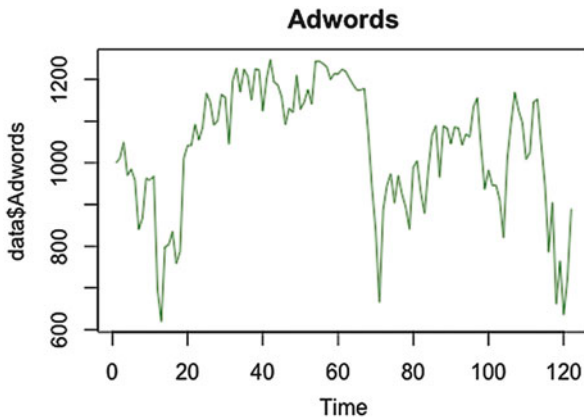
Week	Sales	Flyer	Adwords
1	115792	0.00	999.860
2	183189	17134.21	1011.395
3	210107	0.00	1049.560
4	227392	0.00	969.570
5	282095	0.00	984.850
6	218587	0.00	958.400
7	177226	12079.39	839.670
8	140803	0.00	867.460
9	174974	0.00	962.820
10	145961	0.00	958.420

As you can observe, the firm so far has been having a relatively stable expenditure on AdWords (around 900 each week), while that on flyers it has been much more fluctuating. For example, during the first 10 weeks, the firm spent 17134.21 in week 2 and 12079.39 in week 7, and nothing for the rest 8 weeks. To get a feel of the data patterns, it is a good practice to visually inspect them by plotting sales, flyer expenditure, and Google AdWords expenditure, respectively using *ts.plot*.

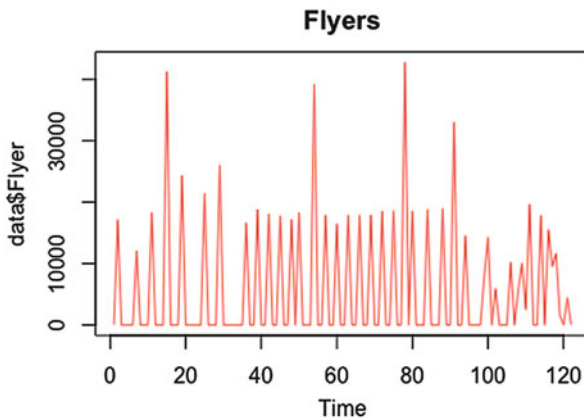
```
ts.plot(data$Sales, col="blue", main="Sales")
```



```
ts.plot(data$Adwords, col="darkgreen", main="Adwords")
```



```
ts.plot(data$Flyer, col="red", main="Flyers")
```



ARIMA Modeling

As elaborated in this chapter, we are going to estimate and predict sales using ARIMA following the procedure summarized below:

- Perform **unit root tests** to check for nonstationary variables and take differences of the variables that are evolving.

- Plot **ACF** and **PACF** to determine the order of lags and hence the specification of ARIMA.
- Split the data into **training** and **testing** set and estimate an **ARIMA** model using the training set and predict sales using the testing set.

Additionally, we will also estimate a **multiple linear regression** (MLR) model using the training set and predict sales using the testing set. This is to compare model performance using ARIMA and MLR method.

Log Transformation

First, through the time-series plots, we observe a high level of data turbulence (or volatility), which, if not treated properly, will lead to false model results and interpretations. It is a typical practice to take the logarithm of each variable to smooth out the series as a preliminary step:

```
data$LSales <- log(data$Sales+1)
data$LAd <- log(data$Adwords+1)
data$LFlyer <- log(data$Flyer+1)
```

Note that we add 1 to each variable during log transformation to avoid having log (0), which equals to negative infinity.

Stationary Tests

It is critical for analysts to make sure that data series being modeled are all stationary (instead of evolving) in order to have reliable model results. As introduced in the chapter, there are multiple tests for series stationarity, including the ADF, KPSS, and Phillips-Perron test that can be executed using R function *adf.test*, *kpss.test*, and *pp.test*, respectively. Here in this section, we demonstrate the procedure of using the ADF test. Under the ADF test, the null and alternative hypotheses are:

- H_0 : The data is not stationary
- H_1 : The data is stationary

Note that for *adf.test* and *pp.test*, we can reject the null hypothesis that the variable is not stationary (i.e., with a unit root) if the p-value is smaller than a certain significance level; yet *kpss.test* works in the opposite way, i.e., the null hypothesis is that the series is stationary without a unit root.

To check for stationarity, we need to first let R know that weekly sales and AdWords and flyer expenditures are time series using *ts* function, and then perform ADF test using *adf.test* function:

```

LSales <- ts(data$LSales, frequency = 52, start = c(1, 1))
LAd <- ts(data$LAd, frequency = 52, start = c(1, 1))
LFLYER <- ts(data$LFLYER, frequency = 52, start = c(1, 1))

adf.test(LSales)

```

```

##
## Augmented Dickey-Fuller Test
##
## data:  LSales
## Dickey-Fuller = -5.2428, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary

```

```
adf.test(LAd)
```

```

##
## Augmented Dickey-Fuller Test
##
## data:  LAd
## Dickey-Fuller = -2.0661, Lag order = 4, p-value = 0.5491
## alternative hypothesis: stationary

```

```
adf.test(LFLYER)
```

```

##
## Augmented Dickey-Fuller Test
##
## data:  LFLYER
## Dickey-Fuller = -6.5352, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary

```

Test results inform us that LAd series (i.e., log-transformed AdWords spending) is evolving with p-value greater than 0.05. We need to take the first-difference of this series to make it stationary. Note that once we first-difference a log-transformed series, the interpretation will be different: now the series refers to growth of weekly AdWords spending, rather than AdWords spending itself.

To take the first-difference of a series, we use R function *diff*:

```

#Take the first difference of Adwords spending series

DLAd <-diff(LAd, differences = 1)

```

Now we can perform ADF test again to make sure that all variables are stationary:

```
adf.test(LSales)
```

```
##  
## Augmented Dickey-Fuller Test  
##  
## data: LSales  
## Dickey-Fuller = -5.2428, Lag order = 4, p-value = 0.01  
## alternative hypothesis: stationary
```

```
adf.test(DLAd) #first-differenced series
```

```
##  
## Augmented Dickey-Fuller Test  
##  
## data: DLAd  
## Dickey-Fuller = -5.0641, Lag order = 4, p-value = 0.01  
## alternative hypothesis: stationary
```

```
adf.test(LFLYER)
```

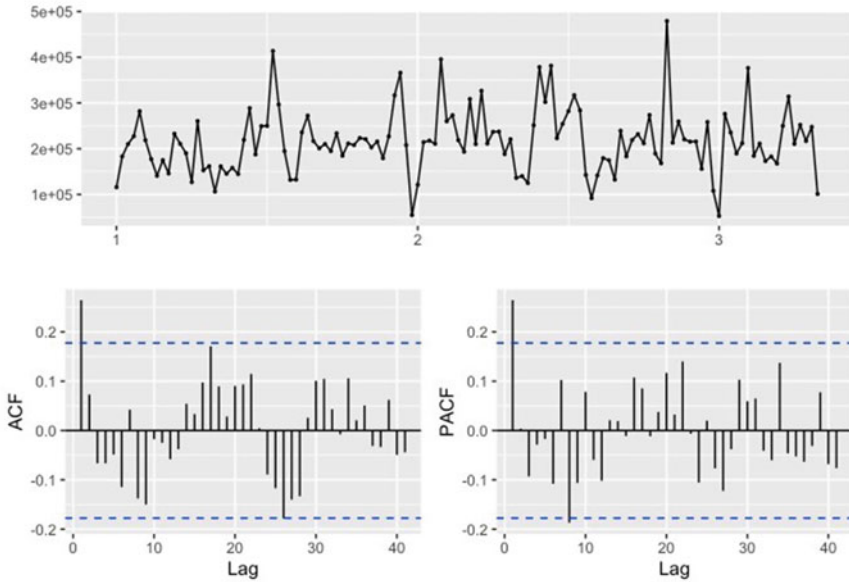
```
##  
## Augmented Dickey-Fuller Test  
##  
## data: LFLYER  
## Dickey-Fuller = -6.5352, Lag order = 4, p-value = 0.01  
## alternative hypothesis: stationary
```

Stationary test results suggest that now the three variables (with AdWords spending first-differenced) are all stationary. Note that to construct ARIMA model, we only need sales series, while all series are needed for MLR and later VAR model.

ACF and PACF for Order of Lags

Once a time series has been stationarized, a systematic way to determine the order of lags of the autoregressive (AR) and moving average (MA) components of ARIMA model is to plot and inspect ACF and PACF. Here we use R function *gtsdisplay*, which can generate (i) the plot of the series over time, (ii) the ACF plot, and (iii) the PACF plot simultaneously and automatically:

```
Sales <- ts(data$Sales, frequency = 52, start = c(1, 1))
ggtsdisplay(Sales) #trend plot and ACF and PACF.
```



Here on the ACF and PACF plots, the dashed horizontal lines represent the critical region (95% confidence level) for the lags. The lag order of the AR and MA component is identified by the number of lags where the PACF and ACF plot displays a clear cutoff, respectively. Here we find both ACF and PACF have a cutoff at lag 1, indicating that we should probably take a lag order of 1 for both MA and AR components. Further, given that we did not take the difference of the sales series, the final specification of our model is ARIMA (1,0,1), or ARMA (1,1).

Construct and Estimate an ARIMA Model

Splitting the Data

To estimate sales and examine model predicting power, we cannot exploit the entire data to construct our model. Instead, we need to split the series into training (in-sample) and testing (out-of-sample) sets. To do this, we apply the most commonly adopted 80–20 scheme, namely we use the first 80% of the observations as the

training set and the rest 20% as the testing set. Given that we have 122 observations in total, we should use the first 96 observations for our training set, and the rest 25 observations as the testing set.

```
#Splitting the data into training and testing sets

train <- data[1:96,]
test <- data[97:122,]
```

Train the Model

ACF and PACF suggest that we estimate an ARIMA (1,0,1) model. To estimate the model using the training set, we use R function *Arima*:

```
#Estimating the ARIMA model using our training set:

fit_arima <- Arima(ts(LSales[1:96], frequency = 52), order = c(1,0,1), seasonal = c(1,0,1))

summary(fit_arima)
```

```
## Series: ts(LSales[1:96], frequency = 52)
## ARIMA(1,0,1)(1,0,1)[52] with non-zero mean
##
## Coefficients:
##      ar1      ma1      sar1      smal      mean
##      0.3046  0.0941  0.4731  -0.0771  12.2368
## s.e.      NaN  0.0477      NaN  0.0634  0.0585
##
## sigma^2 estimated as 0.08607: log likelihood=-20.81
## AIC=53.62  AICc=54.57  BIC=69.01
##
## Training set error measures:
##              ME      RMSE      MAE      MPE      MAPE      MASE
## Training set 0.01038504 0.2856303 0.2119471 0.0265656 1.738597 0.8070106
##              ACF1
## Training set -0.0234848
```

Furthermore, you may estimate ARIMA models with different specifications and compare model performances (e.g., AIC and BIC) to pick the best specification. There are also R functions that can automatically pick the specifications with the lowest AIC and BIC, for example, *auto.arima*. However, analysts should keep in mind that you are the constructor of your model and that it is you that should be the final decision-maker on what model to estimate by considering managerial and strategic factors that model diagnostics cannot inform you. For example, some firms might operate within a certain cycle and would want to evaluate sales using a specific order of lags.

Estimate a Multiple Regression Model

In addition to ARIMA, given a dataset as such, it is also very common for modelers to adopt multiple linear regression method and estimate a linear model to fit and predict sales. This is because MLR allows us to incorporate other exogenous factors, while ARIMA typically only involves the endogenous variable itself.

Again, we need to use the training set for model estimation. To do this, we use the *lm* function in R, referring to “linear model.”

```
fit_lm <- lm(LSales ~ lag_Sales + DLAd + LFLYER, data = train)
summary(fit_lm)
```

```
##
## Call:
## lm(formula = LSales ~ lag_Sales + DLAd + LFLYER, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30041 -0.14924  0.01009  0.15514  0.94615
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.445228   1.165853   6.386 6.84e-09 ***
## lag_Sales    0.389160   0.095216   4.087 9.33e-05 ***
## DLAd         0.088746   0.388587   0.228  0.8199
## LFLYER       0.012172   0.007227   1.684  0.0955 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3095 on 92 degrees of freedom
## Multiple R-squared:  0.1743, Adjusted R-squared:  0.1474
## F-statistic: 6.473 on 3 and 92 DF,  p-value: 0.000504
```

Interpreting the results briefly, a 1% increase in lag sales will lead to 0.39% of increase in current sales; a 1% increase in flyer spending will lead to 0.01% increase in sales. The coefficient of “DLAd” is statistically insignificant.

Validation Set Assessment: ARIMA Versus MLR

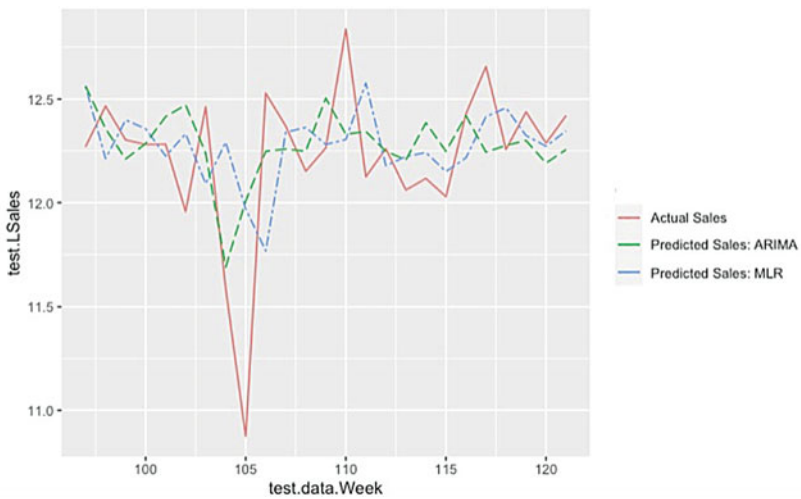
Now it is time for us to contrast model performance using ARIMA and MLR method and predict sales using the testing set. We further plot the actual sales and predicted sales using ARIMA and MLR method, respectively, on the same graph for contrast.

```
# Predict sales based on ARIMA model estimations
predict_arima <- predict(fit_arima, order=c(1,0,1), data = test, n.ahead = 26)

# Predict sales based on Multiple Linear regression model estimations
predict_lm <- predict(fit_lm, test)

#Plot the actual and predicted sales on the same graph
data_frame <- data.frame(test$Week,test$LSales,predict_lm, predict_arima$pred)

ggplot(data_frame, aes(data_frame$test.Week)) +
  geom_line(aes(y = test.LSales, colour="Actual Sales")) +
  geom_line(aes(y = predict_lm, colour="Predicted Sales: MLR")) +
  geom_line(aes(y = predict_arima$pred, colour = "Predicted Sales: ARIMA"))
```



From the graph, both models can mimic (to a certain extent) the pattern of actual sales in the testing set. To determine which method does a relatively more accurate job, we can calculate and compare the root-mean-square deviation (RMSE) of both predictions.

```
# Calculate RMSE of prediction results using MLR and ARIMA method.

rmse.predict_lm <- (sum((test$LSales - predict_lm )^2)/25)^0.5
rmse.predict_arima <- (sum((test$LSales - predict_arima$pred )^2)/25)^0.5

rmse.predict_lm

## [1] 0.3968784

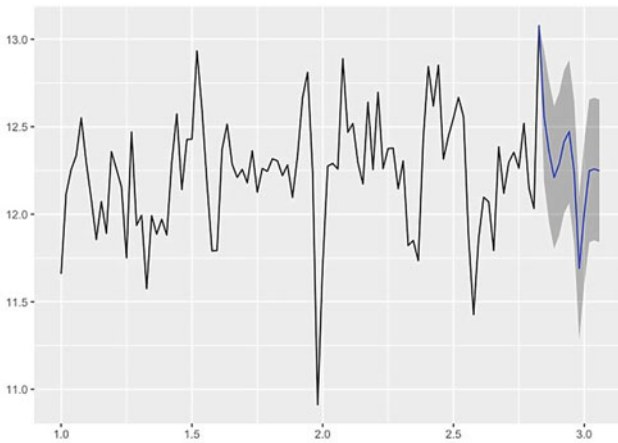
rmse.predict_arima

## [1] 0.3255406
```

Comparing prediction accuracy using both methods, we find that ARIMA managed to capture the dynamics of ABC's sales better, since its RMSE (0.33) is slightly lower than that of MLR (0.40).

To improve the model prediction accuracy of MLR, several other factors should be considered, for example, the weekly price of ABC, and weekly sales of ABC's core competitors. Modelers can incorporate additional variables into the model depending on data availability. Furthermore, answering the first question raised by the CMO, we can predict sales of ABC for the next 3 months (12 weeks) based on our ARIMA model. Here we plot both predicted sales and 95% confidence intervals in the graph below:

```
#Plot predicted sales in the next 3 months using ARIMA method
fit_arima %>% forecast(h=12) %>% autoplot()
```



Reverting the log-transformed predicted values back, we get the predicted sales for the next quarter (i.e., 12 weeks) as below:

```
#predicted sales
forecast<-fit_arima %>% forecast(h=12)
forecast_sales <- exp(forecast$mean)-1
forecast_sales
```

```
## Time Series:
## Start = c(2, 45)
## End = c(3, 4)
## Frequency = 52
## [1] 285957.6 232547.9 200872.0 216504.4 246491.8 260936.2 206717.2 119479.9
## [9] 164058.7 208617.2 210776.9 208660.7
```

VAR Model Steps

Estimating a VAR Model

We are able to set up our VAR model relatively easily since we have already performed model diagnostics on series stationarity through unit root tests in section

“Testing for Evolution Versus Stationarity.” Taking all three variables as endogenous variables, we estimate a VAR model consisting of (log-transformed) weekly sales, lagged Google AdWords expenditure, and flyer expenditure using *VAR* function. We then summarize the results in the table below.

```
#Build a dataset for VAR model
data.ts.dl <- window(cbind( DLAd, LFLYER,LSales), start = c(1, 2))

#VAR estimation
varp <- VAR(data.ts.dl, ic="AIC", lag.max=1, type="const", season=4)

#Summarize and present VAR results
lmp <- varp$varresult
stargazer( lmp$DLAd, lmp$LFLYER,lmp$LSales, column.labels = c( 'DLAd', 'LFlyer', 'LSales'), type = "text", dep.va
r.labels.include = FALSE )
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               DLAd   LFlyer   LSales
##                               (1)    (2)     (3)
## -----
## DLAd.l1                       -0.154   -4.150   0.402
##                               (0.094)  (4.453) (0.339)
##
## LFLYER.l1                      -0.00002 -0.276** 0.003
##                               (0.002)  (0.092) (0.007)
##
## LSales.l1                      -0.062** 0.298   0.310**
##                               (0.026)  (1.215) (0.092)
##
## const                          0.751** 0.444   8.436**
##                               (0.313)  (14.814) (1.127)
##
## sd1                             -0.020   0.816   -0.027
##                               (0.024)  (1.153) (0.088)
##
## sd2                             -0.033   1.077   -0.026
##                               (0.024)  (1.154) (0.088)
##
## sd3                             -0.014   0.673   -0.019
##                               (0.024)  (1.153) (0.088)
##
## -----
## Observations                    120     120     120
## R2                              0.083   0.096   0.107
## Adjusted R2                     0.034   0.048   0.059
## Residual Std. Error (df = 113)  0.094   4.449   0.339
## F Statistic (df = 6; 113)       1.703   1.995*  2.248**
## =====
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

From the VAR output, we find that:

- Direct effects: AdWords (0.402) and flyer (0.003) both have positive direct impact on sales.
- Carryover effects: past AdWords (−0.154), flyer advertising (−0.276), and sales (0.310) all exert impact on their current values, respectively.

- Feedback effects: sales have positive feedback effect on flyer (0.298) and ad spending (0.310), while negative feedback effect on online AdWords spending (-0.062).

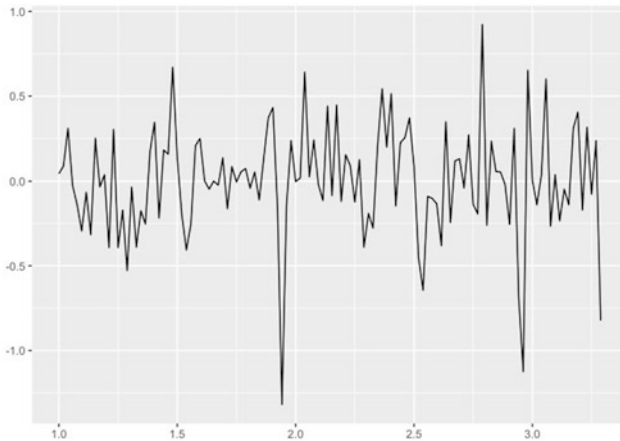
Note that due to the limited sample size and data variation in our sample, some of the coefficients seem statistically insignificant. However, to see the effect of AdWords and Flyer advertising on Revenues over time (e.g., immediate and long-term effects), it is more rigorous to refer to results from impulse response function (IRF) analysis.

After the estimation it is always good practice to check the residuals' normality and the autocorrelation. If there is any misspecification, you may need to search if any anomaly such as outlier and structural break occurs. Here we plot the residuals and inspect their mean.

```
sales.residuals <- data.frame(residuals(varp))$LSales
sales.residuals <- ts(sales.residuals, frequency = 52, start = c(1, 1))
round(mean(sales.residuals),4)
```

```
## [1] 0
```

```
autoplot(sales.residuals)
```



We observe that the residuals seem to vary randomly around zero, with a mean of zero.

Forecast Error Variance Decomposition

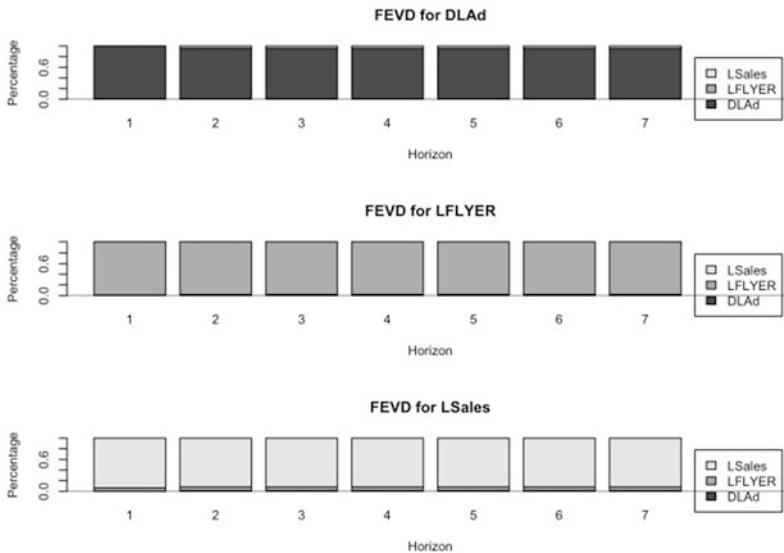
Referring to the second question raised by the CMO, we perform FEVD analysis to evaluate and visualize the relative importance or contribution of flyers, AdWords, and sales inertia using R function *fevd*:

```
fevd <- fevd(varp, n.ahead=7)
round(fevd$LSales, 4)
```

```
##      DLAd LFLYER LSales
## [1,] 0.0005 0.0647 0.9349
## [2,] 0.0109 0.0697 0.9194
## [3,] 0.0110 0.0697 0.9193
## [4,] 0.0110 0.0697 0.9192
## [5,] 0.0110 0.0697 0.9192
## [6,] 0.0110 0.0697 0.9192
## [7,] 0.0110 0.0697 0.9192
```

The table above indicates that ABC’s sales are quite “sticky” in the sense that lagged sales (LSales) contribute to more than 90% to changes in current sales. Offline marketing seems to play a more important role than online AdWords for ABC. The table above corresponds to the bottom panel of the graph below.

```
plot(fevd)
```



IRF Analysis

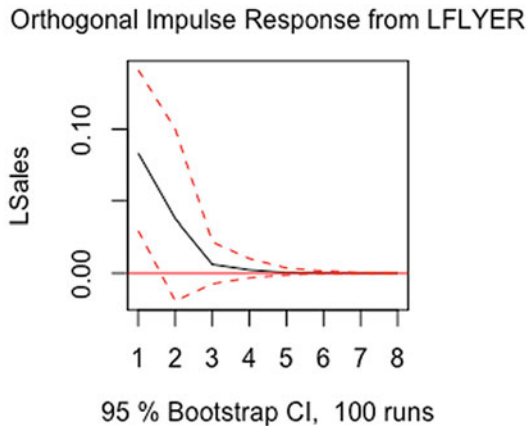
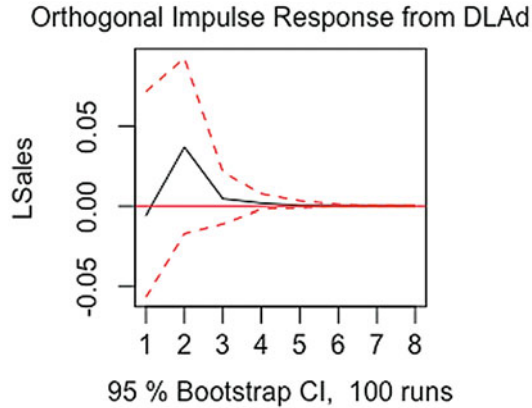
Responding to the third question raised by the CMO, we perform IRF (In this session we are using orthogonalized impulse reaction function for estimation. In R environment, to implement GIRF estimations, we need to estimate a Bayesian VAR model (you may check package “bvartools”) instead, which is beyond the scope of this chapter.) analysis to evaluate the short- and long-run elasticities of flyers and AdWords marketing of ABC using *irf* function in R.

```

irfs <- irf(varp, impulse = c('DLAd', 'LFLYER'), response = 'LSales',
           runs = 100, n.ahead = 7, ortho = TRUE, ci=0.95)

plot(irfs)

```



IRF plots help us visualize when the peak effects occur. On the plots, the solid line refers to IRF coefficients, while the dashed lines refer to lower and upper bound of the IRF coefficient's confidence interval. It seems that increase flyer spending can cause an immediate boost of sales; in contrast, it takes longer time for spending on AdWords to have positive impact on sales. Moreover, we can observe that these impacts all decay fast and gets close to zero over time, mostly within 6 periods (weeks).

Immediate and Long-Term Effects

In order to compute the immediate and long-term effects, we need to evaluate the significance of each IRF coefficient. If the t-statistics of the IRF coefficient is greater than 1 (Here we follow previous research (e.g., Slotegraaf and Pauwels 2008) to set the criteria as $t > 1$. You may apply the $t > 2$ rule if you would like to evaluate coefficient significance at a 95% significance level) ($t > 1$), we treat it as significant and keep the value of that coefficient; otherwise, we treat the coefficient as zero. To calculate the t-statistics, we need to derive the standard error (se) of each coefficient from its confidence interval, since lower bound_{ci} = $\beta - 1.96 * se$, and upper bound_{ci} = $\beta + 1.96 * se$. We then calculate the t-statistics using $t\text{-stat} = \beta / se$.

Based on the above computations, the first period impact is called the immediate effect while the cumulative effect over 8 periods is called the long-run effect.

Now we make a table in R to summarize IRF coefficients and their confidence intervals. You will see in the output that response means the response value at a particular period (there are 8 periods in total), lower and upper refer to the lower and upper bound of the corresponding confidence intervals, respectively.

```
#Make a table to summarize IRF coefficients and their confidence intervals

irf.table.ci <- round(data.frame(period = seq(1, 8),
  response.Adwords = irfs$irf$DLAd,
  Adwords.lower = irfs$Lower$DLAd,
  Adwords.upper = irfs$Upper$DLAd,
  response.flyer = irfs$irf$LFLYER,
  flyer.lower = irfs$Lower$LFLYER,
  flyer.upper = irfs$Upper$LFLYER),4)

colnames(irf.table.ci) <- c('Period', 'DLAdwords', 'DLAdwords Lower', 'DLAdwords Upper', 'LFLYER',
  'LFLYER Lower', 'LFLYER Upper')

knitr::kable(irf.table.ci)
```

Period	DLAdwords	DLAdwords Lower	DLAdwords Upper	LFLYER	LFLYER Lower	LFLYER Upper
1	-0.0058	-0.0601	0.0495	0.0833	0.0259	0.1392
2	0.0369	-0.0223	0.1127	0.0377	-0.0166	0.0855
3	0.0046	-0.0106	0.0305	0.0061	-0.0083	0.0216
4	0.0019	-0.0039	0.0093	0.0023	-0.0037	0.0099
5	0.0003	-0.0023	0.0027	0.0004	-0.0014	0.0028
6	0.0001	-0.0004	0.0015	0.0001	-0.0003	0.0013
7	0.0000	-0.0002	0.0003	0.0000	-0.0002	0.0004
8	0.0000	0.0000	0.0003	0.0000	0.0000	0.0002

Now we apply the $t > 1$ rule to determine coefficient significance and calculate long-term elasticities of AdWords and flyer advertising spending.

```
#Adwords
result_irf_adwords<-matrix(nrow = 8, ncol = 1)

for (i in 1:8) {
  se <- (irfs$Upper$DLAd[i]-irfs$Lower$DLAd[i])/(2*1.96)
  t_irf_adwords<- irfs$irf$DLAd[i]/se

  if (t_irf_adwords>1) {
    result_irf_adwords[i] <- irfs$irf$DLAd[i]
  } else {
    result_irf_adwords[i] <-0
  }
}

result_irf_adwords #print out the results
```

```
##           [,1]
## [1,] 0.0000000
## [2,] 0.03691742
## [3,] 0.00000000
## [4,] 0.00000000
## [5,] 0.00000000
## [6,] 0.00000000
## [7,] 0.00000000
## [8,] 0.00000000
```

```
lr_adwords <- sum(result_irf_adwords)
lr_adwords
```

```
## [1] 0.03691742
```

```
#Flyer spending
result_irf_flyers<-matrix(nrow = 8, ncol = 1)

for (i in 1:8) {
  se <- (irfs$Upper$LFLYER[i]-irfs$Lower$LFLYER[i])/(2*1.96)
  t_irf_flyers<- irfs$irf$LFLYER[i]/se

  if (t_irf_flyers>1) {
    result_irf_flyers[i] <- irfs$irf$LFLYER[i]
  } else {
    result_irf_flyers[i] <-0
  }
}

result_irf_flyers #print out the results
```

```
##           [,1]
## [1,] 0.08333906
## [2,] 0.03768691
## [3,] 0.00000000
## [4,] 0.00000000
## [5,] 0.00000000
## [6,] 0.00000000
## [7,] 0.00000000
## [8,] 0.00000000
```

```
lr_flyers <- sum(result_irf_flyers)
lr_flyers
```

```
## [1] 0.121026
```

After applying the $t > 1$ rule, we figure out that the AdWords advertising has a significant and positive impact on revenues in second period, while flyer advertising has significant and positive impact on revenues in the first and second period. Put it more specifically, after adding up significant coefficients overtime to get the long-term elasticities for both advertisings, we can say that:

- An 1% increase in AdWords advertising spending growth (note that we first-differenced the series) will increase the firm’s revenues by 0.04% in the long run.
- An 1% increase in flyer advertising spending will increase the firm’s revenues by 0.12% in the long run.

Optimal Allocation Between AdWords and Flyer Spending

Finally, we can respond to the final question from the CMO regarding ABC’s budget allocation. To do this, we may first take a look at the current budget allocation of ABC. We just need to review the dataset and calculate the total amount of money that the firm has spent on AdWords and flyers, respectively. Then we create a pie chart to visualize the current budget allocation of the firm.

```
#Current budget allocation

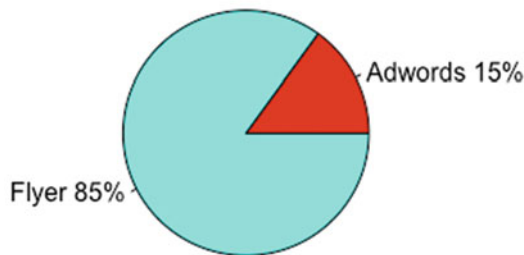
cost_adwords<-sum(data$Adwords)
cost_flyer<-sum(data$Flyer)
cost_total <- cost_adwords + cost_flyer

costshare_adwords<-cost_adwords/cost_total
costshare_flyer<-cost_flyer/cost_total
```

```
# Ingredients for the pie-chart
slices_actual<-c(costshare_adwords, costshare_flyer )
lbls_actual<-c("Adwords", "Flyer")
pct_actual<-round(slices_actual*100)
lbls_actual<-paste(lbls_actual, pct_actual) # add data to labels
lbls_actual<-paste(lbls_actual, "%", sep="") # add % sign to labels

# Get the pie-chart
pie(slices_actual, labels=lbls_actual, col=rainbow(length(lbls_actual)), main="Actual Budget Allocation" )
```

Actual Budget Allocation



We can see that the firm is currently putting far more resources on flyers, since it spends 85% of its budget on it and only 15% on online AdWords.

For the optimal marketing budget allocation, we need to retrieve the impact of AdWords and flyer spending from IRF analysis. More specifically, we will calculate the optimal allocation for each marketing channel using the following formula:

$$\text{Optimal Allocation}_i = \frac{\eta_i}{\sum_{i=1}^I \eta_i}$$

where η is the elasticity of marketing tool i .

As an example, for AdWords spending, we will calculate it as follows:

$$\text{Optimal Allocation}_{\text{Adwords}} = \frac{\eta_{\text{AdWords}}}{\eta_{\text{AdWords}} + \eta_{\text{Flyers}}}$$

Let's do this in R now:

```
#Get the coefficients from IRF results
beta_adwords<-lr_adwords
beta_flyer<-lr_flyers

#The sum of all elasticities
beta_all<-beta_adwords+beta_flyer

#Optimal resource allocation
optim_adwords<-beta_adwords/beta_all
optim_flyer<-beta_flyer/beta_all
```

Having figured out the optimal budget allocation between AdWords and flyer, we can now create another pie chart so that we can compare:

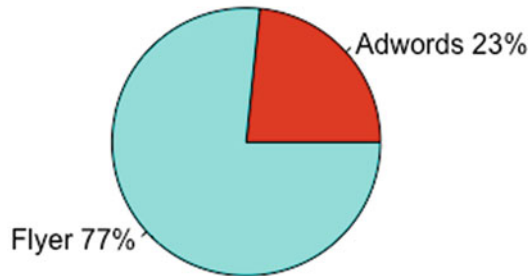
```
## Pie-chart ingredients
optimal_spend<-c(optim_adwords,optim_flyer)
optimal_spend=round(optimal_spend, digits=5)
optimal_spend
```

```
## [1] 0.23374 0.76626
```

```
slices_optim<-c(optim_adwords, optim_flyer)
lbls_optim<-c("Adwords", "Flyer")
pct_optim<-round(slices_optim*100)
lbls_optim<-paste(lbls_optim, pct_optim) # paste variable names to data labels
lbls_optim<-paste(lbls_optim, "%", sep="") # add % sign to labels

# Get the pie-chart
pie(slices_optim, labels=lbls_optim, col=rainbow(length(lbls_optim)), main="Optimal Budget Allocation" )
```

Optimal Budget Allocation



The optimal budget allocation is that the firm should actually spend less of its marketing budget on flyer advertising (77%, instead of 85%), and more on Google AdWords advertising (23% instead of 15%). Contrasting the optimal and actual budget allocation of the firm, it is quite obvious that currently, the firm is underestimating the power of online marketing through AdWords and over-emphasizing the importance of offline flyers.

We can see that without analyzing resource allocation, a firm can be quite far away from what it “should” do. Looking at the optimal budget allocation is quite critical in managers’ decision-making, since utilizing the constrained resource more wisely can potentially make a big difference to firm performance (e.g., revenues).

On a final note, this section talks about the allocation when the sales performance is taken into consideration. Brand managers may pursue different KPIs as well, such as market share, profits, and brand liking. With different KPIs pursued by the brand manager, the allocation would be different. Moreover, instead of keeping the budget the same and reallocating it, the brand manager may want to increase the budget. In such a case, the dynamics between marketing input and financial performance would be altered, leading to different optimal allocation.

To conclude, we responded to the questions raised by ABC’s CMO regarding demand forecasting, marketing effectiveness, and budget allocation ARIMA (and MLR) and VAR (and FEVD and IRF) methods. We hope that our readers can have a better understanding of the materials covered in this chapter by referring to this application exercise.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *2nd international symposium on information theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971* (pp. 267–281). Akadémiai Kiadó: Budapest.
- Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, 96(1), 116–131.
- Bronnenberg, B. J., Mahajan, V., & Vanhonacker, W. R. (2000). The emergence of market structure in new repeat-purchase categories: The interplay of market share and retailer distribution. *Journal of Marketing Research*, 37(1), 16–31.
- Colicev, A., Malshe, A., Pauwels, K., & O'Connor, P. (2018). Improving consumer mindset metrics and shareholder value through social media: The different roles of owned and earned media. *Journal of Marketing*, 82(1), 37–56.
- Colicev, A., Kumar, A., & O'Connor, P. (2019). Modeling the relationship between firm and user generated content and the stages of the marketing funnel. *International Journal of Research in Marketing*, 36(1), 100–116.
- De Haan, E., Wiesel, T., & Pauwels, K. (2016). The effectiveness of different forms of online advertising for purchase conversion in a multiple-channel attribution framework. *International Journal of Research in Marketing*, 33(3), 491–507.
- Dekimpe, M. G., & Hanssens, D. M. (1995). The persistence of marketing effects on sales. *Marketing Science*, 14(1), 1–21.
- Deleersnyder, B., Geyskens, I., Gielens, K., & Dekimpe, M. G. (2002). How cannibalistic is the Internet channel? A study of the newspaper industry in the United Kingdom and the Netherlands. *International Journal of Research in Marketing*, 19(4), 337–348.
- Engle, R. F., Lilien, D. M., & Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica: journal of the Econometric Society*, 55(2), 391–407.
- Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, 55(2), 251–276.
- Esteban-Bravo, M., Vidal-Sanz, J. M., & Yildirim, G. (2017). Can retail sales volatility be curbed through marketing actions? *Marketing Science*, 36(2), 232–253.
- Fischer, M., Shin, H. S., & Hanssens, D. M. (2016). Brand performance volatility from marketing spending. *Management Science*, 62(1), 197–215.
- Franses, P. H., & Van Dijk, D. (1996). Forecasting stock market volatility using (non-linear) Garch models. *Journal of Forecasting*, 15(3), 229–235.
- Franses, P. H., & Van Dijk, D. (2000). *Non-linear time series models in empirical finance*. Cambridge: Cambridge University Press.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 37(3), 424–438.
- Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B: Methodological*, 41, 190–195.
- Hanssens, D. M., & Pauwels, K. H. (2016). Demonstrating the value of marketing. *Journal of Marketing*, 80(6), 173–190.
- Ilhan, B. E., Kübler, R. V., & Pauwels, K. H. (2018). Battle of the brand fans: Impact of brand attack and defense on social media. *Journal of Interactive Marketing*, 43, 33–51.
- Kireyev, P., Pauwels, K., & Gupta, S. (2016). Do display ads influence search? Attribution and dynamics in online advertising. *International Journal of Research in Marketing*, 33(3), 475–490.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1–3), 159–178.
- Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin/New York: Springer Science & Business Media.

- Nijs, V. R., Srinivasan, S., & Pauwels, K. (2007). Retail-price drivers and retailer profits. *Marketing Science*, 26(4), 473–487.
- Pauwels, K. (2004). How dynamic consumer response, competitor response, company support, and company inertia shape long-term marketing effectiveness. *Marketing Science*, 23(4), 596–610.
- Pauwels, K. H. (2017). Modern (multiple) time series models: The dynamic system. In *Advanced methods for modeling markets* (pp. 115–148). Cham: Springer.
- Pauwels, K., Silva-Risso, J., Srinivasan, S., & Hanssens, D. M. (2004). New products, sales promotions, and firm value: The case of the automobile industry. *Journal of Marketing*, 68(4), 142–156.
- Pauwels, K., Demirci, C., Yildirim, G., & Srinivasan, S. (2016). The impact of brand familiarity on online and offline media synergy. *International Journal of Research in Marketing*, 33(4), 739–753.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Slotegraaf, R. J., & Pauwels, K. (2008). The impact of brand equity and innovation on the long-term effectiveness of promotions. *Journal of Marketing Research*, 45(3), 293–306.
- Srinivasan, S., Vanhuele, M., & Pauwels, K. (2010). Mind-set metrics in market response models: An integrative approach. *Journal of Marketing Research*, 47(4), 672–684.
- Srinivasan, S., Rutz, O. J., & Pauwels, K. (2016). Paths to and off purchase: Quantifying the impact of traditional marketing and online consumer activity. *Journal of the Academy of Marketing Science*, 44(4), 440–453.
- Van Dieijen, M., Borah, A., Tellis, G. J., & Franses, P. H. (2019). Big data analysis of volatility spillovers of brands across social media and stock markets. *Industrial Marketing Management*, 88, 465.
- Yildirim, G., Wang, W., & Deleersnyder, B. (2020) Market turbulence following a major new product introduction: Is it really so bad? Working paper.



Modeling Marketing Dynamics Using Vector Autoregressive (VAR) Models

Shuba Srinivasan

Contents

Introduction	516
Vector Autoregressive (VAR) Modeling	518
Unit Root and Cointegration Testing	519
Testing for Evolution Versus Stationarity	519
Cointegration Tests: Does a Long-Run Equilibrium Exist Between Evolving Series?	521
Models of Dynamic Systems of Equations	523
Vector Autoregressive Model with Exogenous Variables (VARX)	523
Structural Vector-Autoregressive Model (SVAR)	529
Vector Error Correction Model	530
Policy Simulation Analysis	530
Impulse Response Functions (IRF)	530
Dynamic Multipliers	535
Granger Causality Tests: Do We Need to Model a Dynamic System?	535
Forecast Error Variance Decomposition (FEVD)	537
Software Programs for VAR Estimation	539
Illustrative Applications of VAR Modeling in Marketing	539
Investor Response Models in the Marketing-Finance	539
Marketing and Mindset Metrics Models	541
Digital Marketing Models	542
Conclusion	543
Cross-References	543
References	543

Abstract

Time-series data include repeated measures of marketing activities and performance that are typically equally spaced in time. In the context of such data, Vector Autoregressive (VAR) models are uniquely suited to capture the time dependence

S. Srinivasan (✉)

Boston University Questrom School of Business, Boston, MA, USA

e-mail: ssrini@bu.edu

of both a criterion variable (e.g., sales performance) and predictor variables (e.g., marketing actions, online consumer behavior metrics), as well as how they relate to each other over time. The objective of this chapter is to provide a foundation in VAR models and to enable the readers to apply them in their own research domain of interest. To this end, the chapter will discuss both the underlying perspectives and differences among alternative VAR models, and the practical issues with testing, model choice, estimation, and interpretation that are common in empirical research in marketing.

From a marketing strategy perspective, both managers and academic researchers pay attention to whether a performance change is temporary (short-term) or lasting (long-term). Establishing the distinction between short-term and long-term marketing effectiveness is central to the understanding of marketing strategy and its implications, which this chapter aims to do. The interaction among appropriate marketing phenomena, modeling philosophy, and contemporary substantive topics sets this work apart from previous treatments on the broader topic of econometrics and time-series analysis in marketing (e.g., Dekimpe and Hanssens, Persistence modeling for assessing marketing strategy performance. In: Lehmann D, Moorman C (eds) *Cool tools in marketing strategy research*. Marketing Science Institute, Cambridge, MA, 2004; Hanssens et al., *Market response models: Econometric and time series analysis*. Springer Science and Business Media, 2001; Pauwels, *Found Trends Market* 11(4):215–301, 2018).

Keywords

Vector autoregressive models · Vector-error correction models · Impulse response functions · Forecast error variance decomposition · Long-term marketing effectiveness

Introduction

In today's data-intensive world, data on marketing actions, market, and firm performance can be gathered in a variety of forms. Managers can examine marketing data over time, for example, through measures such as digital advertising spend per week, brand revenues per month, advertising expenditures over the quarters, firm revenues over the past year, and marketing mindset metrics over several years. Time-series data include repeated measures of marketing activities and performance that are typically equally spaced in time. In the context of such data, Vector Autoregressive models (VAR) are uniquely suited to capture the time dependence of both a criterion variable (e.g., sales performance) and predictor variables (e.g., marketing actions, online consumer behavior metrics), as well as how they relate to each other over time.

The objective of this chapter is to provide a foundation in VAR models and to enable the readers to apply them in their own research domain of interest. To this

end, I will discuss both the underlying perspectives and differences among alternative VAR models, and the practical issues with testing, model choice, estimation, and interpretation that are common in empirical research in marketing. The interaction among appropriate marketing phenomena, modeling philosophy, and contemporary substantive topics sets this work apart from previous treatments on the broader topic of econometrics and time-series analysis in marketing (e.g., Dekimpe and Hanssens 2004; Hanssens et al. 2001; Pauwels 2018).

From a marketing strategy perspective, both managers and academic researchers pay attention to whether a performance change is temporary (short-term) or lasting (long-term). If revenues have declined dramatically this quarter, will they soon rebound to their historic mean and upward trend? Or will they not, in which case firm action may be called for? Likewise, some marketing actions are often considered tactical tools, such as price promotions to boost sales, but may hurt brand performance in the long run (Mela et al. 1997; Pauwels et al. 2002; Srinivasan et al. 2000). Other marketing actions, such as investments in product introductions and advertising may only be justified by promises of future benefits (Srinivasan et al. 2010). Establishing the distinction between short-term and long-term marketing effectiveness is central to the understanding of marketing strategy and its implications.

There has been a rapid growth of applications of VAR models in the marketing literature for a variety of reasons. First, VAR models enable researchers to use a systems approach to explain the multiple channels of influence of marketing variables on each other, and enable the incorporation of customer response, competitive response, and firm actions (e.g., Pauwels et al. 2004; Srinivasan et al. 2004). As such, VAR models are effective in incorporating the combined influence of multiple stakeholders, accounting for the real world of firm and marketing strategy. Second, they allow researchers to make a distinction between short-term, long-term, and cumulative effects of marketing considering differences between temporary, evolving, and structural changes in marketing variables (e.g., Srinivasan et al. 2000). Finally, the availability of online and offline databases (e.g., Srinivasan et al. 2016) with both longitudinal and cross-sectional data has meant that VAR models have great applicability and have enabled various empirical generalizations on marketing phenomena.

The outline of the chapter is as follows. First, I will provide an overview into the VAR modeling process, followed by a discussion on evolution vs stationarity and unit root testing. I then discuss the concept of cointegration and long-run equilibrium among evolving series. Next, I discuss the details of VAR model specification, including reduced form and structural VARs, followed by Vector-Error Correction Models (VECM). Following this, I will review the importance of Impulse Response Analysis, Granger Causality Tests, and Forecast Error Variance Decompositions (FEVD) in order to generate substantive and policy implications from VAR models. Finally, I conclude the chapter with three (illustrative) contemporary applications of VAR to marketing strategy.

Vector Autoregressive (VAR) Modeling

Vector autoregressive modeling involves a multistep process. Table 1 provides an illustration of how a researcher/practitioner should approach a VAR model.

Table 1 Illustrative approach to VAR modeling framework

Managerial/research goal	Key sources	Methodological step
Step 1: Unit Root & Cointegration Tests		
Are variables stationary or evolving? Are evolving variables in long-run equilibrium?	Enders (2003) Engle and Granger (1987) Perron (1989) Perron (1990) Zivot and Andrews (1992) Johansen et al. (2000) Srinivasan et al. (2000)	Dickey Fuller Tests Augmented Dickey-Fuller Test Cointegration Test Structural Break Test
Step 2: Model of Dynamic System		
How do performance and marketing interact in the long run and short run, accounting for the unit root and cointegration results?	Lütkepohl (1993) Dekimpe and Hanssens (1999) Baghestani (1991) Srinivasan et al. (2004)	Vector Autoregressive model VAR in Differences Vector Error Correction model
Step 3: Policy Simulation Analysis		
What is the dynamic impact of marketing on performance? Which actors drive the dynamic impact of marketing?	Pesaran and Shin (1998) Pauwels et al. (2002) Han et al. (2019)	Unrestricted impulse response Restricted policy simulation Dynamic multipliers
Step 4: Granger Causality Tests		
Which variables are temporally causing which other variables?	Granger (1969) Srinivasan et al. (2010)	Granger Causality
Step 5: Drivers of Performance		
What is the importance of each driver's past in explaining performance variance? Independent of causal ordering?	Hanssens (1998) Srinivasan et al. (2004) Nijs et al. (2007)	Forecast Variance Error Decomposition (FEVD) Generalized FEVD

In the first step, unit root tests are used to determine whether the different variables are stable or evolving. If several of the variables are found to have a unit root – that is, if they are found to be evolving – one subsequently tests for cointegration. Depending on the outcome of these two preliminary steps, one estimates a vector autoregressive (VAR) model, in reduced form or structural form in the levels, in the differences, or a vector error correction model (VECM). The parameter estimates from this VAR (or VECM) model are used to derive impulse-response functions and forecast error variance decompositions, from which various summary statistics on the short- and long-run dynamics of the system can be derived. Granger Causality tests help with assessing the temporal causality patterns among the variables. We now briefly elaborate on each of these steps.

Unit Root and Cointegration Testing

Testing for Evolution Versus Stationarity

In the first step, unit root tests are used to determine whether or not the different variables are stable or evolving. When a series may be appropriately modeled as depending on a constant plus a coefficient times a lag of the series plus a random term, testing whether a series is stationary or evolving is accomplished using the well-known test proposed by Dickey and Fuller (1979). When more than one lag is involved the appropriate test is the augmented Dickey Fuller test. If several of the variables are found to have a unit root – that is, if they are found to be evolving – one subsequently tests for cointegration among the evolving series.

Stationarity is the tendency of a time series to revert back to its deterministic components, such as a fixed mean (mean-stationary) or a mean and trend (trend-stationary). Stationary processes have a finite variance and are predictable, while evolving series do not return to a fixed mean (and trend); shocks to these series persist in the future. This distinction is essential in empirically testing for unit roots. Following Dekimpe and Hanssens (1995a), we first consider the simple case where the time-series behavior of the variable of interest (for example, a brand's sales Y_t) is described by a first-order autoregressive process:

$$(I - \varphi L)Y_t = a + u_t \quad (1)$$

where φ is an autoregressive parameter, L the lag operator (i.e., $L^p Y_t = Y_{t-p}$), u_t is a residual series of zero mean, constant variance (σ_u^2), and uncorrelated random shocks, and a is a constant. Note that Eq. (1) may also be written in the following, more familiar form after applying successive backward substitutions:

$$Y_t = [a/(1 - \varphi)] + u_t + \varphi u_{t-1} + \varphi^2 u_{t-2} + \dots, \quad (2)$$

in which the present value of Y_t is explained as a weighted sum of random shocks. Depending on the value of φ , two scenarios are distinguished.

If $|\varphi| < 1$, the effect of past sales (and thus any “shock” that has affected past sales) diminishes as we move into the future. The impact of past shocks diminishes and eventually becomes negligible. Hence, each shock has only a temporary impact. In such a case, the series has a fixed mean $c/(1 - \varphi)$ and a finite variance $\sigma_u^2/(1 - \varphi^2)$. We call such time series stationary, i.e., it has a time-independent mean and variance. This situation is typical for the market performance of established brands in mature markets (e.g., Nijs et al. 2001; Srinivasan et al. 2000).

If $|\varphi| = 1$, sales will not revert to a historical level but will evolve. This situation has been demonstrated for smaller brands and in emerging markets (e.g., Slotegraaf and Pauwels 2008). If $|\varphi| > 1$, the effect of past sales (and thus of past shocks) becomes increasingly important. Such explosive time-series behavior appears to be unrealistic in marketing (Dekimpe and Hanssens 1995b). When $|\varphi| = 1$, Eq. (2) becomes:

$$Y_t = (a + a + \dots) + u_t + u_{t-1} + \dots, \quad (3)$$

In this case, each random shock has a permanent effect on the subsequent values of Y . Sales do not revert to a historical level, but instead wander freely in one direction or another, and are evolving.

Distinguishing between stationarity versus evolution therefore involves checking if φ in Eq. (1) is smaller than or equal to 1. However, for the t -statistic special tables need to be used in lieu of the standard distribution. The generalization of the Dickey-Fuller test to an AR (p) process yields the Augmented Dickey-Fuller test. This test is based on a reformulation of the AR (p) process as:

$$(1 - L)Y_t = \Delta Y_t = \alpha_0 + \beta Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \dots + \alpha_{2p} \Delta Y_{t-p} + u_t \quad (4)$$

The Augmented Dickey-Fuller (ADF) test can be used to test the null hypothesis. The rationale behind adding lagged first differences is that the error should be approximately white noise. In addition, depending on the assumptions of the underlying process, the test may be performed with or without the model intercept. Enders (2003) offered an iterative procedure to implement these different test specifications, as implemented in several marketing papers (e.g., Slotegraaf and Pauwels 2008; Srinivasan et al. 2004).

While the Augmented Dickey Fuller (ADF) method is the most popular unit root test in marketing, it has the limitation of the low power of the test under certain conditions; see Maddala and Kim (2007) for an excellent discussion. Because it has been argued that conventional unit root tests (e.g., ADF) tend to underreject the null of unit root, researchers tend to use the alternative such as the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (Kwiatkowski et al. 1992), which uses the null of stationarity. Most applications in marketing, however, show that the ADF and KPSS tests lead to similar conclusions (e.g., Pauwels and Weiss 2008; Villanueva et al. 2008). Other papers have built on these topics; for example, Franses et al. (1999) developed an outlier-robust unit root test and the logical consistency requirement when modeling market shares has also been incorporated in unit root tests by Franses et al. (2001).

Typically, unit root tests are based on the assumption of no structural breaks in the series. If a structural break is identified in a series, the stability of the model is crucial to the task of evaluating the impact of structural changes within the system. It could be that there are large shocks called “structural breaks.” Failure to account for such shocks in testing biases the unit root tests toward reporting evolution. In marketing, such structural breaks could correspond to permanent changes to prices, the introduction of new products and channels of distribution (Deleersnyder et al. 2002; Kornelis et al. 2008; Pauwels and Srinivasan 2004), and other similar permanent changes in marketing. Researchers typically define a structural break in terms of a parameter change in the deterministic part of the model, i.e., in the slope and/or intercept of the deterministic growth path (Perron 1989).

The timing of the structural break is either known or unknown. In the case of known structural breaks there are shifts in the data generating process, such as price increases or decreases (Srinivasan et al. 2000), introducing a new store brand (Pauwels and Srinivasan 2004), or channel (Kornelis et al. 2008), or changing the pricing structure from free to fee (Pauwels and Weiss 2008). Perron (1990) developed the most widely used test for a single break, which has been the focus of most marketing applications (Deleersnyder et al. 2002; Lim et al. 2005; Nijs et al. 2001; Pauwels and Srinivasan 2004). Zivot and Andrews (2002) propose testing for structural breaks that are unknown which may be a common occurrence, instead of eyeballing the time series for where a structural break should be and then testing for it. For instance, if players anticipate changes, they may even react before the time the researcher dates the event (e.g., Pauwels and Srinivasan 2004), which will be picked up by unknown structural break tests.

Cointegration Tests: Does a Long-Run Equilibrium Exist Between Evolving Series?

Cointegration describes the existence of an equilibrium or stationary relationship among two or more time series, each of which is individually nonstationary. An equilibrium relationship would imply that, even if they diverge from each other in the short run, such deviations are stochastically bounded or diminishing over time. Figure 1 shows an illustrative example of cointegration between annual advertising spending and sales revenues for the Lydia Pinkham from 1907 to 1960.

To illustrate cointegration, we consider an example in which a brand’s sales (*SALES*), its own marketing (*MKTG*), and its competitors’ marketing (*CMKTG*) are all evolving. The existence of a cointegrating relationship between these three variables would imply (see Srinivasan et al. (2000) for a more in-depth discussion) the following:

$$SALES_t = \beta_0 + \beta_1 MKTG_t + \beta_2 CMKTG_t + e_t \quad (5)$$

A simple testing procedure for cointegration, proposed by Engle and Granger (1987), is to estimate Eq. (5) using Ordinary Least Squares and test the residuals e_t

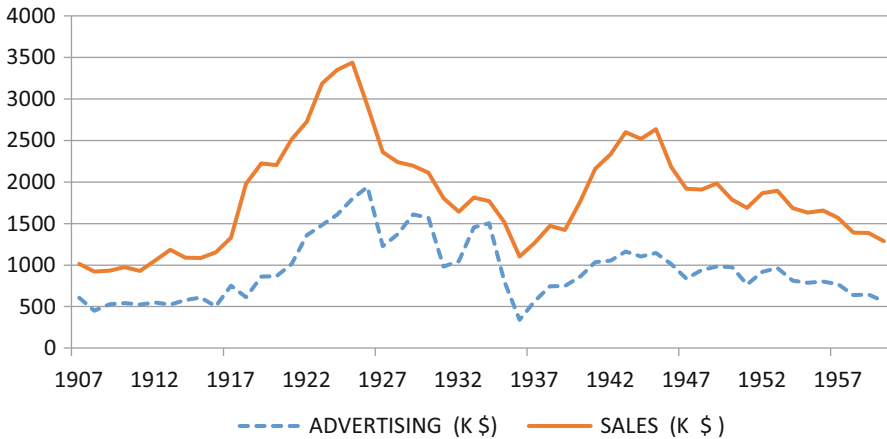


Fig. 1 Cointegrating relationship between advertising and sales – Lydia Pinkham Data

for a unit root using standard unit root tests, and with the updated critical values listed in Engle and Yoo (1987). A marketing application of the Engle and Granger approach to cointegration testing can be found in Baghestani (1991) using the Lydia Pinkham data. This procedure is simple and consistent, but can be biased in small samples and is not possible with more than one cointegrating vector. Johansen's (1988) Full Information Maximum Likelihood (FIML) test does not suffer from these limitations, and has been used extensively in marketing applications. Dekimpe and Hanssens (1999) applied the latter test in their analysis of a prescription drugs market. They found that even though each of the individual series (prescriptions, advertising, sales calls, and price differential) was evolving, the four variables were related together in a long-run cointegrating equilibrium that prevented them from wandering too far apart from one another.

Such long-run cointegrating equilibria can emerge for a variety of reasons. First, cointegration can arise from stationary linear combinations of category and brand sales, for instance. If such regressions exist, then they are consistent with market shares being stationary or stable. Srinivasan and Bass (2000) propose that if brand sales and category sales are cointegrated, this also implies that the market is in long-run equilibrium. From a strategic perspective, this implies that firms are unable to improve their relative position despite improving their absolute long-run performance with respect to sales. Srinivasan and Bass (2000) show that evolution occurs for a majority of brand sales series, and that a vast majority of the market share time-series models are stationary. This is consistent with arguments of Bass and Pilon (1980) that many markets are in a long-run equilibrium. The relative position of the brands is only temporarily affected by marketing activities. Even for brand sales (and category sales) series, later studies suggest that stationarity, and not evolution is the norm for mature brands in mature categories (e.g., Nijs et al. 2001; Pauwels et al. 2002). Emerging markets and brands show a substantially higher potential for evolution (Osinga et al. 2010; Slotegraaf and Pauwels 2008). More recently, Kireyev

et al. (2016) use the Johansen et al. (2000) test allowing for structural breaks in the relationship between online display and online search clicks.

Second, certain budgeting rules (e.g., advertising as a percentage-of-sales allocation rules) imply that sales successes eventually translate into higher marketing spending, which may result in sales and marketing mix variables being cointegrated. For instance, Srinivasan et al. (2000) note that one rationale for a long-run cointegrating relationship between own prices, competitive prices, and sales is that different price levels correspond to different long-run demand or sales levels, and these, in turn, are associated with different levels of shares. This long-run cointegrating equilibrium among sales and prices is consistent with the underlying idea that customers' limited budgets may cause different price levels to be associated with different long-run demand levels.

Finally, competitive decision rules can result in firms' marketing spending levels never deviating too far from each other (Dekimpe and Hanssens 2004). Table 2 summarizes the findings from an illustrative set of papers that link sales to marketing mix.

If there is a cointegrating relationship, we estimate a Vector-Error Correction Model (VECM). If the series are stationary, we estimate a VAR model in levels. However, if there are unit-roots but there is no cointegrating relationship, we estimate a VAR in differences since regressions on the levels of evolving variables may produce spurious results.

Models of Dynamic Systems of Equations

Vector Autoregressive Model with Exogenous Variables (VARX)

The dynamic interactions and feedback effects among marketing variables are captured in Vector-Autoregressive (VARX) models with exogenous variables (e.g., Dekimpe and Hanssens 1999). The endogenous treatment of marketing actions implies that they are explained by both past marketing actions and past performance variables. VARX models can capture complex feedback loops that may impact brand performance over time. For instance, an increase in advertising in each week may generate a high level of consumer awareness, inducing some consumers to consider the brand, and try it. Their subsequent purchases may not only increase brand sales, but also awareness by their family, friends, and colleagues who see them use the brand and follow suit themselves. Because of such chains of events, the full performance implications of the marketing actions, such as advertising, may extend well beyond the immediate effects. By capturing these feedback loops, VARX models yield a comprehensive picture of how marketing mix actions affect the full dynamic system including sales performance.

Motivations for using and estimating VARX models stem from an interest in explaining the dynamics and interrelationships among multiple variables. For instance, sales may be driven by feedback from past performance (*SALES* at period $t-1$) as well as by own marketing actions and competitive marketing actions.

Table 2 Illustrative marketing studies using vector auto-regressive models

Study	Data	Model	Research focus
Baghestani (1991)	54 yearly observations	VECM	Cointegration of advertising and sales
Bronnenberg et al. (2000)	1991–1996; 257 weeks of ready-to-drink tea.	VAR	Relation between distribution and long-run share, through early growth and later stages of product life cycle
Dekimpe and Hanssens (1995a)	Database of 400 prior analyses	Unit root tests	Evolution of sales and stationarity of market shares
Dekimpe and Hanssens (1995b)	Monthly advertising and sales data	VAR Persistence analysis	Long-term effect of advertising on sales
Dekimpe and Hanssens (1999)	(1) Monthly pharmaceutical data for 5 years (2) Brandaid data	VECM	Long-run impact of sales calls, advertising, prices and promotions on sales
Dekimpe et al. (1999)	113 Weeks scanner data for four categories	VAR Impulse response analysis	Impact of promotions on sales in stationary and non-stationary markets
Franses et al. (1999)	Weekly scanner data for consumer-packaged goods	Cointegration analysis	Impact of distribution, advertising and promotions on sales
Franses et al. (2001)	Weekly scanner data for 2 years in three categories	Unit root tests	Propose unit-root and cointegration tests that take the logical-consistency properties of market-share series into account
Srinivasan and Bass (2000)	Weekly scanner data for grocery products	Cointegration analysis/VECM	Evolution of sales with stationary market shares: brand sales and category sales are cointegrated.
Nijs et al. (2001)	4 years for 560 categories	VAR Persistence analysis	Persistent effect of promotions on category sales
Pauwels et al. (2002)	Scanner data for 2 years in two categories	VAR Persistence analysis	Quantifies the long-term effect on category incidence, brand choice and purchase quantity
Srinivasan et al. (2004)	75 brands in 25 categories for 7 years	VAR Persistence analysis	Quantifies effects of price promotions on manufacturer revenues, retailer revenues and margin
Pauwels and Srinivasan (2004)	75 brands in 25 categories for 7 years	VAR with structural break	Effect of store brand entry on (1) the retailer, (2) the manufacturers, and (3) the consumers

(continued)

Table 2 (continued)

Study	Data	Model	Research focus
Pauwels et al. (2004)	Weekly automotive data from 1996 to 2001	VAR analysis, FEVD and persistence analysis	Short- and long-term impact of promotions and new product introduction on revenues, profits and stock market performance
Steenkamp et al. (2005)	Weekly data for 4 years in 442 consumer product categories	VAR analysis and persistence analysis	Competitive reaction elasticities due to price promotion or advertising attacks, both in the short and the long run
Nijs et al. (2007)	Weekly data for 24 categories for 8 years	VAR and FEVD	Drivers of retail prices: competitive retailer prices, pricing history, brand demand, wholesale prices, and retailer category-management considerations
Villanueva et al. (2008)	Internet firm data for a 70-week period	VAR and Persistence analysis	Impact of marketing vs. word of mouth customer acquisition on customer equity
Srinivasan et al. (2008)	Weekly data for 24 categories for 8 years	VAR	Retailers choice of demand-based pricing vs. inertia
Trusov et al. (2009)	36 weeks of sign-ups, referrals, media and marketing events for internet firm	VAR	Effect of word-of-mouth (WOM) vs. marketing on member growth
Heerde et al. (2010)	Data for Lexus RX300 introduction	Time-varying VEC model	Estimate cannibalization, brand switching, and primary demand expansion for a pioneering innovation
Srinivasan et al. (2010)	Weekly data for 60 consumer brands in four categories for 96 weeks	VAR with IRFs, GFEVD and Granger Causality	Analyze the added explanatory value of customer mindset metrics vs. marketing mix in a sales response model
Wiesel et al. (2011)	Daily data from office supply firm: transaction, marketing, online and off-line activities	VAR + experimentation	Marketing communication effects on offline and online purchase funnel metrics and the magnitude and timing of the profit impact of firm-initiated and customer-initiated contacts
Pauwels et al. (2016)	50 weekly observations of brand performance, online and offline media for four companies	Bayesian VAR with IRFs	Assess how within-online synergy and cross-channel synergy vary across familiar and unfamiliar brands
Srinivasan et al. (2016)	CPG data on marketing mix, online media, and sales for 40 weeks	VAR with IRFs, Granger Causality	Effects of consumer activity in online media (paid, owned

(continued)

Table 2 (continued)

Study	Data	Model	Research focus
			and earned) vs. marketing mix on sales.
Colicev et al. (2018)	Daily data for 45 brands in 21 sectors on mindset metrics and firm performance	VAR with IRFs, Granger Causality	Role of mindset metrics on social media – shareholder value link.
Han et al. (2019)	Weekly survey data for 4 years on customers’ attitudes for computer and automobile brands	VAR with IRFs, GFEVD	Impact of negative buzz on awareness and purchase intent

Competitive marketing actions may be explained by their historical patterns and their reaction to competitive performance (i.e., performance feedback) and/or to the focal firm’s actions. For instance, a higher click-through and thus spending on paid search may be induced by the firm’s offline marketing actions and by higher sales in previous periods, e.g., due to positive word-of-mouth by previous customers (Srinivasan et al. 2016; Wiesel et al. 2011). In turn, word-of-mouth referrals may be driven by the firm’s paid marketing actions (Trusov et al. 2009).

Next we outline the VARX (vector autoregressive model with exogenous variables) . The common practice in marketing is to allow the most relevant variables to be endogenous and to control for the effects of other variables by considering them exogenously (Dekimpe and Hanssens 1999; Horváth et al. 2005; Nijs et al. 2001; Srinivasan and Bass 2000; Srinivasan et al. 2000). This, i.e., the imposition of exogeneity, can imply a reduction of the number of parameters and also an improved precision of forecasting. For expository purposes, we first consider a model in the levels and focus on a simple three-equation model linking own sales performance (*SALES*), own marketing spending (*OMKT*), and competitive marketing spending (*CMKT*). The corresponding VAR model in matrix notation the model given is by:

$$\begin{aligned}
 \begin{bmatrix} SALES_t \\ OMKT_t \\ CMKT_t \end{bmatrix} &= \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix} + \sum_{i=1}^p \begin{bmatrix} \Phi_{11,i} & \Phi_{12,i} & \Phi_{13,i} \\ \Phi_{21,i} & \Phi_{22,i} & \Phi_{23,i} \\ \Phi_{31,i} & \Phi_{32,i} & \Phi_{33,i} \end{bmatrix} \times \begin{bmatrix} SALES_{t-i} \\ OMKT_{t-i} \\ CMKT_{t-i} \end{bmatrix} + \Psi \\
 &\times \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} + \begin{bmatrix} u_{Sales,t} \\ u_{OMKT,t} \\ u_{CMKT,t} \end{bmatrix} \tag{6}
 \end{aligned}$$

which can be written as

$$Y_t = A + \sum_{i=1}^p \Phi_i Y_{t-i} + \Psi X_t + \Sigma_t, \quad t = 1, 2, \dots, T, \tag{7}$$

where A is a 3×1 vector of intercepts, Y_t is an 3×1 vector of the endogenous variables (*SALES*, *OMKT*, *CMKT*), and X_t is a vector of exogenous control variables. The exogenous variables often include terms such as (1) a deterministic-trend t to capture the impact of omitted, gradually changing variables, and (2) dummy variables to account for seasonal fluctuations in sales, or any other endogenous variable, and Σ_t is the covariance matrix of the residuals. Dekimpe and Hanssens (1995a) provide a good summary of the multiple channels of influence that are captured by VAR models. These are enumerated below. First, VAR models allow the capture of *contemporaneous effects* of variables on each other. For example, suppose a brand manager intends to launch a costly promotional campaign with a view to lifting the brand's declining sales performance. The anticipated immediate effect of the promotion on sales is an important factor in the launch of the promotional campaign. Many marketing actions including advertising, promotions, price changes, distribution changes, etc., often have a considerable immediate impact on performance. In the reduced form of the VAR, these effects are reflected in the contemporaneous correlation terms in the variance covariance matrix.

Second, there are *carryover effects* of marketing activity on sales, reflected in the parameter ϕ_{12} in Eq. (6). Numerous studies have argued that the effect of advertising in one period may be carried over, at least partially, into future periods (see Hanssens et al. 2014, for example). Consumers are expected to remember past advertising messages and create "goodwill" toward the brand that only gradually deteriorates because of forgetting.

Third, *purchase reinforcement effects* suggest that the dynamic impact of marketing actions on sales can also work indirectly through purchase reinforcements: a given outlay may create a new customer who will not only make an initial purchase, but also repurchase in the future. Purchase reinforcement is reflected in the parameter ϕ_{11} in Eq. (6). Using a similar logic, Horsky and Simon (1983) assume that advertising gives innovators an incentive to try the product after which an imitation effect takes over, creating a larger customer base and higher future sales. Current advertising should receive credit for these subsequent sales (e.g., Bass and Clarke 1972; Hanssens et al. 2001). Villanueva et al. (2008) in their study of customer equity note that an increase in the number of customers acquired through word-of-mouth might influence future word-of-mouth acquisitions because these customers may generate more referrals than customers acquired through marketing.

Fourth, *feedback effects* suggest that future marketing actions (e.g., advertising) may be influenced by past sales as well as current sales. Feedback from sales to marketing action is reflected in the parameter ϕ_{21} in Eq. (6). Dekimpe and Hanssens (1995a) point out that this is highly likely when percentage-of-sales budgeting rules are applied. Such feedback effects manifest due to a chain reaction where an increase in advertising in period t results in an increase in sales in period t , which in turn results in increased advertising in period $t+1$ and so on. The profit implications of such advertising increases should consider both the revenue impact as well as the additional expenditures in advertising due to the initiating increase in advertising. Using persistence modeling approaches such as VARs, credit is given to the initial

advertising increase for the subsequent sales increases since without it, none of the subsequent effects would have occurred.

Fifth, the VAR model allows the researcher to capture *firm-specific decision rules*, which typically model the dependence of current marketing activity such as advertising and prices on previous advertising and prices. Such a decision rule is reflected in the parameter ϕ_{22} in Eq. (6). For instance, Nijs et al. (2007) find that a brand's past prices are the dominant driver of current retail prices in each category studied, and account anywhere from 50% to 60% of the variation in current prices, confirming the powerful tendency to rely on past prices in determining future prices (Srinivasan et al. 2008).

Last but not the least, VAR models allow the capture of *competitive reaction effects*. Competitive reaction effects are reflected in the parameter ϕ_{23} and the parameter ϕ_{32} in Eq. (6). For instance, competitive activities may change own marketing effectiveness drastically. Similarly, in the context of retail pricing, competitive retailer activity is expected to influence retailer prices and performance (Srinivasan et al. 2008). For instance, price promotions by competing retailers may reduce store traffic, inducing the retailer to respond (Srinivasan et al. 2004).

The representation in Eq. (7) is the reduced-form of the VAR model in which the errors are contemporaneously correlated, with the contemporaneous correlations reflected in the off-diagonal terms of the variance-covariance matrix. The residual correlation matrix can be used to establish the presence and the direction of the effects. Various procedures have been used in the marketing literature to deal with identification, such as by imposing an *a priori* imposition of a certain causal ordering on the variables. When researchers work on model identification, they assert a connection between the reduced form and the structure so that estimates of reduced form parameters translate into structural parameters. In the words of Sims (1986, p.2), "*Identification is the interpretation of historically observed variation in data in a way that allows the variation to be used to predict the consequences of an action not yet undertaken.*"

An important feature of the reduced-form model is that all variables on the right-hand side (RHS) of Eq. (7) are predetermined at time t , and the system can be estimated without imposing restrictions or a causal ordering. Moreover, because all RHS variables are the same in each equation, there is no efficiency gain in using Seemingly Unrelated Regression (SUR) estimation. In that case, the Ordinary least squares (OLS) estimates are consistent and asymptotically efficient even if the errors are correlated across equations (Srivastava and Giles 1987, Chap. 2) and asymptotically efficient in using OLS estimation, equation by equation. This feature is especially valuable in marketing applications with many endogenous variables. For example, a 6-equation VAR model requires estimation of $6 \times 6 = 36$ additional parameters for each lag added to the model. This does not bode well for estimations with the typical weekly scanner panel data of 104 observations for 2 years, due to over-parameterization concerns. In contrast, OLS estimation equation-by-equation implies that only six additional parameters have to be estimated.

Several authors (e.g., Pesaran et al. 1993; Dekimpe and Hanssens 1995a; Pauwels et al. 2004) have restricted all parameters with $|t| < 1$ to zero as a step toward parsimony. Others such as Nijs et al. (2007) and Srinivasan et al. (2008), also for parsimony of estimation, reduce the number of parameters by eliminating the insignificant ones and reestimating the model. In this case, when researchers reduce the number of parameters by eliminating the insignificant lagged parameters and reestimating the model, they need to use SUR for estimation as the RHS variables are no longer identical across equations. While such strategies accomplish the goal of parsimony and may alleviate the problem of estimating and interpreting so many parameters, they are unlikely to fully eliminate it. As a consequence, VAR modelers typically do not interpret the individual parameters themselves, but rather focus on the impulse-response functions (IRFs) derived from these parameters.

Structural Vector-Autoregressive Model (SVAR)

The structural VAR (SVAR) representation of the reduced form VAR in Eq. (7) is written as:

$$B_0 Y_t = A + B_1 Y_{t-1} + B_2 Y_{t-2} + \dots + B_p Y_{t-p} + \Sigma_t \quad (8)$$

The vector of endogenous variables Y is regressed on constant terms (which may include a deterministic time trend) and on its own past, with p being the number of lags, and B coefficient matrix of a given lag. Note that the contemporaneous effects are captured in the B_0 matrix; as a result, the structural errors ε are uncorrelated (orthogonal) across equations. This structural form of the VAR model is directly interesting for decision makers, as it generates predictions of results of various kinds of actions, by calculating the conditional distribution given the action (Sims 1986). It is also the appropriate form for imposing restrictions, typically on the B_0 matrix (e.g., Amisano and Giannini 1997). For instance, researchers provide theory-based reasons for why one group of variables does not cause another group of variables (e.g., Bermanke 1986).

Structural VARs have seen quite a few applications in marketing (DeHaan et al. 2016; Gijsenberg et al. 2015; Horváth et al. 2005). Horváth et al. (2005) show that the inclusion of competitive reaction and feedback effects is more important in the tuna category but not in the shampoo category where competitive interactions are limited due to differentiated brand positioning. Gijsenberg et al. (2015) develop a Double-Asymmetric Structural VAR (DASVAR) model that allows for asymmetric effects of increases versus decreases and for a different number of lags in each equation. In their analysis of the effect of service on customer satisfaction, they find that losses (service failures) not only have stronger (the first asymmetry), but also longer-lasting (the second asymmetry) effects on satisfaction than gains. Finally, DeHaan et al. (2016) find support for their proposed restriction of both immediate and dynamic feedback loops within the online funnel of a retailer; they find that increases in product page visits increase checkouts, but not vice versa.

Vector Error Correction Model

If some of the variables have a unit root, the VAR model is specified in differences. If there is a cointegrating relationship, we estimate a Vector-Error Correction Model (VECM). Srinivasan et al. (2000) estimate a VECM model, among market share and prices of four brands of beer, which have a long-run cointegrating relationship; the rationale is that different price levels correspond to different long-run demand or sales levels, and these, in turn, are associated with different levels of shares. Specifically, if $SALES_t$, $OMKT_t$, and $CMKT_t$ are cointegrated, then the VECM model in differences includes the error-correction term to capture the long-run cointegrating relationships as shown below:

$$\begin{aligned} \begin{bmatrix} \Delta SALES_t \\ \Delta OMKT_t \\ \Delta CMKT_t \end{bmatrix} &= \begin{bmatrix} \alpha_{SALES} & 0 & 0 \\ 0 & \alpha_{OMKT} & 0 \\ 0 & 0 & \alpha_{CMKT} \end{bmatrix} \begin{bmatrix} e_{Sales,t-1} \\ e_{OMKT,t-1} \\ e_{CMKT,t-1} \end{bmatrix} + \sum_{i=1}^p \Phi_i \\ &\times \begin{bmatrix} \Delta SALES_{t-i} \\ \Delta OMKT_{t-i} \\ \Delta CMKT_{t-i} \end{bmatrix} + \Psi \times \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} + \begin{bmatrix} u_{Sales,t} \\ u_{OMKT,t} \\ u_{CMKT,t} \end{bmatrix} \end{aligned} \quad (9)$$

The addition of the error correction terms reflects the fact that in each period the system adjusts toward the long-run cointegration relationship, with the coefficients of the error-correction term reflecting the speed of adjustment toward the equilibrium. When the variables are evolving, omitting the long run cointegrating relationship will underestimate the effects of marketing mix variables on performance (Vanden Abeele 1994). Incorporating the long-run relationship among variables will allow an estimation of long-run elasticities as well as short-run elasticities (Srinivasan et al. 2000).

Typically, VAR modelers do not interpret the individual parameters themselves, but tend to focus on the impulse-response functions derived from these parameters. Impulse-response functions trace the incremental performance and spending implications of an initial one-period change in one of the support variables, over time. In so doing, they provide a concise summary of the information contained in the multitude of VAR parameters, a summary that lends itself well to a graphical and easy-to-interpret representation.

Policy Simulation Analysis

Impulse Response Functions (IRF)

An impulse-response function (IRF) traces the incremental effect of a one-unit (or one-standard-deviation) shock in one of the variables on the future values of the other endogenous variables (e.g., Srinivasan et al. 2000, 2004). IRFs can also

be viewed as the difference between two forecasts: a first forecast, based on an information set that does not take the marketing shock into account, and another prediction based on an extended information set that takes this shock into account. As such, IRFs trace the incremental effect of the marketing action reflected in the shock. Note that marketing actions (such as, for example, a price promotion) are operationalized as deviations from a benchmark, which is derived as the expected value of the marketing-mix variable (for example, the price) as predicted through the dynamic structure of the VAR model. Response functions are based on the estimated parameters of the full VARX model. Note from Eqs. (6) and (7) that VARX models capture immediate as well as lagged, direct as well as indirect interactions among the endogenous variables. Based on all these estimated reactions, the impulse response function estimates the net result of a “shock” to a marketing variable on the performance variables relative to their baselines (their expected values in the absence of the marketing shock). Specifically, it measures the long-term performance response to a one-unit shock (Pauwels et al. 2002; Nijs et al. 2001; Srinivasan et al. 2004).

Starting from the reduced-form model specification in Eq. (7), we can substitute each lag of each endogenous variable by the same equation and thus express the right-hand side as a function of only current and lagged values of the error terms. This yields the Vector Moving Average (VMA) representation:

$$Y_t = A + (I - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p)^{-1} \Sigma_t \tag{10}$$

In words, each endogenous variable is explained by a weighted average of current and past errors or “shocks” both to itself and to the other endogenous variables. Therefore, we operationalize a change to a variable (e.g., a price promotion) as a shock to the variable series (e.g., to sales). An impulse response function then tracks the impact of that shock to each variable in the system (price, sales, competitive price, etc.) during the shock (typically denoted as period 0) and for each period thereafter.

For a simple illustration of the IRFs, let’s consider the VAR model below:

$$\begin{bmatrix} SALES_t \\ OMKT_t \\ CMKT_t \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \begin{bmatrix} SALES_{t-1} \\ OMKT_{t-1} \\ CMKT_{t-1} \end{bmatrix} + \begin{bmatrix} u_{Sales,t} \\ u_{OMKT,t} \\ u_{CMKT,t} \end{bmatrix} \tag{11}$$

For a unit shock to own marketing action *OMKT*, one sets $[u_{Sales}, u_{OMKT}, u_{CMKT}] = [0, 0, 0]$ prior to *t*; to $[0, 1, 0]$ at time *t*; and to $[0, 0, 0]$ after time *t*. One then computes (simulates) the future values for the various endogenous variables:

$$\begin{bmatrix} SALES_t \\ OMKT_t \\ CMKT_t \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} SALES_{t+1} \\ OMKT_{t+1} \\ CMKT_{t+1} \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \Phi_{12} \\ \Phi_{22} \\ \Phi_{32} \end{bmatrix}$$

$$\begin{aligned} \begin{bmatrix} SALES_{t+2} \\ OMKT_{t+2} \\ CMKT_{t+2} \end{bmatrix} &= \begin{bmatrix} \Phi_{11} & \Phi_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & \Phi_{23} \\ \Phi_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \begin{bmatrix} \Phi_{12} \\ \Phi_{22} \\ \Phi_{32} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \Phi_{11}\Phi_{12} + \Phi_{12}\Phi_{22} + \Phi_{13}\Phi_{32} \\ \Phi_{21}\Phi_{12} + \Phi_{22}\Phi_{22} + \Phi_{23}\Phi_{32} \\ \Phi_{31}\Phi_{12} + \Phi_{32}\Phi_{22} + \Phi_{33}\Phi_{32} \end{bmatrix} \end{aligned}$$

etc.

A plot of these forecasts against time yields the impulse response function; allowing all endogenous variables to respond according to the historically observed reaction patterns, as captured by all estimated VAR-coefficients. In case the affected variable is evolving (has a unit root), the shock may (but does not need to) have a permanent impact, i.e., the variable does not return to its pre-shock level. In the typical case of stationary variables, these shock effects die out, i.e., the permanent impact is 0 and the variable returns to its steady-state, i.e., pre-shock level.

How can IRFs show permanent effects if the VAR-model can only include stationary variables? If the performance variable (e.g., sales) is evolving, we indeed include it in the model in first differences, i.e., in *changes* to the variable. Therefore, the impulse response function on this variable (change in sales) will die out. To derive the effect on sales itself, we must accumulate the IRF values starting from the first period. Thus, the IRFs for evolving variables will converge to the persistent impact and the IRFs for stationary variables will return to their baseline with permanent impact of 0. What happens when the impulse (i.e., marketing) is the evolving variable? Again, we include the marketing action (e.g., price) in first differences in the model, and its IRF therefore shows the performance impact not of a temporary shock but of a permanent change (e.g., reduction in regular price). We need to keep this in mind when interpreting the result.

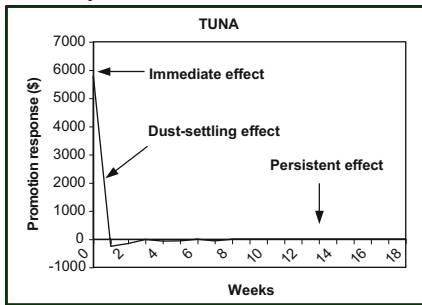
Researchers often derive the three summary statistics from IRFs using the VAR model: (1) the immediate or short-term performance impact of own marketing (*MKT*) or competitive marketing (*CMKT*) on brand sales (*SALES*), which is readily observable to managers, and may therefore receive considerable managerial scrutiny; (2) the persistent or the long-term impact (i.e., the value to which the IRF converges); and (3) the total or cumulative impact, which combines the immediate effect with all effects across the dust-settling period. In the absence of permanent effects, this total impact becomes the relevant metric to evaluate performance outcomes (Pauwels et al. 2002; Pauwels and Srinivasan 2004).

Srinivasan et al. (2004) answer the question of whether of price promotions benefit manufacturers and retailers by analyzing 7 years of scanner data, covering

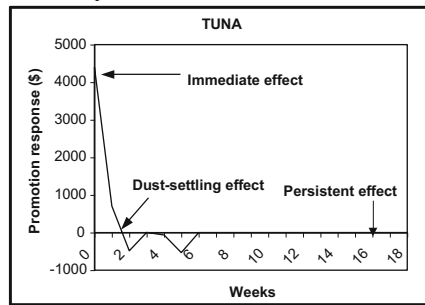
25 product categories and 75 brands, from the Chicago area’s second-largest supermarket chain, Dominick’s Finer Foods. Previous research showed that price promotions tend to have little long-term effect on sales volume. Their research found that the same is true for revenues and margins. Figure 2 shows the impulse response functions of the impact of price promotion on manufacturer and retailer revenues for two categories, tuna and cheese, from the 25 categories that they study. It outlines the immediate, the dust-settling, and the persistent effects of promotions for these two categories.

During the 1-week promotion, the cheese manufacturer saw an immediate revenue increase as customers bought more of its brand. But the retailer saw a loss, because gains from increased sales of the promoted brand were more than offset by loss of sales from regularly priced brands. During the dust-settling weeks, 2–6, the manufacturer saw a negative impact on revenue as customers switched back to their usual brands and toward competing brands that had launched their own promotions.

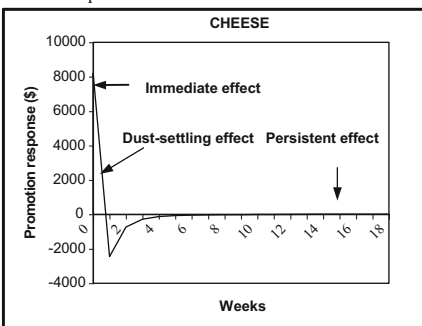
A: Impulse response function of a price promotion of one cent per ounce on manufacturer revenue



B: Impulse response function of a price promotion of one cent per ounce on retailer revenue



C: Impulse response function of a price promotion of one cent per ounce on manufacturer revenue



D: Impulse response function of a price promotion of one cent per ounce on retailer revenue

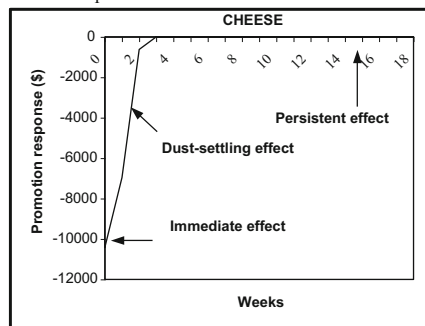


Fig. 2 Impulse-response functions. (a) Impulse response function of a price promotion of 1 cent per ounce on manufacturer revenue. (b) Impulse response function of a price promotion of 1 cent per ounce on retailer revenue. (c) Impulse response function of a price promotion. (d) Impulse response function of a price promotion of 1 cent per ounce on manufacturer revenue of 1 cent per ounce on retailer revenue

Meanwhile, retailer revenue for the cheese category gradually moved back to baseline as the promotion effects tailed off. By week 6, manufacturer and retailer revenues had returned to their prepromotion levels and remained stable through week 26. They find that promotions of frequently promoted brands, for example, tend to have a positive short-term effect on both retailers' and manufacturers' revenues but a negative impact on retailers' profit margins. Thus, the interests of manufacturers and retailers may well be aligned for one financial metric, such as revenue, but not for another, such as profit. Impulse Response Functions can therefore be used to obtain insights into impact on relevant performance metrics. Researchers can also obtain the wear-in time of each driver's effect on performance as the period with the highest (in absolute value) impulse response coefficient (Pauwels and Hanssens 2007; Srinivasan et al. 2010).

An approach to identify the shocks of a VAR model is to use orthogonal impulse responses where the basic idea is to decompose the variance-covariance matrix so that $\Sigma = PP'$, where P is a lower triangular matrix with positive diagonal elements, which is often obtained by a Cholesky decomposition. Note that the output of the Cholesky decomposition is a lower triangular matrix so that the variable in the first row will never be sensitive to a contemporaneous shock of any other variable and the last variable in the system will be sensitive to shocks of all other variables. Therefore, the results of orthogonal impulse response analysis might be sensitive to the order of the variables and it is advised to estimate the VAR model with different orders to see how strongly the resulting IRFs are affected by ordering.

Generalized IRFs (GIRF) can be obtained with the simultaneous-shocking approach (Evans and Wells 1983; Dekimpe and Hanssens 1999), in which the information in the residual variance-covariance matrix of Eq. (7) is used to derive a vector of *expected* instantaneous shock values. The advantage of this approach is that it does not require selecting a temporal ordering among the variables of interest. Standard errors are subsequently derived using the Monte Carlo simulation approach with 250 runs in each case (see Horváth et al. 2005). The GIRF estimates, given a one-unit shock to variable I , the expected value for shocks occurring simultaneously to the other variables j ($i \neq j$) and is shown in Eq. (12) below:

$$E [u_j | u_i = 1] = \sigma_{ij} / \sigma_{ii}; \quad \text{with } \sigma_{ij}; \quad \sigma_{ii} \text{ elements of } \Sigma \quad (12)$$

(from Eq. 7).

As explained in Dekimpe and Hanssens (1999) and Nijs et al. (2001), we can now calculate generalized IRFs (GIRFs), which do not depend on a causal ordering. When the causal ordering is clear (e.g., retail prices that cannot be changed by manufacturers for weeks; Leeflang and Wittink 1996), GIRFs lead to the same inferences as IRFs. GIRFs are particularly important when prior theory or observation does not suggest a clear causal ordering, e.g., among different marketing actions, competitors, or online customer actions (e.g., DeHaan et al. 2016), and typically most marketing papers tend to use GIRFs.

Dynamic Multipliers

The marginal impact of changes in the exogenous variables can be investigated with the help of dynamic multiplier analysis. For example, if the exogenous variables are marketing actions, such as promotional display or feature variables, the consequences of changes in these actions can be analyzed (if they are endogenous we apply impulse response analysis described earlier). VAR models with exogenous variables (i.e., VARX) as shown in Eqs. (6) and (7) can be expressed in the following way:

$$B(L)Y_t = A + C(L)X_t + u_t, \quad t = 1, \dots, T \tag{13}$$

where X_t is an h -dimensional vector of exogenous variables and $C(L)$ is a matrix polynomial with lag operator L : $C(L) = C_0 - C_1L - C_2L^2 - \dots - C_SL^S$ and C_i are $h \times k$ coefficient matrices, $i = 0, \dots, S$. The model is referred to as a *VARX* (P, S) process. If u_t is an *MA*(Q) process the model becomes a *VARMAX* (P, S, Q) process.

Dynamic multiplier analysis can be used for policy simulation. For instance, a brand manager may want to know about the consequences of an increase in promotions of its brand over time. Or, the effects of changes in exogenous variables that are not under control of any decision maker may be of interest (see Horvath 2003). The dynamic effects of exogenous variables on the endogenous variables are captured by the dynamic multipliers (Lütkepohl 1993, p. 338):

$$D(L) = \sum_{i=0}^{\infty} D_i L^i = B(L)^{-1} C(L) \tag{14}$$

where $B(L)$ and $C(L)$ are defined in Eq. (13). From this representation, the response of $Y_{i,t+\tau}$ to a unit change in X_{jt} can easily be obtained:

$$\frac{\partial Y_{i,t+\tau}}{\partial X_{j,t}} = d_{ij,\tau}, \quad t = 1 \dots \tau \tag{15}$$

where $\delta_{ij,\tau}$ is the row i th, column j th element of the $h \times k$ matrix of coefficients D_τ , the coefficient matrix of the τ -th lag in Eq. (14), $\tau = 0, \dots, \infty$.

Granger Causality Tests: Do We Need to Model a Dynamic System?

Granger causality implies that knowing the history of a variable X helps explain a variable Y , beyond Y 's own history. This temporal causality is the closest proxy for causality that can be gained from studying the time series of variables (i.e., in the absence of manipulating causality in controlled experiments). Granger causality tests, for instance, allow the testing of the traditional market response assumption that brand performance is driven by marketing actions. Srinivasan

et al. (2010) use Granger causality tests to assess whether marketing performance could be driven by past performance, by own and competitive mindset metrics, and by own and competitive marketing actions. Franses (1998) notes that not accounting for this marketing endogeneity may lead to substantially wrong conclusions about marketing effectiveness. Unfortunately, marketing (and economic) theory is typically insufficient to correct a priori specification of such models, and their identification thus requires “incredible identifying restrictions” to which Sims (1980) objected. In the absence of strong theoretic rationale to exclude specific directions of causality, he prefers to establish them empirically using the available data (see Pauwels 2018).

Specifically, one can test for the presence of endogeneity among all variables with Granger causality tests (Granger 1969). As Pauwels (2018) insightfully notes “temporal causality” is the closest proxy for causality that can be gained from studying the time series of the variables (i.e., in the absence of manipulating causality in controlled experiments). In other words, a variable x Granger causes a variable y if knowing the past of x improves the forecast for y based on only the past of y . Formally, x Granger causes y if, at the 5% significance level:

$$\text{MSFE}(y_t|y_{t-1}, \dots, y_{t-k}, x_{t-1}, \dots, x_{t-m}) < \text{MSFE}(y_t|y_{t-1}, \dots, y_{t-k}) \quad (16)$$

with MSFE = mean squared forecast error and k and m the maximum lags for y and x . An important caveat is that Granger causality tests are pairwise, i.e., they can indicate variable x is Granger causing y while instead they are both being driven (x earlier than y) by a third variable z . Accordingly, the researcher should develop a full understanding of the web of Granger causality and how performance can be affected.

Several marketing papers have applied Granger Causality tests to assess the issue of temporal causality.

Hanssens (1980) applied Granger Causality tests to sort out patterns of competitive interactions in the airline industry. Trusov et al. (2009) find that offline events organized by a large social media company increased the number of online friend referrals they received. Therefore, these organized events had a higher total ROI than would be calculated from their direct performance effects. Srinivasan et al. (2010), in their Granger-causality tests, show that marketing actions and mind-set metrics more often Granger-cause sales than vice versa. Awareness, consideration, and liking Granger-cause sales for, respectively, 73%, 71%, and 63% of all brands, and sales Granger-causes the mind-set metrics for, respectively, 52%, 60%, and 51% of all brands. Recently, Kireyev et al. (2016) find that display impressions Granger cause Search Impressions and Search Clicks.

Granger Causality tests are useful to assess whether managers react to market performance when deciding on marketing actions. As Pauwels (2018) notes the feedback could be from actions to performance and vice versa since high sales may induce marketing spending (e.g., sales may Granger cause advertising) and low sales may induce management to take course corrections by changing marketing spending (e.g., low sales may Granger cause price promotions). Horváth et al. (2005) consider the relative importance of such feedback in their model. Overall, Granger Causality

has been used extensively in time-series applications in marketing to shed light on the direction of causality, as outlined above.

Forecast Error Variance Decomposition (FEVD)

Based on the VARX parameters, researchers also obtained the Forecast Error Variance Decomposition (FEVD) estimates to investigate whether, for example, own and competitive actions explain brand sales performance beyond the impact of past brand sales performance. FEVD quantifies the dynamic explanatory value on performance of each endogenous variable. Akin to a “dynamic R^2 ,” the FEVD provides a measure of the relative impact over time of shocks initiated by each of the individual endogenous variables in a VARX model (Pesaran and Shin 1998; Nijs et al. 2007).

The idea behind FEVD is to stimulate a “typical” shock on the fully estimated system, realize a forecast up to a chosen horizon, and then decompose the variance of the forecast error. The “typical” shock is simulated on the residuals which are contemporaneously correlated. As a result, the impact of a simulated shock is likely to incorporate the degree of correlation between the error terms. Therefore, the influence of a shock cannot be completely attributable to a precisely defined variable of the model. Cholesky orthogonalization of the error terms offers a way to overcome this problem by rewriting the system to impose a causal ordering. The procedure is however sensitive to the way in which variables enter the system. The first variable in the model is allowed to affect all the variables whereas the second variable affects all the variables except the first one, and so on. This is equivalent to imposing a hierarchy of effects to aid FEVD interpretation. I refer the interested reader to Valenti et al. (2020) who perform a Cholesky FEVD to investigate advertising’s hierarchy of effects by imposing a causal structure to advertising and intermediate factors.

The Generalized Forecast Error Variance Decomposition (GFEVD) is order-invariant like the GIRFs are and can be derived using the following equation:

$$\theta_{ij}^g(n) = \frac{\sum_{l=0}^n (\psi_{ij}^g(l))^2}{\sum_{l=0}^n \sum_{j=0}^m (\psi_{ij}^g(l))^2}, \quad i, j = 1, \dots, m. \quad (17)$$

where $\psi_{ij}^g(l)$ is the value of a Generalized Impulse Response Function (GIRF) following a one-unit shock to variable i on variable j at time l (Pesaran and Shin 1998). In GFEVD an initial shock can (but need not, depending on the size of the corresponding residual correlation) affect all other endogenous variables instantaneously. This has been applied in a marketing setting by Nijs et al. (2007). Importantly, the GFEVD attributes 100% of the forecast error variance in performance to either (1) the past values of the other endogenous variables or (2) the past of performance itself. The former (e.g., does a past change in awareness drives current

sales) is much more managerially and conceptually interesting than the latter (a past change in sales drives current sales, but we do not know what induced that past change in sales). One can assess the dynamic explanatory value of the marketing and competitive marketing by the extent to which they increase the sales forecast error variance explained by the potential drivers of sales, and thus reduce the percentage explained by past sales. The relative importance of the drivers established is typically based on the GFEVD values at 6 months, which reduces sensitivity to short-term fluctuations. Studies have shown that a period of 26 weeks (6 months) is sufficient for stationary series in consumer-packaged goods to capture dynamic effects (Pauwels and Srinivasan 2004; Srinivasan et al. 2004).

To evaluate the accuracy of the GFEVD estimates, standard errors can be obtained using Monte Carlo simulations (see Benkwitz et al. 2001). While GFEVD is the appropriate method to assess the dynamic R-squared, it does come at a cost: it only allows comparable analyses of brands with stationary variables because the variance for evolving variables is (theoretically) infinite (Pesaran and Shin 1998; Srinivasan et al. 2008).

(Generalized) Forecast Error Variance Decomposition always sums up to 100%, with typically the own past of the focal variable, explaining most of its variance. In some marketing applications, the % of “inertia” is of special importance, e.g., indicating price inertia in Nijs et al. (2007) and Srinivasan et al. (2008). In most others though, it is the least interesting of % categories, and thus gets reduced from the 100% to yield the % of performance explained by the other groups of variables. For instance, Srinivasan et al. (2010) show how adding mindset metrics improves the % of the GFEVD not explained by own past performance, while Srinivasan et al. (2016) and Colicev et al. (2018) show how adding online behavior metrics (owned, earned, and paid) does the same in a multichannel context.

Turning to the applicability of VAR models, researchers have successfully applied VAR models to a variety of datasets, including annual observations (e.g., Baghestani 1991), decades of quarterly observations (e.g., Pauwels et al. 2004), several years of weekly observations (e.g., Srinivasan et al. 2010), and a number of months of daily observations (e.g., Colicev et al. 2018). VAR models require an adequate number of degrees of freedom to provide reliable estimates (e.g., Colicev and Pauwels 2020). A rule of thumb is to have at least five observations per parameter (Leeftang et al. 2015), which practically translates into a minimum of about 50 time periods (e.g., 12 years of quarterly data or 5 years of monthly data) per firm.

Researchers additionally need to consider the sample size and the sampling frequency. Mitchell and James (2001) provide an overview of how data frequency can allow or jeopardize how researchers can establish causality among variables. For example, finance and marketing fields often deal with weekly, daily, and hourly data. Tellis and Franses (2006), in the context of the duration of advertising carryover effect on sales, argue that the optimal data interval for researchers to collect data is the unit exposure time. For example, if firms change pricing strategies once a quarter, the quarterly level data is appropriate for studying the impact of pricing changes.

Colicev and Pauwels (2020) advocate that the best approach would be to measure the variables as frequently as possible.

Finally, when the time-series data is available over an extended period, the variables may exhibit time-varying volatility, i.e., periods of swings interspersed with periods of relative calm. ARCH-type models are appropriate to investigate this explicitly, given that they model the variance of the current error term as a function of the size of previous error terms (see Colicev and Pauwels 2020 for a good discussion on this topic).

Software Programs for VAR Estimation

Modern time series packages are included in software as Stata, Matlab, Gauss, and R. Dedicated software packages include Time Series Processor (TSP), OX/PcGive, Eviews, and RATS. Stata is the most commonly used while Matlab has embedded functions not available in Stata, but requires more coding. R is versatile in its time series functions while TSP and RATS are dedicated to time series. Eviews and OX/PcGive provide an easy-to-use interface with click-and-find program options. Moreover, Eviews has the most typical VAR option as the default in its software and offers regular updates, adding the state-of-the-art tests and deleting less relevant ones. Its student version is a low-cost option to get started for a novice. I refer the interested reader to Colicev and Pauwels (2020) who employ a dataset, which combines public social media data from Facebook with corporate reputation data from a private data source, to illustrate the VAR model by explaining the key methodological steps needed to estimate and interpret the results through a software tutorial in R and STATA.

Illustrative Applications of VAR Modeling in Marketing

Empirical work using VAR models has expanded and is now part of the mainstream in marketing applications. Next, I will outline three substantive applications of VAR models in the following domains: (1) Investor Response Models, (2) Marketing Mix and Mindset Metrics Models, and (3) Digital Marketing Models.

Investor Response Models in the Marketing-Finance

VAR models are useful in modeling the marketing-finance interface since they use a system's representation (e.g., Dekimpe and Hanssens 1995a; Pauwels et al. 2002), in which each equation tracks the behavior of an important agent; for example, the consumer (demand equation), the manager (decision rule equations), competition (competitive reaction equation), and the investor (stock price equation). The long-run behavior of each endogenous variable is obtained from a shock-initiated chain reaction across the equations. For instance, a successful new-product introduction

will generate higher revenue, which may prompt the manufacturer to reduce sales promotions in subsequent periods. The combination of increased sales and higher margins may improve earnings and ultimately stock price. Because of such chains of events, the full performance implications of the initial product introduction may extend well beyond its immediate effects. As an example, a persistence model estimated as a vector autoregressive model (VAR) in differences can be specified for each brand (two in the illustration) of firm i , as follows:

$$\begin{bmatrix} \Delta MBR_{it} \\ \Delta INC_{it} \\ \Delta REV_{it} \\ MKT1_{it} \\ MKT2_{it} \end{bmatrix} = C + \sum_{n=1}^N B_n \times \begin{bmatrix} \Delta MBR_{it-n} \\ \Delta INC_{it-n} \\ \Delta REV_{it-n} \\ MKT1_{it-n} \\ MKT2_{it-n} \end{bmatrix} + \Gamma \times \begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} + \begin{bmatrix} uMBR_{it} \\ uINC_{it} \\ uREV_{it} \\ uMKT1_{it} \\ uMKT2_{it} \end{bmatrix} \quad (18)$$

with B_n , Γ vectors of coefficients, $[u_{MBR_{it}}, u_{INC_{it}}, u_{REV_{it}}, u_{MKT1_{it}}, u_{MKT2_{it}}]' \sim N(0, \Sigma_u)$, N the order of the system based on Schwartz' Bayes Information Criterion (SBIC), and all variables expressed in logarithms or their changes (Δ). In this system, the first equation is an expanded version of the stock-return response model (Srinivasan and Hanssens 2009a, b). The second and third equations explain the changes in, respectively, bottom-line (INC) and top-line financial performance (REV) of firm i . The fourth and fifth equations represent firm's marketing actions (e.g., for each brand), i.e., ($MKT1_{it}$) and ($MKT2_{it}$). For example, Pauwels et al. (2004) considered a brand's new-product introductions and sales promotions. The exogenous variables in this dynamic system (X_{1t} , X_{2t} , X_{3t}) could include controls such as the Carhart four factors and the impact of stock-market analyst earnings expectations. The impact of contemporaneous shocks is incorporated through the elements of Σ_u . As described earlier, such models provide baseline forecasts of each endogenous variable, along with estimates of the shock or surprise component in each variable.

Several applications of the VAR model exist in the marketing-finance domain. Pauwels et al. (2004) assess investor reactions to auto companies' new product introductions with price promotions to find that new product introductions have a gradually increasing influence on stock price, all else being equal, while price promotions generally lower firm value, even though they may successfully stimulate sales. Thus, investors view new product activity as generating long-term value and promotions as destroying long-term value. They show that investors in the automotive industry need about 6 weeks to fully incorporate the impact of a new-product introduction on stock returns. Joshi and Hanssens (2010) have found that advertising in the PC industry has a small but positive long-term effect on stock prices, again after controlling for advertising's direct impact on sales and profits.

In a recent application of VARX models to the marketing-finance interface, Colicev et al. (2018) examine the effects of owned and earned social media on brand awareness, purchase intent, and customer satisfaction and link these consumer mindset metrics to shareholder value metrics including abnormal returns and idiosyncratic risk. Analyzing daily data for 45 brands in 21 sectors using vector

autoregression models, they find that brand fan following improves customer mindset metrics, and that purchase intent and customer satisfaction positively affect shareholder value. This chapter represents an application of VAR modeling that combines mindset metrics with the marketing-finance interface.

When authors want to combine the results from multiple entities (e.g., firms) together and provide an overall picture of the analysis, Panel Vector Autoregression (PVAR) specifications are useful (e.g., Holtz-Eakin et al. 1988). Kang et al. (2016) use a structural panel VAR to show that corporate social responsibility actions are not driven by slack resources, but by an (only partially successful) attempt to make up for past social irresponsibility.

Marketing and Mindset Metrics Models

Papers in this stream research the question of whether it is useful or not to include marketing mix actions and customers mindset metrics into one overall model to explain brand performance. Specific questions addressed include: (1) Does the addition of mindset metrics to a sales model that already includes marketing mix actions enhance explanatory power? (2) And, if so, does this inclusion help in understanding how marketing actions drive sales? Using a 4-weekly data set with comprehensive information on performance metrics, marketing mix, and mindset metrics for over 60 brands in four fast-moving consumer goods categories over a period of 7 years, Srinivasan et al. (2010) estimate Vector Autoregressive (VARX) models to find that addition of mindset metrics to a sales model that already includes marketing mix significantly enhances explanatory power in predicting brand sales.

In a recent application, Valenti et al. (2020) examine where there is a dominant hierarchical sequence on how advertising influences purchases, given that it changes how consumers think and feel about brands. While the Hierarchy of Effects (HoE) model has guided advertising decisions for decades, there is little support for any hierarchy, thus suggesting the death of HoE. To answer the question of whether a hierarchical sequence holds for advertising effects, they undertake a large-scale VAR-based econometric analysis in which they compare 13 alternative hierarchies, each in two different versions (correlated and orthogonal errors), leading to 26 models proposed in previous literature. These hierarchies come in three types: The Classical HoE, a Simultaneous HoE (based on Vakratsas and Ambler 1999), and an Integrated HoE (based on Bruce et al. 2012). They estimate the corresponding models (involving restrictions on the VAR parameters depending on the sequence) for the top brands over 150 brands in different consumer packaged goods categories to find that the death of the HoE has been greatly exaggerated. They find that the Integrated HoE sequence fits better than any alternative. The sequence of the hierarchy differs by brand, with Affect→Cognition→Experience being the most common, and they identify moderators of the sequence including brand and category characteristics. This chapter offers important findings for brand managers, especially as it counters a prevailing belief in the advertising literature that there exists little support for a hierarchical sequence.

Digital Marketing Models

With the growth of digital marketing, there is considerable interest in examining the impact of offline and online marketing on market performance, and their interactions among each other. Srinivasan et al. (2016) quantify the role of online consumer activity measured by paid, owned, earned, and unearned media metrics in driving sales within the context of traditional marketing mix variables price, distribution, and advertising. Based on the Vector Autoregressive (VAR) model, they derive Generalized Forecast Error Variance Decompositions (GFEVD) and Generalized Impulse Response Functions (GIRF) to quantify the elasticity and relative influence of consumer activity and traditional marketing mix actions on sales for a consumer-packaged good product. Beyond establishing online consumer activity metrics as leading sales indicators, their study also shows that even small changes to online engagement metrics can lead to sales declines. Pauwels and van Ewijk (2013) also show how adding online behavior metrics (owned, earned, and paid) does the same, also using VAR models. As a potential wellspring of strategic intelligence, tracking online consumer activity metrics could prove instrumental in expanding the role of marketing in corporate decision making in practice.

Consumers are regularly exposed to negative information about brands through word-of-mouth, news, reviews, and social media. Prior literature on consumers' response to negative brand information has shown that when more negative information is available about a brand, sales are depressed. In contrast, Han et al. (2019) find that an increase in negative information about a brand may lead to an increase in brand awareness and purchase intent for the brand. Using 4 years of weekly survey data tracking customers' attitudes toward computer and automobile brands, they estimate VARX models that relate a survey measure of exposure to negative information about a brand (negative buzz) with brand awareness, positive feeling toward the brand, and purchase intent for the brand. As expected, for automotive brands, they find that a shock in negative buzz leads to higher brand awareness and negative effects on positive feeling and purchase intent. However, for computers, they find that an increase in negative buzz is followed by increases in awareness, positive feeling, and purchase intent. This research therefore suggests there are circumstances when negative buzz should not be suppressed.

Pauwels et al. (2016) have a novel application of Bayesian VARs in a multichannel (offline/online) setting. Unrestricted estimation of VAR models risks over-parametrization because the parameter space proliferates with the number of endogenous variables. In a standard VAR model, a large number of parameters may produce a good model fit, but still result in multicollinearity and loss of degrees of freedom, which in turn may lead to inefficient estimates and poor performance in the impulse-response functions. Bayesian models alleviate such issues thanks to shrinkage, which imposes restrictions on the parameters of the VAR model. Bayesian Vector Autoregressive (BVAR) models are formulated in Litterman (1986) and Doan et al. (1984) but have seen little application in marketing (for an exception see Horvarth and Fok 2013; Pauwels et al. 2016). Several priors have been used in the econometrics literature to estimate the Bayesian VAR models, including Minnesota prior and NormalWishart prior. Pauwels et al. (2016) estimate the BVAR model through the

“mixed estimation” technique developed by Theil and Goldberger (1961), which involves supplementing data with prior information on the distributions of the coefficients. Their results indicate that within-online synergy is higher than online-offline synergy for familiar brands but not for unfamiliar brands. Managers of unfamiliar brands may obtain substantial synergy from offline marketing spending, even though its direct elasticity pales in comparison with that of online media while managers of familiar brands can generate more synergy by investing in different online media.

Conclusion

Vector-autoregressive models have come a long way in terms of their scope and scale of applications in marketing, not only because more extensive data sets have become available, but also because of growing interest in marketing on research (1) that potentially involve multiple dynamic feedback loops, and (2) where marketing theory is insufficiently developed to specify *a priori* all temporal precedence relationships (Dekimpe and Hanssens 2018). In those instances, the flexibility of VAR models to capture dynamic inter-relationships, and to quantify the short- and long-run net effects of the various influences at hand, renders them more appropriate. Such models allow researchers to obtain the short-run and the long-run impact of marketing on business performance. Furthermore, VAR models acknowledge and incorporate the idea that the impact of marketing actions is determined by the interplay between the responses of consumers, firm, competitors, investors, and other stakeholders. Given the growth of data availability particularly in digital settings, the marketing field will see a continued growth of digital applications of VAR models in the coming years. I hope the current chapter will contribute to a further adoption and diffusion of these techniques in the marketing community.

Cross-References

- ▶ [Applied Time-Series Analysis in Marketing](#)
- ▶ [Assessing the Financial Impact of Brand Equity with Short Time-Series Data](#)
- ▶ [Return on Media Models](#)

Acknowledgment I am grateful to my coauthors including Frank Bass, Dominique Hanssens, Koen Pauwels, Gokhan Yildirim, Marnik Dekimpe, Philip Hans Franses, Albert Valenti, and Elea Feit, among others for insights that I have gathered over the years in our joint work on VAR models.

References

- Amisano, G., & Giannini, C. (1997). *Topics in structural VAR economics*. Berlin: Springer.
- Baghestani, H. (1991). Cointegration analysis of the advertising-sales relationship. *The Journal of Industrial Economics*, 671–681.
- Bass, F. M., & Clarke, D. G. (1972). Testing distributed lag models of advertising effect. *Journal of Marketing Research*, 9(3), 298–308.

- Bass, F. M., & Pilon, T. L. (1980). A stochastic brand choice framework for econometric modeling of time series market share behavior. *Journal of Marketing Research*, 486–497.
- Benkowitz, A., Lütkepohl, H., & Wolters, J. (2001). Comparison of bootstrap confidence intervals for impulse responses of German monetary systems. *Macroeconomic Dynamics*, 5(1), 81–100.
- Bernanke, B. S. (1986). Alternative explanations of the money-income correlation. In *Carnegie-Rochester conference series on public policy* (Vol. 25, pp. 49–99). North-Holland.
- Bronnenberg, B. J., Mahajan, V., & Vanhonacker, W. R. (2000). The emergence of market structure in new repeat-purchase categories: The interplay of market share and retailer distribution. *Journal of Marketing Research*, 37(1), 16–31.
- Bruce, N. I., Peters, K., & Naik, P. A. (2012). Discovering how advertising grows sales and builds brands. *Journal of Marketing Research*, 49(6), 793–806.
- Colicev, A., Malshe, A., Pauwels, K., & O'Connor, P. (2018). Improving consumer mind-set metrics and shareholder value through social media: The different roles of owned and earned media. *Journal of Marketing*, 82(1), 37–56.
- Colicev, A., & Pauwels, K. (2020). Multiple Time Series Analysis for organizational research. *Long Range Planning*, forthcoming. <https://doi.org/10.1016/j.lrp.2020.102067>.
- De Haan, E., Wiesel, T., & Pauwels, K. (2016). The effectiveness of different forms of online advertising for purchase conversion in a multiple-channel attribution framework. *International Journal of Research in Marketing*, 33(3), 491–507.
- Dekimpe, M. G., & Hanssens, D. M. (1995a). The persistence of marketing effects on sales. *Marketing Science*, 14(1), 1–21.
- Dekimpe, M. G., & Hanssens, D. M. (1995b). Empirical generalizations about market evolution and stationarity. *Marketing Science*, 14(3 Suppl), G109–G121.
- Dekimpe, M. G., & Hanssens, D. M. (1999). Sustained spending and persistent response: A new look at long-term marketing profitability. *Journal of Marketing Research*, 397–412.
- Dekimpe, M. G., & Hanssens, D. M. (2004). Persistence modeling for assessing marketing strategy performance. In C. Moorman & D. R. Lehmann (Eds.), *Assessing marketing strategy performance*. Cambridge, MA: Marketing Science Institute.
- Dekimpe, M. G., & Hanssens, D. M. (2018). Time series models of short-run and long-run marketing impact. In N. Mizik & D. M. Hanssens (Eds.), *Handbook of marketing analytics: Methods and applications in marketing management, public policy, and litigation support*. Edward Elgar.
- Dekimpe, M., Hanssens, D., & Silva-Risso, J. (1999). Long-run effects of price promotions in scanner markets. *Journal of Econometrics*, 89(1), 2.
- Deleersnyder, B., Geyskens, I., Gielens, K., & Dekimpe, M. G. (2002). How cannibalistic is the Internet channel? A study of the newspaper industry in the United Kingdom and the Netherlands. *International Journal of Research in Marketing*, 19(4), 337–348.
- Dickey, D., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427–431.
- Doan, T., Litterman, R. B., & Sims, C. A. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3, 1–100.
- Enders, W. (2003). *Applied econometric time series*. Wiley.
- Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 251–276.
- Engle, R. F., & Yoo, B. S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, 35(1), 143–159.
- Evans, L., & Wells, G. (1983). An alternative approach to simulating VAR models. *Economic Letters*, 12(1), 23–29.
- Franses, P. H. (1998). *Time series models for business and economic forecasting*. Cambridge: Cambridge University Press.
- Franses, P. H., Kloek, T., & Lucas, A. (1999). Outlier robust analysis of long-run marketing effects for weekly scanner data. *Journal of Econometrics*, 89(1–2), 293–315.

- Franses, P. H., Srinivasan, S., & Boswijk, P. (2001). Testing for unit roots in market shares. *Marketing Letters*, 12(4), 351–364.
- Gijsenberg, M. J., Van Heerde, H. J., & Verhoef, P. C. (2015). Losses loom longer than gains: Modeling the impact of service crises on perceived service quality over time. *Journal of Marketing Research*, 52(5), 642–656.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 424–438.
- Han, J. A., Feit, E. M., & Srinivasan, S. (2019). Can negative buzz increase awareness and purchase intent? *Marketing Letters*, 1–16.
- Hanssens, D. M. (1980). Market response, competitive behavior, and time series analysis. *Journal of Marketing Research*, 470–485.
- Hanssens, D. M. (1998). Order forecasts, retail sales, and the marketing mix for consumer durables. *Journal of Forecasting*, 17(34), 327–346.
- Hanssens, D. M., Parsons, L. J., & Schultz, L. (2001). *Market response models: Econometric and time series analysis*. Springer Science and Business Media.
- Hanssens, D. M., Pauwels, K. H., Srinivasan, S., Vanhuele, M., & Yildirim, G. (2014). Consumer attitude metrics for guiding marketing mix decisions. *Marketing Science*, 33(4), 534–550.
- Heerde, H. V., Srinivasan, S., & Dekimpe, M. (2010). Estimating cannibalization rates for pioneering innovations. *Marketing Science*, 29(6), 1024–1039.
- Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating vector autoregressions with panel data. *Econometrica: Journal of the econometric society*, 1371–1395.
- Horsky, D., & Simon, L. S. (1983). Advertising and the diffusion of new products. *Marketing Science*, 2(1), 1–17.
- Horváth, C. (2003). *Dynamic analysis of marketing systems*. Doctoral Thesis, University of Groningen. Alblasterdam: Labyrinth Publication.
- Horváth, C., & Fok, D. (2013). Moderating factors of immediate, gross, and net cross-brand effects of price promotions. *Marketing Science*, 32(1), 127–152.
- Horváth, C., Leeflang, P. S., Wieringa, J. E., & Wittink, D. R. (2005). Competitive reaction-and feedback effects based on VARX models of pooled store data. *International Journal of Research in Marketing*, 22(4), 415–426.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2), 231–254.
- Johansen, S., Mosconi, R., & Nielsen, B. (2000). Cointegration analysis in the presence of structural breaks in the deterministic trend. *The Econometrics Journal*, 3(2), 216–249.
- Joshi, A., & Hanssens, D. M. (2010). The direct and indirect effects of advertising spending on firm value. *Journal of Marketing*, 74(1), 20–33.
- Kang, C., Germann, F., & Grewal, R. (2016). Washing away your sins? Corporate social responsibility, corporate social irresponsibility, and firm performance. *Journal of Marketing*, 80(2), 59–79.
- Kireyev, P., Pauwels, K., & Gupta, S. (2016). Do display ads influence search? Attribution and dynamics in online advertising. *International Journal of Research in Marketing*, 33(3), 475–490.
- Kornelis, M., Dekimpe, M. G., & Leeflang, P. S. (2008). Does competitive entry structurally change key marketing metrics? *International Journal of Research in Marketing*, 25(3), 173–182.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1–3), 159–178.
- Leeflang, P., Wieringa, J. E., Bijmolt, T. H., & Pauwels, K. (2015). *Modeling markets*. New York: Springer.
- Leeflang, P. S., & Wittink, D. R. (1996). Competitive reaction versus consumer response: Do managers overreact? *International Journal of Research in Marketing*, 13(2), 103–119.
- Lim, J., Currim, I. S., & Andrews, R. L. (2005). Consumer heterogeneity in the longer-term effects of price promotions. *International Journal of Research in Marketing*, 22(4), 441–457.

- Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business and Economic Statistics*, 4, 25–38.
- Lütkepohl, H. (1993). *Introduction to multiple time series*. Berlin: Springer.
- Maddala, G. S., & Kim, I. M. (2007). *Unit roots, cointegration, and structural change* (No. 4). Cambridge University Press.
- Mela, C. F., Gupta, S., & Lehmann, D. R. (1997). The long-term impact of promotion and advertising on consumer brand choice. *Journal of Marketing Research*, 248–261.
- Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review*, 26(4), 530–547.
- Nijs, V. R., Dekimpe, M. G., Steenkamps, J. B. E., & Hanssens, D. M. (2001). The category-demand effects of price promotions. *Marketing Science*, 20(1), 1–22.
- Nijs, V. R., Srinivasan, S., & Pauwels, K. (2007). Retail-price drivers and retailer profits. *Marketing Science*, 26(4), 473–487.
- Osinga, E. C., Leeflang, P. S., & Wieringa, J. E. (2010). Early marketing matters: A time-varying parameter approach to persistence modeling. *Journal of Marketing Research*, 47(1), 173–185.
- Pauwels, K. (2018). Modeling dynamic relations among marketing and performance metrics. *Foundations and Trends in Marketing*, 11(4), 215–301.
- Pauwels, K., Demirci, C., Yildirim, G., & Srinivasan, S. (2016). The impact of brand familiarity on online and offline media synergy. *International Journal of Research in Marketing*, 33(4), 739–753.
- Pauwels, K., & Hanssens, D. M. (2007). Performance regimes and marketing policy shifts. *Marketing Science*, 26(3), 293–311.
- Pauwels, K., Hanssens, D. M., & Siddarth, S. (2002). The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. *Journal of Marketing Research*, 39(4), 421–439.
- Pauwels, K., Silva-Risso, J., Srinivasan, S., & Hanssens, D. M. (2004). New products, sales promotions, and firm value: The case of the automobile industry. *Journal of Marketing*, 68(October), 142–156.
- Pauwels, K., & Srinivasan, S. (2004). Who benefits from store brand entry? *Marketing Science*, 23(3), 364–390.
- Pauwels, K., & Van Ewijk, B. (2013). Do online behavior tracking or attitude survey metrics drive brand sales? An integrative model of attitudes and actions on the consumer boulevard. *Marketing Science Institute Working Paper Series*, 13.118, 1–49.
- Pauwels, K., & Weiss, A. (2008). Moving from free to fee: How online firms market to change their business model successfully. *Journal of Marketing*, 72(3), 14–31.
- Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 1361–1401.
- Perron, P. (1990). Tests of joint hypotheses in time series regression with a unit root. *Advances in Econometrics: Co-integration, Spurious Regression and Unit Roots*, 8, 10–20.
- Pesaran, M. H., Pierse, R., & Lee, K. C. (1993). Persistence, cointegration and aggregation: A disaggregated analysis of output fluctuations in the U.S. economy. *Journal of Econometrics*, 56, 57–88.
- Pesaran, H. H., & Shin, Y. (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters*, 58(1), 17–29.
- Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 1–48.
- Sims, C. A. (1986). Are forecasting models usable for policy analysis? *Federal Reserve Bank of Minneapolis Quarterly Review*, 10(1), 2–16.
- Slotegraaf, R. J., & Pauwels, K. (2008). The impact of brand equity and innovation on the long-term effectiveness of promotions. *Journal of Marketing Research*, 45(3), 293–306.
- Srinivasan, S., & Bass, F. M. (2000). Cointegration analysis of brand and category sales: Stationarity and long-run equilibrium in market shares. *Applied Stochastic Models in Business and Industry*, 16(3), 159–177.
- Srinivasan, S., & Hanssens, D. M. (2009a). Marketing and firm value: Metrics, methods, findings and future directions. *Journal of Marketing Research*, 46(3), 293–312.

- Srinivasan, S., & Hanssens, D. M. (2009b). Marketing et valeur de l'entreprise: mesures, méthodes, résultats et voies futures de recherche. *Recherche et Applications en Marketing*, 24(4), 97–130.
- Srinivasan, S., Pauwels, K., Hanssens, D. M., & Dekimpe, M. G. (2004). Do promotions benefit manufacturers, retailers, or both? *Management Science*, 50(5), 617–629.
- Srinivasan, S., Pauwels, K., & Nijs, V. (2008). Demand-based pricing versus past-price dependence: A cost-benefit analysis. *Journal of Marketing*, 72(2), 15–27.
- Srinivasan, S., Popkowski Leszczyc, P., & Bass, F. M. (2000). Market share response and competitive interaction: The impact of temporary, evolving and structural changes in prices. *International Journal of Research in Marketing*, 17(4), 281–305.
- Srinivasan, S., Rutz, O. J., & Pauwels, K. (2016). Paths to and off purchase: Quantifying the impact of traditional marketing and online consumer activity. *Journal of the Academy of Marketing Science*, 44(1), 440–453.
- Srinivasan, S., Vanhuele, M., & Pauwels, K. (2010). Mind-set metrics in market response models: An integrative approach. *Journal of Marketing Research*, 47(4), 672–684.
- Srivastava, V. K., & Giles, D. E. A. (1987). *Seemingly unrelated regression equations models*. New York: Marcel Dekker.
- Steenkamp, J. B. E., Nijs, V. R., Hanssens, D. M., & Dekimpe, M. G. (2005). Competitive reactions to advertising and promotion attacks. *Marketing Science*, 24(1), 35–54.
- Tellis, G. J., & Franses, P. H. (2006). Optimal data interval for estimating advertising response. *Marketing Science*, 25(3), 217–229.
- Theil, H., & Goldberger, A. S. (1961). On pure and mixed statistical estimation in economics. *International Economic Review*, 2, 65–78.
- Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an-internet social networking site. *Journal of Marketing*, 73(5), 90–102.
- Vakratsas, D., & Ambler, T. (1999). How advertising works: What do we really know? *Journal of Marketing*, 63(1), 26–43.
- Valenti, A., Yildirim, G., Vanhuele, M., Srinivasan, S., & Pauwels, K. (2020). *Is the hierarchy of effects in advertising dead or alive?*. Working paper. Cambridge, MA: Marketing Science Institute.
- Vanden Abeele, P. (1994). Commentary to: Diagnosing competition: Development and findings. In G. Laurent, G. L. Lillien, & B. Pras (Eds.), *Research traditions in marketing* (pp. 79–105). Boston: Kluwer Academic.
- Villanueva, J., Yoo, S., & Hanssens, D. M. (2008). The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. *Journal of Marketing Research*, 45(1), 48–59.
- Wiesel, T., Pauwels, K., & Arts, J. (2011). Practice prize paper-marketing's profit impact: Quantifying online and off-line funnel progression. *Marketing Science*, 30(4), 604–611.
- Zivot, E., & Andrews, D. W. (1992). Oil-price shock, and the unit-root. *Journal of Business and Economic Statistics*, 10(3).
- Zivot, E., & Andrews, D. W. (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics*, 20(1), 25–44.



Structural Equation Modeling

Hans Baumgartner and Bert Weijters

Contents

Introduction	550
The Core Structural Equation Model and Its Submodels	551
Model Estimation	556
Testing the Global Fit of Models	558
Respecifying Models That Do Not Pass the Global Fit Test	562
Assessing the Local Fit of Models	564
Measurement Model	564
Latent Variable Model	567
The Problem of Endogeneity	567
Extensions of the Core Structural Equation Model	569
Measurement Model Extensions	569
Latent Variable Model Extensions	571
Models That Incorporate Population Heterogeneity	572
Empirical Illustration of Structural Equation Modeling	574
Conceptual Model	574
Measurement Model	575
Latent Variable Model	579
Multi-Sample Analysis	581
Concluding Comments	581
Cross-References	583
References	584

H. Baumgartner (✉)

Smeal College of Business, The Pennsylvania State University, State College, PA, USA

e-mail: HansBaumgartner@psu.edu

B. Weijters

Faculty of Psychology and Educational Sciences, Department of Work, Organization and Society,
Ghent University, Ghent, Belgium

e-mail: bert.weijters@ugent.be

Abstract

This chapter presents an overview of the process of structural equation modeling, involving the steps of model specification, model estimation, overall fit evaluation, model respecification, and local fit assessment (including interpreting the parameters of the model). Various extensions of the core structural equation model are described to enable more general representations of measurement and latent variable models as well as applications of the model to heterogeneous populations. An empirical example is provided to illustrate the process of structural equation modeling and to demonstrate some of the complexities that may arise in practical applications.

Keywords

Structural equation modeling · Confirmatory factor analysis · Measurement models · Path analysis · Multi-sample analysis

Introduction

Only a few decades ago structural equation modeling (SEM) was regarded as an advanced statistical methodology that was used primarily by academic researchers to conduct sophisticated measurement analyses or to test the validity of theoretical models based on empirical data (Bagozzi 1980). Nowadays, SEM is a standard data analysis method that is employed widely by both academic and industry researchers (e.g., Chapman and Feit 2019). The success and rapid adoption of SEM is likely due to the following three reasons. First, SEM allows researchers to take into account measurement error (both random and systematic) when estimating correlations or structural relationships between constructs. Since observed measures in practical applications are usually measured with error, the suspect assumption of perfect measurement in conventional correlation and regression analysis is circumvented (see chapter ► [“Regression Analysis”](#) by Skiera, Reiner, and Albers, this volume, for a discussion of regression analysis). For example, it is unlikely that constructs such as the quality or value of a product or the satisfaction experienced by a customer can be measured well with single items, and even when multiple measures of these constructs are averaged (which corrects for unreliability of measurement to some extent), this does not provide much insight into the quality of measurement of the constructs by their measures. Second, researchers are often interested in estimating and testing models in which the dependence of multiple constructs on different sets of antecedents is modeled simultaneously and the process through which one construct influences another is investigated. SEM enables researchers to study complex conceptual frameworks in an integrative fashion and avoids the piecemeal testing of chains of effects as in conventional regression analysis. For example, a researcher may want to investigate both the antecedents of customer satisfaction such as expectations, perceived quality, and perceived value, and the consequences

of customer satisfaction such as loyalty and complaints, as well as the mediating role of satisfaction in this process. Third, researchers who want to study the invariance of model parameters across discrete populations (e.g., different demographic groups, industries, or countries) or test hypotheses about specific group differences can specify models for multiple populations in which both the homogeneity and heterogeneity of parameters can be investigated in a straightforward manner.

Although SEM is widely used, it is not always used well. A first goal of this chapter is to present an overview of the methodology, with special emphasis on issues that are sometimes misunderstood in applications of the technique (e.g., global fit testing). SEM is a rather vibrant research domain, and the core model has been extended in a variety of ways. A second goal of this chapter is to bring these developments to the attention of a wider audience and to encourage additional applications of SEM. Although we will not be able to cover advanced topics in using SEM in any detail, we will raise several important issues and point interested readers to the relevant literature. The final goal is to offer an empirical illustration of many of the issues discussed in this chapter, which will hopefully demonstrate the power of SEM for data analysis and convince researchers who have not used SEM to apply it in their own research. The data set used in the illustration and the code necessary to run the models in R, using the package *lavaan* (Rosseel 2012) and various supporting packages, are available for download on Github (<https://github.com/HansBaum129/SEM>).

The Core Structural Equation Model and Its Submodels

A full structural equation model consists of two parts: a model specifying the structural relationships between the substantive variables or constructs of interest (called the latent variable model because the constructs in one's model often cannot be observed and measured directly), and a model specifying the relationships between the constructs and their hypothesized observed (manifest) measures or indicators (called the measurement model). We will assume that the latent variable underlying a set of observed variables is equal to the construct of interest, and we will therefore use the terms construct and latent variable interchangeably. However, a researcher should carefully evaluate whether this assumption is justified when a structural equation model is specified for a particular substantive context. In the simplest case, the latent variable model is like a regression model (although the variables in the model are usually unobserved or latent), but in more complex models, the latent variable model could be comprised of a series of regression models, one for each construct to be explained in one's conceptual framework or theory. If a researcher is willing to assume that a construct is measured perfectly by a single indicator (where the single indicator could be an average of several observed measures), then an explicit measurement model is superfluous (i.e., the construct is identical to the observed measure). However, in many (most) cases, this is a tenuous assumption, and often (usually) researchers will want to specify a measurement

model which enables a thorough investigation of the measurement quality of the indicators of the constructs of interest.

Two types of measurement models can be formulated (e.g., MacKenzie et al. 2011). In a reflective measurement model, the observed variables (effect indicators) are specified as a function of (and thus a reflection of) hypothesized latent variables, which presumably represent the substantive variables the researcher is interested in. For example, a customer's satisfaction with a product may be measured with semantic differential scales such as satisfied-dissatisfied or happy-sad and these observed measures are assumed to be (fallible) reflections of respondents' satisfaction. In a formative measurement model, the observed variables (cause indicators) are specified as determinants of hypothesized latent variables, which means that constructs are formed by their indicators. For example, a customer's satisfaction with a service may depend on the friendliness, knowledgeable, and responsiveness of the sales staff and these service attributes are all assumed to contribute to a respondent's overall satisfaction. In general, a formatively measured construct is not completely captured by its indicators (i.e., the construct is measured with error), but sometimes it is assumed that the formative construct is equal to a linear combination of its indicators; Bollen (2011) refers to the two types of indicators as cause and composite indicators, respectively.

Formative indicators are frequently misspecified as reflective indicators (Jarvis et al. 2003; MacKenzie et al. 2005), and such measurement model misspecifications can have various negative consequences (see Diamantopoulos et al. 2008 for a summary). Indicators should therefore be evaluated carefully before a reflective measurement model is specified (usually almost by default). Jarvis et al. (2003) and MacKenzie et al. (2005) recommend that researchers ask themselves the following four questions about each indicator: Is the indicator a manifestation of the underlying construct (rather than a defining characteristic of it)? Is a given indicator conceptually interchangeable with the other indicators of the same construct? Will the indicators of the construct necessarily covary? And does each indicator have the same antecedents and consequences as the other indicators of the same construct? If the answer to these questions is yes, the measurement model is reflective; if the answer is no, it is formative. Although formative indicators should not be misspecified as being reflective (Rhemtulla et al. 2020), it is difficult to recommend formative measurement models for general use because they give rise to many difficult problems (see the recent discussion in Baumgartner and Weijters 2019). Since formative measures can sometimes be reformulated to make them reflective, it might be preferable to ensure that the measures used are truly reflective, rather than specifying a formative measurement model. Alternatively, the presumed formative indicators can be specified as (possibly errorful) determinants of the overall (formative) construct, although the construct has to be directly measured by reflective indicators in this case. In the sequel, we will focus on reflective measurement models, but we will briefly return to the difference between reflective and formative measurement models when discussing how to assess the quality of construct measurement.

Formally, a structural equation model can be specified as follows:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad (1)$$

$$\mathbf{y} = \boldsymbol{\Lambda}^y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (2)$$

$$\mathbf{x} = \boldsymbol{\Lambda}^x \boldsymbol{\xi} + \boldsymbol{\delta} \quad (3)$$

Equation (1) is the latent variable model in which the vector $\boldsymbol{\eta}$ (eta) contains the endogenous constructs (i.e., constructs that are functions of other constructs in the proposed model) and the vector $\boldsymbol{\xi}$ (ksi) contains the exogenous constructs (i.e., constructs that are not affected by other constructs in the model). The matrix \mathbf{B} (Beta) contains the coefficients representing the effects of endogenous on other endogenous constructs (i.e., β_{ij} is the effect of η_j on η_i , and the diagonals of this matrix are zero since a variable cannot influence itself); the matrix $\boldsymbol{\Gamma}$ (Gamma) contains the coefficients representing the effects of exogenous on endogenous constructs (i.e., γ_{ij} is the effect of ξ_j on η_i). The vector $\boldsymbol{\zeta}$ (zeta) contains the errors in equations (structural disturbances) associated with each endogenous construct. If \mathbf{B} is subdiagonal (i.e., all the coefficients above and on the diagonal are zero) and the error terms in $\boldsymbol{\zeta}$ are pairwise uncorrelated, the model is called recursive. In a recursive model, there are no reciprocal effects between the endogenous constructs and no feedback loops from one endogenous construct to itself, and the errors in equations are uncorrelated (i.e., there are no unobserved variables causing the endogenous variables to be correlated). Most structural models encountered in practice are recursive models, although they are often not realistic representations of reality.

Equations (2) and (3) are the measurement models (confirmatory factor models) for the endogenous ($\boldsymbol{\eta}$) and exogenous ($\boldsymbol{\xi}$) constructs, respectively; \mathbf{y} is a vector containing the (mean-centered) measures of the endogenous constructs, and \mathbf{x} is a vector containing the (mean-centered) measures of the exogenous constructs. The coefficients expressing the effects of the endogenous and exogenous constructs on their observed measures (called factor loadings) are contained in the matrices $\boldsymbol{\Lambda}^y$ and $\boldsymbol{\Lambda}^x$ (Lambda-y and Lambda-x), respectively. The vectors $\boldsymbol{\varepsilon}$ (epsilon) and $\boldsymbol{\delta}$ (delta) contain the unique factors (measurement errors) corresponding to the observed measures. The variance-covariance matrices of $\boldsymbol{\xi}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\delta}$ are called $\boldsymbol{\Phi}$ (Phi), $\boldsymbol{\Theta}^\varepsilon$ (Theta-epsilon) and $\boldsymbol{\Theta}^\delta$ (Theta-delta), respectively. We will not discuss the model assumptions in detail, but it is important that $\boldsymbol{\zeta}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\delta}$ are uncorrelated with $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$.

If only a single measure is available for each construct (or multiple measures are averaged to form a single composite) and measurement error in observed variables is ignored, Eqs. (2) and (3) are not needed and the analysis is based on Eq. (1); this is the conventional econometric simultaneous equation model. If a researcher is only interested in conducting a measurement analysis, Eq. (3) is sufficient.

Many different measurement models can be specified, depending on whether an observed variable is allowed to load on multiple constructs, whether method factors

are considered in addition to the substantive factors (as in multi-trait multi-method analyses), and whether measurement errors (unique factors) are specified to be correlated or uncorrelated (so-called correlated uniquenesses). Usually (at least in the first step), each observed measure is hypothesized to load on a single latent variable (the so-called target factor, which is thought to represent the substantive construct of interest), and while the variances of the unique factors are allowed to differ across observed measures, all unique factors are specified to be pairwise uncorrelated. This is called a congeneric factor model.

The model specification in Eqs. (1) to (3) is very general, but if the model parameters are to be unique, it is necessary to impose some restrictions on the model (i.e., a researcher has to make sure that the model is identified). Identification rules for structural equation models in general do not exist, but some guidelines can be offered. First, a necessary condition for a model to be identified is that the number of free model parameter not be greater than the number of distinct elements in the variance-covariance matrix of the observed variables. If this is the case, the degrees of freedom of the model will be nonnegative. Second, it is useful to break down model identification into two parts. In the first step, ignore the specific structural specification expressed by Eq. (1) and consider a (congeneric) measurement model for all the constructs and observed measures, in which the constructs are freely correlated. If there are no directed relationships between the constructs, all constructs can be treated as exogenous constructs and the measurement model can be specified using Eq. (3). Since all the variables on the right-hand side of Eq. (3) are latent, the scale in which each construct in ξ is measured has to be specified; this can be done by setting either the loading of one observed measure per construct or the variance of each construct in ξ to one. If there are at least two indicators per construct, a (congeneric) measurement model with at least two correlated factors is identified. If a construct is measured by a single indicator, the variance of the unique factor corresponding to this indicator has to be set to zero (or another assumed value). If there are at least three indicators per construct, the constructs do not have to be correlated, and when a construct is measured by at least four indicators, even a single-factor model is overidentified (i.e., the model has a positive number of degrees of freedom, which implies that the fit of the model to data can be tested).

In the second step, once the measurement model has been shown to be identified, the identification status of the structural specification of interest should be checked. Recursive models (see the earlier discussion) are known to be identified, but demonstrating identification for more complex models (e.g., by using the rank rule) is more difficult. Frequently, researchers rely on empirical identification strategies, which basically means that they trust that the computer program used for estimation will issue a warning when a model is not identified.

Figure 1 is a graphical depiction of the SEM model that will be used later in the chapter to illustrate the process of structural equation modeling. The model represents the core constructs in the so-called Technology Acceptance Model (TAM) (Davis 1989) and consists of two endogenous latent variables (or *etas*), perceived usefulness (PU) and behavioral intention to use the new technology (BI), and one

exogenous construct (or ksi), perceived ease of use (PEOU). By convention, latent variables of substantive interest are shown as ellipses (or circles). Directed arrows show causal effects, so the model assumes that PU is caused by PEOU (the strength of this relationship is expressed by γ_{11}) and BI is caused by PEOU and PU (the strength of these relationships is expressed by γ_{21} and β_{21} , respectively). PEOU is not expected to account for all the variation in PU, and PEOU and PU are not expected to account for all the variation in BI, so errors in equations (structural disturbances) are associated with each endogenous variable; the variances of these errors (zetas) are called psis and are shown as double-headed arrows. Since the structural errors are not connected with two-headed arrows (which can refer to either variances or covariances), it means that they are specified to be uncorrelated. This is a highly restrictive (and unrealistic) assumption, since it implies that there are no other unobserved variables that may cause PU and BI to be correlated. Unfortunately, a model with correlated structural errors is not identified in the present case since the latent variable model is saturated (see below), so the assumption cannot be relaxed. Because there are no feedback loops (as in PEOU \rightarrow PU \rightarrow BI \rightarrow PEOU) or reciprocal relationships (e.g., PU \rightleftharpoons BI), and since the errors in equations are uncorrelated (i.e., there is no double-headed arrow between ζ_1 and ζ_2), the model is recursive.

PEOU and PU are each measured by four indicators (PEOU1-PEOU4 and PU1-PU4), and BI is measured by two indicators (BI1-BI2). By convention, observed measures are shown as rectangles or squares. The strength of the relationships between the latent variables and their indicators is given by the lambdas (λ_{ij}), which are the factor loadings. Associated with each observed variable is a unique factor (or error of measurement), either epsilon or delta, and the variances of the unique factors are called thetas (again indicated by double-headed arrows). All unique factors are pairwise uncorrelated.

The graphical model specification shown in Fig. 1 is equivalent to the algebraic model formulation shown in Table 1. There are two latent model equations corresponding to the two endogenous latent variables (PU and BI) and 10 measurement equations corresponding to the 10 observed variables.

Since the latent variable model is saturated, the model in Fig. 1 is equivalent to a confirmatory factor model in which PEOU, PU, and BI are freely correlated. Each

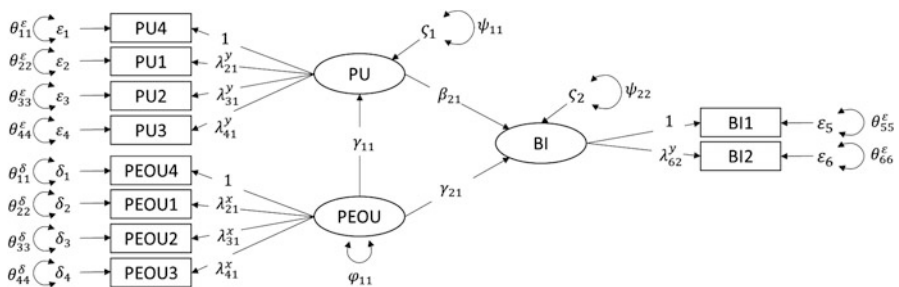


Fig. 1 Illustrative structural equation model

Table 1 Algebraic formulation of the model in Fig. 1

Latent variable model:
$PU = \gamma_{11}PEOU + \zeta_1$
$BI = \beta_{21}PU + \gamma_{21}PEOU + \zeta_2$
with $VAR(PEOU) = \varphi_{11}$, $VAR(\zeta_1) = \psi_{11}$, $VAR(\zeta_2) = \psi_{22}$, and $COV(\zeta_1, \zeta_2) = 0$.
Measurement model:
$PEOU4 = PEOU + \delta_1$
$PEOU1 = \lambda_{21}^x PEOU + \delta_2$
$PEOU2 = \lambda_{31}^x PEOU + \delta_3$
$PEOU3 = \lambda_{41}^x PEOU + \delta_4$
$PU4 = PU + \varepsilon_1$
$PU1 = \lambda_{21}^y PU + \varepsilon_2$
$PU2 = \lambda_{31}^y PU + \varepsilon_3$
$PU3 = \lambda_{41}^y PU + \varepsilon_4$
$BII = BI + \varepsilon_5$
$BI2 = \lambda_{62}^y BI + \varepsilon_6$
with $VAR(\delta_i) = \theta_{ii}^x$, $VAR(\varepsilon_i) = \theta_{ii}^y$, and all δ_i , ε_i , and ζ_i mutually uncorrelated. All observed and latent variables are assumed to be mean-centered, so intercepts are not included in the model specification.

observed variable loads on a single factor and the unique factors are uncorrelated, so the confirmatory factor (or measurement) model is congeneric. For identification, one loading per factor (e.g., λ_{11}^x , λ_{11}^y , and λ_{52}^y) has to be constrained to one, or the factor variances have to be set to unity. In a confirmatory factor model, it is best to set the factor variances to one because this yields significance tests and readily interpretable factor loadings for all indicators, and the factor covariances are actually correlations; in a full structural equation model, the variances of PU and BI are functions of other model parameters, so it is better to set one loading per factor to one. Since all constructs are measured with at least two indicators and the measurement model is congeneric, the confirmatory factor model is identified as long as the correlation of BI with the other constructs is nonzero. The latent variable model is also identified because it is a recursive model (i.e., there are no feedback loops or reciprocal effects between PEOU, PU, and BI, and the covariance between the structural disturbances ζ_1 and ζ_2 is zero). The variance-covariance matrix of the 10 observed variables consists of 55 distinct elements ($10 \times 11/2 = 55$, i.e., 10 variances and 45 covariances), and since the model in Fig. 1 (or Table 1) contains 23 free parameters (7 loadings, 10 unique factor variances, the variance of PEOU, three structural coefficients, and 2 structural disturbance variances), the model is overidentified with 32 degrees of freedom.

Model Estimation

The goal of estimation is to find values for all model parameters such that the variance-covariance matrix implied by the estimated parameters is as close as possible to the sample variance-covariance matrix. Structural equation models are

usually estimated using the maximum likelihood (ML) method under the assumption that the observations are sampled independently from an underlying multivariate normal distribution. Before estimating the model, a researcher should ascertain that the assumption of multivariate normality is not too badly violated by graphically examining the data (e.g., checking the symmetry of the variable distributions via histograms, doing normal probability plots, etc.) and computing statistics such as skewness and kurtosis (both univariate and multivariate), and possibly conducting formal tests of normality. In order for normality to hold, the variables have to be continuous, which is rarely the case, so the hope is that the results based on ML will be robust to (moderate) violations of underlying assumptions. Although estimation procedures are available that do not require normality (e.g., so-called asymptotically distribution-free procedures), they have been shown not to perform well unless the sample size is very large. There are also estimation procedures that lessen the influence of unusual observations (outliers), but these are not used very often. Structural equation models can also be estimated using partial least squares (PLS) estimation, but we will not discuss PLS estimation in this chapter because it is based on a different statistical model; the reader is referred to the chapter ▶ [“Partial Least Squares Structural Equation Modeling”](#) by Sarstedt, Ringle, and Hair (this volume).

In theory (i.e., when the underlying assumptions are satisfied), ML estimation is attractive because the ML estimator is consistent, asymptotically unbiased, asymptotically efficient, and asymptotically normally distributed. In practice, the parameter estimates themselves tend to be reasonably robust (close to the true parameters, at least in simulations), but this robustness does not hold for the overall test of model fit and the estimates of the standard errors of the parameters, which are needed for the statistical tests conducted on individual parameters or sets of parameters. Possible solutions to this problem are discussed below in the context of model testing.

Since the desirable properties of the ML estimator only hold asymptotically, the question arises how large the sample size should be so that one can have confidence in the estimates and statistical tests. Few reliable guidelines are available. Existing sample size recommendations are based on rules of thumb such as the sample size should be at least 200, there should be at least 5–10 observations per estimated parameter, or at least 10 cases should be available per observed variable (see Muthén and Muthén 2002 and Wolf et al. 2013 for references). Since there is no magic cutoff above which the desirable properties of ML suddenly kick in, the usefulness of these rules of thumb may be questioned. Furthermore, research has shown that the required sample size depends on a host of factors, including the number of observed and latent variables, the distribution of the observed variables, the reliability of the measures, the strength of the relationships between the latent variables, the type of model (CFA models vs. models with a latent variable specification), and the amount of missing data (Muthén and Muthén 2002; Wolf et al. 2013). Finally, the required sample size depends on the magnitude of the bias in parameter and standard error estimates that the researcher is willing to tolerate, the desired accuracy of the coverage rate of confidence intervals (how often, say, a 95% confidence interval includes the true parameter value), and the power for detecting specific effects or sets of effects that the researcher wants to achieve.

Instead of relying on questionable rules of thumb, a researcher can use a Monte Carlo analysis to determine the required power. To provide a specific example, consider the model in Fig. 1. Assume that each indicator has a reliability of 0.64 (i.e., with standardized observed variables, all loadings are 0.80), and the standardized path coefficients between PEOU and PU, PEOU and BI, and PU and BI are all 0.5. Thus, the amounts of variance explained in PU and BI are 25% and 75%, respectively. To conduct the analysis, a large number of samples of a given size is generated from the assumed population, for each data set the model is estimated, and the performance of the estimator at a given sample size is assessed. Specifically, Muthén and Muthén (2002) and Wolf et al. (2013) propose the following two-step procedure for determining sample size. First, the bias of both the parameter and standard error estimates should not exceed a certain percentage (e.g., the ratio of the value estimated for a given sample minus the population value over the population value should be within, say, 5% of the true value on average). In addition, the 95% confidence interval should include the true value in, say, at least 90% of the replications. Second, the estimated power for the parameter(s) of interest should be, say, at least 80% (i.e., the estimated parameter should be significant in at least 80% of samples). For the illustration, we used Mplus to conduct an analysis for 10,000 replications (the results are available on Github). Even at a sample size of only 100, most parameters had little bias (with the exception of the effects of PEOU and PU on BI, as well as the indirect effect). The same was true for standard error bias. The minimum coverage accuracy was 0.929, and power was at least 95%. At a sample size of 200, bias, coverage, and power were at acceptable levels for all parameters.

Two problems may arise during the estimation of the model. First, the estimation procedure may fail to converge within a given number of iterations or within a given time limit. For example, in the sample size simulation, the nonconvergence rate was 0.33% at a sample size of 100. Second, even if the estimation does converge, the solution may not be admissible. For example, estimated correlations may be greater than one in absolute magnitude or variance estimates may be negative. The causes of both problems include poorly specified models, few observed indicators per factor, small sample sizes, and bad starting values. It is possible to constrain questionable parameters so that inadmissible estimates are impossible, but this may lead to nonconvergence. Another possibility is to fix inadmissible estimates to a certain value (e.g., +1 or -1 for correlations that exceed 1 in absolute value or 0 for negative variance estimates); this may be defensible when the confidence interval about the estimated parameter includes the boundary of the parameter space (+1 or -1 for correlations, 0 for negative variance estimates).

Testing the Global Fit of Models

When a model is overidentified (i.e., when the degrees of freedom are positive), the fit of the model to data can be tested with a chi-square goodness-of-fit test. The null hypothesis for the test is that the model fits perfectly, whereas the alternative

hypothesis is that the fit is less than perfect (i.e., the model fits less well than the saturated model with zero degrees of freedom). The test statistic (which we will call T_{ML} in the case of maximum likelihood estimation based on the multivariate normal distribution) is compared to the critical value of a reference distribution (e.g., a chi-square distribution with df degrees of freedom), and if T_{ML} exceeds the critical value, the null hypothesis of perfect fit is rejected. The test assumes that (a) the assumptions on which the chosen estimation procedure is based are satisfied (e.g., in the case of maximum likelihood, it is assumed that the observations are independently and identically distributed and are sampled from a multivariate normal distribution) and (b) the sample size is large (because the test is only asymptotically valid). In practice, the chi-square test may not be useful because (a) it is not robust to violations of the underlying assumptions; (b) it requires a large sample size and the available sample may not be large enough to yield a trustworthy test; and, maybe most importantly, (c) the null hypothesis assumes that the specified model fits perfectly in the population, which is likely not a realistic assumption since most models are at best only approximately true. If the model is only an approximation, a large enough sample will invariably lead to the decision that the model does not fit perfectly. Sample size is thus a two-edged sword: on the one hand, a large sample is required for the chi-square test to be valid, but on the other hand, when the specified model is not literally true in the population, a large sample will lead to the rejection of the specified model. If the reasons for the departure from perfect fit could be reliably detected and sensibly corrected, this would not be a problem, but unfortunately, there are often small discrepancies in many different parts of the model that defy easy rectification.

Modified test statistics for assessing overall model fit have been suggested to deal with both a violation of normality and a small sample size. When multivariate normality is violated, the Satorra-Bentler rescaled (mean-adjusted) test statistic (T_{MLM}) is commonly used (Satorra and Bentler 2001). Unfortunately, research has shown that correct models are rejected too frequently by this test statistic when the number of variables is large and/or the sample size is small (i.e., the type I error rate is too high). Other modifications of T_{ML} to correct for non-normality (e.g., mean- and variance-adjusted test statistics) have been proposed as well.

To deal with small samples, Bartlett suggested multiplying the minimum of the ML discrepancy function not by $(N-1)$ but by a different factor that depends on the number of observed variables and the number of factors. This modification was initially introduced for exploratory factor models, but it has been applied to confirmatory factor models and more general structural equation models as well. Furthermore, the correction has also been applied to T_{MLM} . Modified versions of the Bartlett correction have been proposed as well.

Yang, Jiang and Yuan (2018) studied the performance (in terms of type I error rates) of 11 modifications of the conventional T_{ML} statistic (including T_{MLM} and the Bartlett correction) for a correctly specified SEM model across different numbers of observed variables (15–80), different sample sizes (70–2500), and different population distributions (normal, elliptical, skewed, and rescaled skewed). The degrees of freedom for the models studied ranged from 76 to 3066, so model complexity was

much higher than in many models encountered in practice. Yang et al. found that T_{MLM} showed the worst performance among all the modifications of T_{ML} and the Bartlett correction applied to T_{MLM} showed the best performance (although some other modifications performed similarly well). The performance of the test statistics that performed well depended on all three factors studied (number of observed variables, sample size, and distribution of the observed variables), although when the number of observed variables did not exceed 30, the performance of T_{MLM} with the Bartlett correction was good in general. When the number of observed variables was large and the sample size small, none of the test statistics yielded trustworthy results.

A different small-sample correction was recently proposed by McNeish (2020), who suggested that an F-distribution (rather than a chi-square distribution) be used as the reference distribution. Specifically, the chi-square statistic has to be divided by the degrees of freedom of the model (df) and the resulting ratio is then compared to an F-distribution with df and (sample size minus 1) degrees of freedom. This is based on the fact that as the denominator degrees of freedom of an F-distribution go toward infinity (i.e., as the sample size becomes very large), the F-distribution converges to a chi-square distribution divided by the numerator degrees of freedom. When the data are not normal, T_{MLM} , rather than T_{ML} , should be compared to an F-distribution.

An alternative to tests of overall model fit based on T (either T_{ML} or the various modifications discussed above) is to rely on various alternative fit indices. Generally, these do not enable inferential tests of model fit (although some, such as RMSEA, do) and instead quantify the degree of fit on a continuous scale. In order to judge fit based on these alternative fit indices, researchers need guidelines on how to interpret the scale on which fit is measured. Different researchers have proposed various cutoff values for different fit indices, based on either personal experience or simulation evidence. Initially, researchers hoped that general cutoff criteria could be developed that would be independent of model and data characteristics. For example, based on extensive simulations in which they varied sample size, distributional characteristics (normal distribution and different types of non-normality), estimation method, and type of misspecification (of either the factor loadings or the factor correlations), Hu and Bentler (1998, 1999) made recommendations about (a) which estimation procedure researchers should use (ML is preferable to generalized least squares or GLS and asymptotically distribution-free or ADF methods); (b) which fit indices researchers can rely on (among the preferred fit indices are the standardized root mean residual [SRMR], the root mean squared error of approximation [RMSEA], the confirmatory fit index [CFI], and the Tucker-Lewis index [TLI], although some caution is required for RMSEA and TLI when the sample size is 250 or smaller); and (c) what cutoffs researchers should employ to evaluate fit. In particular, they suggested that researchers use a two-index presentation strategy in which SRMR is combined with one of the other recommended fit indices, primarily because SRMR was particularly effective in detecting structural model misspecification (misspecification of the factor correlations), while the other fit indices were more effective in detecting measurement model (loading) misspecification. In

their words, “our results suggest a cutoff value close to .95 for the ML-based TLI . . . [and] CFI . . . ; a cutoff value close to .08 for SRMR; and a cutoff value close to .06 for RMSEA, before one can conclude that there is a relatively good fit between the hypothesized model and the observed data” (Hu and Bentler 1998, p. 449). Note that low values of RMSEA and SRMR and high values of CFI and TLI are indicative of good fit.

Unfortunately, subsequent research has shown that the distribution of the various fit indices depends on many different model and data characteristics, including the number of indicators per factor, the number of factors, the magnitude of the factor loadings, and the sample size (in addition to degree of misspecification, degree of normality, etc.). This implies that these model and data characteristics have to be taken into account when formulating guidelines for model fit. This has led to the development of so-called flexible cutoffs (Niemand and Mai 2018). In addition, researchers must take into account to what extent they are concerned about type I (rejecting a correct model) or type II (failing to reject an incorrect model) errors. A website (www.flexiblecutoffs.org) is available that readers can use to derive cutoffs for specific models of interest.

To summarize the discussion, it is unfortunate that, in spite of the voluminous research on the topic, generally applicable guidelines for overall fit assessment (“golden rules”) remain elusive (Marsh et al. 2004). It also seems doubtful that flexible cutoffs will prove to be a fully satisfactory solution to the fit conundrum. Furthermore, since even a well-fitting model based on the most stringent standards (e.g., the standard chi-square goodness-of-fit test) is not necessarily the “true” model, an acceptable overall model fit should never be used as the sole (or primary) arbiter of whether a proposed theory is correct (especially since usually most of the overidentifying restrictions come from the measurement model, not the latent variable model). This is especially true when alternative fit indices are used to adjudicate model fit because the cutoffs for the alternative fit indices are essentially arbitrary.

Our recommendation is twofold. First, lack of fit based on the chi-square goodness-of-fit test (or more robust alternatives) should instigate a search for major and correctable misspecifications. Second, if a respecified model still does not meet the stringent requirements of the chi-square test (after readily correctable misspecifications have been implemented), alternative fit indices (probably in combination with flexible cutoff values) may be used to justify the conclusion that the hypothesized (or respecified) model is good enough to evaluate the local fit of the model and interpret the parameters of interest. The guidelines offered by Hu and Bentler (1998) may be used as rough rules of thumb to judge model fit (i.e., if $RMSEA < 0.06$, $SRMR < 0.08$, $CFI > 0.95$, and $TLI > 0.95$, the model fits reasonably well), but they should not be employed dogmatically and supplemented with flexible cutoff values. Finally, instead of evaluating a single model, it is usually preferable to compare several plausible competing models and to determine which model is most consistent with the data. Information criteria such as the Bayesian Information Criterion (BIC) can be helpful when comparing alternative models (Bollen et al. 2014), especially when non-nested models are to be compared.

Respecifying Models That Do Not Pass the Global Fit Test

Based on our own experience with estimating numerous structural equation models, it is rare that the originally hypothesized model will provide an acceptable fit to the data. If all the fit indices indicate that the fit is poor, the model should not be interpreted before appropriate modifications are introduced. Frequently, the situation is complicated by the fact that some fit indices indicate that the fit is acceptable, while others suggest that the fit is questionable. It is not unusual to read papers in which the authors' primary concern seems to be to defend their favored model by selectively focusing on the fit indices that imply acceptable fit. Obviously, a better approach is for researchers to be skeptical of their own models and to thoroughly investigate potential sources of misfit before moving on to a substantive interpretation of the results.

The two primary tools for model modification are residual analysis and inspection of the modification indices and expected parameter changes. Residuals are the differences between the observed covariances or correlations and the covariances or correlations implied by the estimated model. A positive residual indicates that an observed covariance or correlation is underfitted, whereas a negative residual signals overfitting. In our experience, modification indices are easier to use and provide more useful information, so we will focus on them. A modification index (MI), also called Lagrange multiplier or LM statistic in some programs, is the expected decrease in the chi-square statistic (e.g., T_{ML}) when a previously fixed parameter is freely estimated or an equality constraint is relaxed. If a MI exceeds the critical value of a chi-square distribution with one degree of freedom (e.g., 3.84 for a significance level of 0.05), relaxing the constraint in question will significantly improve the fit of the model. Associated with each MI is an expected parameter change (EPC) statistic, which is the predicted estimate when a parameter is freely estimated in the revised model. When the model is reasonably complex and the sample size relatively large, many MI's can be significant, and it may not be straightforward to decide how to modify the model. Parameters should be freed one at a time in a stepwise fashion, and model modifications should be strongly guided and tempered by conceptual considerations. Sometimes, potential revisions to the model suggested by highly significant MI's make no conceptual sense, and simulations have shown that data-driven specification searches frequently fail to identify known misspecifications (e.g., MacCallum 1986; MacCallum et al. 1992). It is also important to check the EPC's associated with significant MI's to ascertain whether a suggested model modification is practically relevant and substantively interpretable. For example, the MI for a nontarget loading or an error covariance may be highly significant, but if the EPC is negligible (e.g., a standardized nontarget loading of 0.05), it is probably not meaningful to add an additional parameter to the model. Finally, researchers should compare the substantively important parameter estimates in the original model (or the model in which the major misspecifications, if any, were corrected) with those in the final, modified model, ideally one that is judged acceptable on all (or most) fit criteria. If there are no substantively important

differences, the simpler model might be preferable even when it fits the data less well than the more complex model, especially if some of the parameter estimates in the more complex model are difficult to interpret or explain.

The most common model modifications are the following. In the measurement model, observed variables often have nonzero loadings on factors other than the target factor on which each observed variable is supposed to load. It can happen that these nontarget loadings are actually stronger than the target loading, in which case the measurement model has to be revised or the offending indicator has to be dropped from the model (although this may cause other problems, such as a restriction of the domain of content of the construct). Another problem in the measurement model could be that some correlations between the unique factors (error terms) associated with different observed variables are nonzero, or that additional factors are needed to fully account for the correlations between the observed variables. For example, if some indicators are coded such that a higher score indicates a higher standing on the construct of interest (so-called regular items), whereas for other items a lower score indicates a higher standing on the underlying construct (so-called reversed items), the keying direction of the items may lead to correlated uniquenesses or require the inclusion of a method factor (or several method factors) to model this source of covariation (see Baumgartner and Weijters 2019).

In the latent variable model, the covariances between the exogenous variables are usually freely estimated, so there should be no misspecification in this part of the model. However, a saturated structural model in which all possible pairwise relationships between constructs are estimated is not parsimonious, and since researchers prefer simple models, usually some relationships between exogenous and endogenous constructs, or between endogenous constructs, are specified to be zero. For example, construct M may be hypothesized to fully mediate the relationship between constructs X and Y, in which case the direct path from X to Y should be zero. However, it is possible that the mediation is only partial in the data analyzed. This means that the MI for the direct path from X to Y will be significant (see also chapter ► “Mediation Analysis in Experimental Research” by Koschate-Fischer and Schulle, this volume). Such a misspecification is easily rectified. However, significant modification indices are not always informative. For example, in a panel data set in which construct X is measured at time t and construct Y is measured at $(t + 1)$, the MI for the path from Y to X might be significant, but such a relationship is of course impossible. The goal is to find a latent variable specification that is as simple as possible and as complex as required (see Anderson and Gerbing 1988 for details).

Often, some of the covariances or paths between constructs will be nonsignificant. If a relationship was hypothesized a priori, it is best to retain the nonsignificant path in the final model. If a relationship is not of explicit interest, one may prune the model by dropping the nonsignificant path, but the overall goodness-of-fit test is no longer interpretable as an a priori test. Ideally, modified models should be tested on new data to avoid that misleading conclusions are derived from data sets that happen to contain idiosyncratic associations.

Assessing the Local Fit of Models

Even if a model explains the covariances between the observed variables very well (as shown by a nonsignificant chi-square goodness-of-fit test), this does not mean that the constructs are measured validly and reliably, that the relationships between the constructs are consistent with the researcher's hypotheses, or that a significant proportion of the variance in the endogenous variables is accounted for. Answers to these questions require a more detailed assessment of the local fit of the estimated model. We recommend a two-step process in which the quality of the measurement model is evaluated first and then the latent variable model is investigated in detail (Anderson and Gerbing 1988). Unless the constructs are measured appropriately, it will be difficult to interpret the relationships between the constructs with confidence. Although a measurement analysis can be conducted for the model in which a particular structure is imposed on the relationships between the constructs, it is advantageous to start with a measurement model in which both the endogenous and exogenous variables are allowed to be freely correlated. In such a model, the latent variable model is saturated so that a structural misspecification will not distort the measurement relations. Once the measurement model is deemed acceptable (using the measurement model modification strategies described earlier), the hypothesized structural specification can be implemented, and after the latent variable model has been adapted, if necessary, the parameters of substantive interest can be interpreted.

Measurement Model

We will assume that a congeneric factor model fits the data adequately. If the model contains nontarget loadings or correlated uniquenesses (correlated errors), some of the discussion below may not be applicable (e.g., the computation of composite reliability).

The first step is to check the parameter estimates for the loadings, factor correlations, and error variances (i.e., the variances of the unique factors). There should be no improper solutions (e.g., the factor correlations should not exceed one in absolute magnitude, the error variances should be nonnegative), and the factor loadings should be positive (assuming all items are keyed such that higher scores reflect a higher standing on the construct of interest), significant, and substantial.

The second step is to compute various statistics related to reliability and convergent validity. Conceptually, reliability refers to the degree of convergence of measures that are very similar (they have perfectly correlated true scores) but may be distorted by random error; convergent validity refers to the degree of convergence between measures that are less similar (e.g., they might be based on different methods for measuring a construct) and may contain nonrandom error. It is often difficult to draw a sharp distinction between the two, and we will mostly use the term reliability to refer to both reliability and convergent validity. Three measures of reliability are commonly reported. As the name suggests, individual-item reliability

(IIR) refers to the reliability of a single indicator as a measure of the target construct; conceptually, it is the squared correlation between an indicator and the underlying construct, and it is computed as the square of the completely standardized loading (i.e., the loading from a factor model in which both the constructs and the observed measures are standardized to a variance of one). A summary measure for the average individual-item reliability of all measures of a construct is called average variance extracted (AVE; Fornell and Larcker 1981). For example, if a construct is measured by four items, the AVE would be the average of the four IIRs. The final reliability measure is called composite reliability (CR), and it refers to the squared correlation between an unweighted sum (or average) of all measures of a construct and the construct. It can be computed as follows:

$$CR_{\Sigma v_i} = \frac{(\sum \lambda_{ij})^2 \varphi_{jj}}{(\sum \lambda_{ij})^2 \varphi_{jj} + \sum \theta_{ii}}, \quad (4)$$

where the subscripts i and j refer to the i^{th} measure of construct j . If, in the previous example, the four items measuring the construct of interest were averaged (or summed), CR would be the estimated reliability of that composite. Composite reliability can be computed for both observed variables in their original metric or standardized observed variables; the results will differ (corresponding to the difference between a coefficient alpha based on the original or standardized variables), and if the observed variables are measured on very different scales, standardization is preferable. Since multiple measures are generally more reliable than single measures, CR will usually be larger than IIR or AVE.

We hesitate to provide guidelines about desirable levels of reliability, because reliability depends greatly on various item characteristics that do not necessarily reflect differences in measurement quality. For example, a series of more or less identical items administered one after the other will likely exhibit high reliability because respondents will fail to see the difference between the items and there are strong demands for consistency; in contrast, items that cover the domain of an intended construct more broadly and comprehensively may demonstrate less consistency. Of course, reliability assumes that the items are exchangeable, but in practice, convergent validity is probably the more appropriate concept (because measures should not be obviously redundant). Available recommendations also differ widely, particularly with respect to IIR. For IIR and AVE, 0.5 is frequently proposed as a lower limit of acceptability (i.e., observed measures should contain, on average, at least 50% trait variance). This criterion may not sound very stringent, but it frequently is not satisfied in practice. For CR, the same guidelines as for coefficient alpha apply (the two tend to be very similar in magnitude); thus, values below 0.6 are probably unacceptable, and values of 0.8 or higher are often deemed desirable.

The third step is to assess the discriminant validity of the constructs in one's model. The idea is that constructs should not correlate too highly, otherwise they may not be distinct. An important goal of discriminant validity assessment is to avoid construct proliferation. The primary and most defensible test of discriminant

validity is that the disattenuated correlation between each pair of constructs (i.e., the factor correlation corrected for the downward bias in observed correlations due to measure unreliability) should be significantly different from unity (i.e., constructs should not be perfectly correlated). The most straightforward way to conduct this test is to construct confidence intervals around the estimated factor correlations; if the confidence interval does not include one, discriminant validity is satisfied. The major problem with this criterion is that rather high correlations will differ from one when the test is sufficiently powerful and that constructs may not be distinct for practical purposes (i.e., statistical significance is not the same as practical significance). Ultimately, it is up to the researcher to decide whether a conceptual distinction between two highly correlated constructs is justified; the final arbiter of this decision should not be a statistical test. Many researchers also evaluate discriminant validity using a criterion originally proposed by Fornell and Larcker (1981). It is not a statistical test (although a statistical test could be conducted), but a numerical comparison of the squared (disattenuated) correlation between two constructs and the AVEs of the constructs involved in the correlation. If the squared correlation between two constructs is smaller than the AVE of both constructs, discriminant validity is said to be satisfied. Alternatively (and equivalently), the factor correlation can be compared with the square root of AVE. The idea is that a construct should share more variance with its own measures (as assessed by AVE) than with other (supposedly distinct) constructs. On the one hand, since the squared correlation between constructs is compared to AVE rather than unity, the Fornell and Larcker criterion is more stringent than the test of whether two constructs are perfectly correlated. On the other hand, since the Fornell and Larcker criterion usually involves only a numerical comparison, it is less stringent than a statistical test of whether two constructs are perfectly correlated (which takes into account the uncertainty involved in this decision).

If two constructs lack discriminant validity, the model has to be respecified. The two constructs may be combined (if a conceptual argument can be constructed supporting this integration), one construct may be dropped, or better measures for one of the constructs (or both constructs) may have to be developed.

So far we have only discussed discriminant validity at the construct level. At the item level, discriminant validity means that an item is solely (or at least primarily) related to its target construct, not to other, related constructs. In general, nontarget loadings are undesirable, and if they are too high, the indicator in question is probably not a good measure of the intended construct.

Once an appropriate measurement model is in place, the restrictions contained in the latent variable model can be implemented and the reliability statistics should be recomputed. The differences in the values of these statistics between the two specifications should be minor, but the measurement analysis should be reported for the final model.

It should be noted that the measurement analysis described for reflective measurement models is inappropriate for formative indicators models. Formative indicators need not be highly (positively) correlated, and error resides in the latent variable, not the indicators, so the conventional reliability indices are not applicable. Furthermore,

the notion of reliability is questionable with formative indicators, and convergent validity is the more relevant concept. Convergent validity of individual formative indicators can be assessed by the strength of the relationship between each formative indicator and the construct it measures, and the convergent validity of the formative indicators as a set can be expressed by the variance accounted for in the formatively measured construct by its indicators. Discriminant validity at the construct level can be assessed by testing whether the correlations between the constructs differ from one (as with reflective measurement models), but the conventional Fornell and Larcker criterion is not applicable. More detail is provided in Baumgartner and Weijters (2019) and MacKenzie et al. (2011), as well as the references given there.

Latent Variable Model

In studies in which SEM is used to test conceptual frameworks, the latent variable model will be of primary substantive interest. The research was probably motivated by the desire to investigate particular relationships between constructs, so the researcher will look at the sign and magnitude of the relevant parameter estimates and their statistical significance (or, preferably, the confidence interval around the estimated parameters). To get a sense of the explanatory power of the proposed framework, it is also useful to look at the variance accounted for in each endogenous construct. The chi-square goodness-of-fit test is sometimes used for evaluating the explanatory power of a framework, but as stated earlier, this is inappropriate because (a) usually most of the overidentifying restrictions tested by the chi-square test are derived from the measurement model and the chi-square test does not directly test the overidentifying restrictions contained in the latent variable model, and (b) the “explanatory” variables may explain little variation in the endogenous constructs even when the model fits well based on the chi-square test.

Sometimes, hypotheses to be tested involve indirect effects. For example, if it is hypothesized that M mediates the effect of X on Y , the indirect effect of $X \rightarrow M$ and $M \rightarrow Y$ is of interest. All programs used for SEM enable the estimation and testing of indirect effects. However, the tests are usually based on normal-theory approximations (similar to the Sobel test), which are inferior to other alternatives such as bootstrapping or Bayesian procedures. These should be used in preference to the normal-theory tests. MacKinnon et al. (2002) compared 14 methods to test the statistical significance of an indirect effect and they concluded that the “best balance of Type I error and statistical power . . . is the test of the joint significance of the two effects comprising the intervening variable [indirect] effect” (p. 83).

The Problem of Endogeneity

A model may have to be respecified not only when it fails to pass a global fit test, but also when local fit tests indicate that something is amiss. A key assumption for both the measurement and latent variable models is that the error term in each equation is

uncorrelated with the explanatory variables in that equation. If this assumption is violated, a so-called endogeneity problem exists (see chapter ► “[Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)” by Ebbes, Papies, and Van Heerde, this volume, for details). Common causes of endogeneity include measurement error in the explanatory variables; omitted variables that are correlated with the included explanatory variables and become part of the error term when not considered in the model; and reciprocal effects between the dependent variable and an explanatory variable. The usual way to deal with endogeneity is to use instrumental variables for the endogenous explanatory variables (Wooldridge 2016). An instrumental variable is a variable that is correlated with the endogenous explanatory variable (instrument relevance) but does not have a direct effect on the dependent variable and is uncorrelated with the error term in the equation of interest (instrument exogeneity). When a researcher anticipates that there might be an endogeneity problem, a so-called auxiliary variable (Bollen 2012) can be used as an instrument. For example, assume that a researcher is interested in the effect of schooling on earnings (Wooldridge 2016). When (unobserved) ability is not included in the regression, schooling is likely endogenous because ability is expected to be related to both schooling and earnings. A researcher might use proximity to a college/university or father’s (mother’s) education as an instrument, although one may question both choices because it is not clear that these variables are actually uncorrelated with the error term in the earnings-schooling regression. Bollen (2012, 2018) shows that observed variables included in a model can also serve as instrumental variables (so-called model-implied instrumental variables or MIIVs). We do not have the space to discuss MIIVs in detail but want to briefly present the basic idea. Consider the model in Fig. 1 and Table 1 and rewrite the measurement equations for the marker variables (i.e., the observed variables whose loading is set to 1) so that the latent variable appears on the left-hand side and the observed variable on the right-hand side (e.g., for PEOU this yields $PEOU = PEOU4 - \delta_1$). Then substitute this expression for all occurrences of PEOU. Do the same for PU4 and BI1. For the two latent variable equations in the model this yields the following two equations:

$$PU4 = \gamma_{11}PEOU4 - \gamma_{11}\delta_1 + \varepsilon_1 + \varsigma_1 = \gamma_{11}PEOU4 + u_1 \quad (5)$$

and

$$\begin{aligned} BI1 &= \beta_{21}PU4 + \gamma_{21}PEOU4 - \gamma_{21}\delta_1 - \beta_{21}\varepsilon_1 + \varepsilon_5 + \varsigma_2 \\ &= \beta_{21}PU4 + \gamma_{21}PEOU4 + u_2. \end{aligned} \quad (6)$$

Note that if we regressed PU4 on PEOU4 and BI1 on PU4 and PEOU4, the coefficient estimates would be inconsistent because the explanatory variables in both equations are correlated with the (composite) error terms u_1 and u_2 . However, we might be able to use other observed variables in the model as instruments for the endogenous explanatory variables PEOU4 and PU4. For example, an instrumental variable for PEOU4 in Eq. (5) would have to be correlated with PEOU4 but

uncorrelated with u_1 (i.e., δ_1 , ε_1 , and ζ_1). It turns out that PEOU1, PEOU2, and PEOU3 are suitable instruments for PEOU4 in Eq. (5), and PEOU1, PEOU2, PEOU3, PU1, PU2, and PU3 are suitable instruments for PEOU4 and PU4 in Eq. (6), assuming that the specified model is correct. The MIIVsem package in R (Fisher et al. 2020) can be used to identify model-implied instrumental variables and to estimate the coefficients in both the measurement and latent variable model equations using two-stage least squares (2SLS). The advantage of using 2SLS estimation rather than a system-wide procedure such as ML is that the 2SLS estimator is less dependent on the normality assumption and that misspecifications in other parts of the model are less likely to affect the estimation of the parameters in a specific model equation (Bollen 2018). Furthermore, if an equation is overidentified (i.e., there are more instruments than endogenous explanatory variables), a χ^2 test is available to test the null hypothesis that the MIIVs are uncorrelated with the equation error (the so-called Sargan test). If the null hypothesis is rejected, the assumptions embedded in the specified model must be questioned (e.g., because the MIIVs implied by the model are apparently not valid instrumental variables) and the model may have to be respecified. Further details will be provided in the empirical section.

Extensions of the Core Structural Equation Model

We do not have the space to discuss in detail the many extensions of the core structural equation model that have appeared in the literature. However, we will briefly mention various models that significantly expand the scope of SEM and point the interested reader to the relevant literature.

Measurement Model Extensions

The extension from reflective to formative indicator models has already been mentioned and we will not discuss it further. The interested reader can consult sources such as Baumgartner and Weijters (2019), Bollen (2011), Diamantopoulos (2011), Diamantopoulos et al. (2008), Diamantopoulos and Winklhofer (2001), Edwards (2011), Howell et al. (2007), Jarvis et al. (2003), Kline (2013), MacCallum and Browne (1993), MacKenzie et al. (2005), MacKenzie et al. (2011), and Wilcox et al. (2008).

The (reflective) measurement model described so far assumes that the observed variables are continuous and measured on an interval scale. This assumption is violated in nearly all applications of SEM. Extensions to ordered-categorical (discrete-ordinal) observed variables (e.g., Likert-type scales) are available and deserve more widespread use. The conventional measurement model for continuous observed variables still applies, but the assumption is that continuous observed variables are not directly observed and only discretized versions of these variables are available. Although the statistical theory underlying these models has been around for a long

time, and even though it is relatively easy these days to estimate these models in existing computer programs, they do introduce various complications, including the fact that the interpretation of the parameters of the model is less straightforward.

Fortunately, in many circumstances, methods developed for continuous data based on normal theory maximum likelihood estimation provide acceptable answers when robust corrections are applied to the test statistics and standard errors (this is necessary because categorical data are by definition non-normal). Specifically, based on an extensive simulation study, Rhemtulla et al. (2012) concluded that for scales with at least five answer categories, “reliance on continuous methodology in the presence of ordinal data will produce acceptable results” (p. 371).

One way in which researchers can improve the continuousness of their indicators is to form item parcels (i.e., sums or averages of sets of individual items within scales or subscales) and to use these item parcels as indicators. Measurement experts commonly agree that parceling should not be used in scale development and/or validation studies, or when the factor structure of a set of items is not well-understood, but when the number of items used to measure a construct is relatively large (e.g., greater than, say, 5), it may be impractical or even infeasible to specify a measurement model for the individual items. Furthermore, item parceling has several advantages, including better variable to sample size ratios, improved distributional properties and reliability of the parceled indicators, and more stable parameter estimates (Bandalos and Finney 2001). Usually, items are allocated to parcels in a (quasi-)random fashion, but there are situations in which strategic parceling is preferable (Weijters and Baumgartner, [forthcoming](#)). One complication that arises when using parceling is that, depending on how the parcels are formed (i.e., which specific items are allocated to a given parcel), the results may differ, and research has shown that the resulting parcel allocation variability can be non-negligible (Sterba 2011; Sterba and Pek 2012; Sterba and Rights 2017). It is therefore necessary to investigate parcel allocation variability, for example, by computing the average goodness-of-fit and the average parameter estimates across many different parcel allocations. The `semTools` package in R can be used for this purpose.

In the congeneric factor model, each indicator is allowed to load on a single construct and nontarget loadings are restricted to zero. This is a rather strong assumption that is frequently violated. Two extensions weaken this assumption. One is exploratory structural equation modeling (ESEM), where the usual confirmatory (congeneric) factor (measurement) model is replaced with an exploratory factor model (see Marsh et al. 2014). The other is Bayesian structural equation modeling (BSEM), where nontarget loadings are freely estimated but informative priors with a mean of zero and small variance are specified for the nontarget loadings to identify the model (see Muthén and Asparouhov 2012). Readers are referred to Baumgartner and Weijters (2019) and the original sources for more detail.

A final extension of the measurement model relates to situations in which substantive factors are not the only source of covariation between the indicators. Frequently, there are systematic, non-substantive influences on observed measures that are due to the method of measurement, which can cause dependencies between the items. Collectively, these are called method effects (MacKenzie and Podsakoff

2012; Podsakoff et al. 2003; Podsakoff et al. 2012), and the concern is that shared method variance may distort substantive relationships (i.e., common method bias). For example, some respondents may have a tendency to use certain response categories (e.g., the extremes, the positive or negative side, or the midpoint of the response scale), regardless of what they are being asked; shared characteristics of items, such as their keying direction (i.e., whether the item is a regular or reversed item), may lead to variance overlap; and common features of the measurement instrument or the context in which an instrument is administered may induce correlations between (some of) the items (see Podsakoff et al. 2003). To avoid common method bias, method factors or correlated uniquenesses can be included in the measurement model; method effects can be explicitly measured and accounted for in the measurement model or modeled implicitly via method factors or correlated uniquenesses; and method effects can be considered at the factor level or the level of individual items (see Baumgartner and Weijters 2019; Baumgartner and Weijters forthcoming; and Podsakoff et al. 2003 for details, as well as chapter ► “Crafting Survey Research: A Systematic Process for Conducting Survey Research” by Vomberg and Klarmann, this volume, for a discussion of survey research more generally).

Latent Variable Model Extensions

So far we have assumed that the relationships in the latent variable model are linear. This limits the applicability of SEM because theoretical frameworks sometimes specify nonlinear relationships between constructs. Here we will briefly discuss one type of nonlinear relationship in the latent variable model, namely, interactions between the exogenous latent variables (although quadratic effects could be considered as well). For concreteness, assume that instead of hypothesizing that PU partially mediates the effect of PEOU on BI, a researcher instead wants to test whether PEOU and PU have a multiplicative effect on BI, that is,

$$BI = \gamma_0 + \gamma_1 PU + \gamma_2 PEOU + \gamma_3 PEOU * PU + \varsigma_1 \quad (7)$$

The problem that arises in this type of model is that products of normally distributed variables do not have a normal distribution, which implies that BI and the indicators of BI are also non-normal. Starting with the work of Kenny and Judd (1984), many different approaches for modeling latent interaction effects have been considered (see Cortina et al. 2021 for a recent review). The most promising approach appears to be one suggested by Klein and Moosbrugger (2000), which has been implemented in Mplus and is also available in the nlsem package in R (Umbach et al. 2017). Basically, Klein and Moosbrugger (2000) show that the density of the observed variables can be expressed as a continuous mixture of normal densities and that this density can be approximated by a finite mixture of normal densities. The parameters can then be estimated with the EM algorithm. In contrast to other methods, this approach does not require that products of observed variables

be used as indicators of the latent interaction term, and it has performed well in simulation studies (maybe partly because it minimizes non-normality since products of observed variables are not used as indicators).

Models That Incorporate Population Heterogeneity

The single-sample structural equation model assumes that the observations are sampled from a single homogeneous population. One way the core model can be extended to multiple populations is to assume that there are G populations, and even though the same measurement and latent variable model applies to each of the G populations, the values of the model parameters may differ across populations. The model can be written as follows:

$$\eta^g = \alpha^g + B^g \eta^g + \Gamma^g \xi^g + \zeta^g \quad (8)$$

$$y^g = \tau^{yg} + A^{yg} \eta^g + \varepsilon^g \quad (9)$$

$$x^g = \tau^{xg} + A^{xg} \xi^g + \delta^g \quad (10)$$

where the superscript g refers to the g^{th} population ($g = 1, \dots, G$). This is the multi-sample analogue of the model in Eqs. (1)–(3), except that the model includes a latent variable model intercept term α^g (alpha) and measurement intercept terms τ^{yg} and τ^{xg} (tau). In single-sample models, the latent and observed variables are assumed to be mean-centered, but in multi-sample models, it is possible to specify a mean structure, which expresses the means of the observed variables as a function of the latent means of ξ^g (denoted by κ^g) and which requires the inclusion of intercepts in the three equations. For identification, the measurement intercept of the indicator whose loading on the target construct is set to one is restricted to zero (although other identification constraints are possible).

There are two primary uses for this model. One is to assess the invariance of parameters across groups. This is particularly important for multi-sample measurement models, because comparing construct means and relationships between constructs across groups is only meaningful if the measurements are comparable across groups. The details are spelled out in Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000), as well as other sources, but stated briefly: (a) if relationships between constructs are to be compared across groups, at least two loadings per construct have to be invariant and (b) if latent means are to be compared across groups, at least two loadings and two intercepts per construct have to be invariant. Chi-square difference tests can be used to check whether the loadings and intercepts are invariant. For example, for invariance of the loadings, the model in which all loadings are freely estimated is compared with the model in which the loadings are constrained to be equal across groups. If the first model fits significantly better than the second model, the hypothesis of invariance of all loadings has to be rejected, and modification indices can be used to free the loadings that are not invariant. When the sample sizes are rather large, it may be more meaningful to

base model comparisons on alternative fit indices; information criteria such as BIC can be especially useful for this purpose.

The second use of multi-sample models, already hinted at in the previous paragraph, is to test substantive hypotheses about differences in latent means and structural relationships between different groups. For example, a researcher may want to test whether US respondents are more individualistic (less collectivistic) than Chinese respondents, or whether attitudes are a stronger influence on behavioral intentions than social norms for US respondents, whereas the opposite is the case for Chinese respondents. Multi-sample analysis thus enables the testing of moderator effects as long as the moderator is discrete (see also chapter ► [“Challenges in Conducting International Market Research”](#) by Engelen, Engelen, and Craig, this volume, for further details about international marketing research).

In multi-sample structural equation models, the model parameters are treated as fixed effects. A second way in which population heterogeneity can be modeled is to assume that the groups for which data are available are randomly sampled from a larger number of populations and that the parameters in a particular population are specific realizations of a parameter distribution with a certain mean and variance (see chapter ► [“Multilevel Modeling”](#) by Haumann, Kassemeyer, and Wieseke, this volume). Such hierarchical (or multilevel) random effect models, in which the individual observations are nested within higher-level groups, are usually used when the number of groups is relatively large, because they only require the estimation of the means and (co)variances of the parameters and are thus more parsimonious than fixed-effect models, in which separate parameters have to be estimated for all groups (see Muthén and Asparouhov 2011 for details).

Several special cases of hierarchical models deserve mention. When repeated observations for the same units are available over time, where the number of units is relatively large and the number of time periods is limited (e.g., respondents' materialism is measured on several occasions, or sales data are recorded across a large cross-section of brands for several years), a latent (growth) curve model can be specified. In this case, the repeated observations over time are nested within some higher-level unit (e.g., respondents, brands). Latent curve models simultaneously model both the aggregate change trajectory across all units and individual differences in this average trajectory across entities. The factors representing the individual curve parameters can also be related to other variables of interest so that it becomes possible to investigate hypotheses about systematic influences on individual change processes and to specify individual differences in change as antecedents of other constructs. For example, latent curve models provide answers to the following types of research questions: What is the average trajectory of materialism over time for a sample of respondents (where different functional forms can be specified for this average trajectory)? How much individual variation is there about this average trajectory? Do the trajectories of different variables covary (e.g., does increasing loneliness over time lead to an increase in materialism)? Which unit-level covariates (e.g., gender, social class) can explain different trajectories over time? See Bollen and Curran (2006) for further details.

Hierarchical models are also useful in a cross-sectional context when respondents provide ratings of multiple stimuli. An application that is particularly relevant for marketers is conjoint analysis (see chapter ► [“Choice-Based Conjoint Analysis”](#) by Eggers et al., this volume). In a typical conjoint study, respondents’ rate (or choose from) multiple product profiles, where the product profiles (orthogonally) vary attributes such as price, brand name, or quality. In this case, the multiple ratings per respondent are nested within respondent, and a hierarchical model produces the distribution (means, variances, and covariances) of the part-worth utilities expressing the influence of the different levels of the design attributes on respondents’ overall ratings (or choices). The individual-level part-worth utilities can also be related to various antecedents and consequences. See Weijters and Baumgartner (2019) for details.

Multi-sample analysis and hierarchical modeling represent situations in which the group membership of different observations is known a priori. This is called observed population heterogeneity. In models for unobserved population heterogeneity, the goal of the analysis is to uncover the number of populations from which the observations are sampled and to determine the likely membership of observations in each group. Such models may be valuable in areas such as segmentation analysis, although they are not used much in theory-guided research. See Muthén (2001) for details.

Empirical Illustration of Structural Equation Modeling

In this section, the concepts and procedures described above are illustrated with an empirical example. The example uses publicly available data from a study by Diop et al. (2019), who surveyed 762 Chinese respondents (i.e., drivers who held a valid driver license at the time the study was conducted) about various issues related to road guidance through Variable Message Sign (VMS) information. The R code for all analyses reported below (which also includes access to the data file directly from the PLOS ONE website) is available at <https://github.com/HansBaum129/SEM>.

Conceptual Model

A VMS system uses electronic traffic signs that can be dynamically updated to provide travelers with information about such things as road blockages, congestion, and alternative routes to get to a destination. In line with the Technology Acceptance Model or TAM (Davis 1989), a driver’s behavioral intention (BI) to use electronic message signs can be explained by the perceived usefulness (PU) and perceived ease of use (PEOU) of VMS information, with PU additionally acting as a partial mediator of the effect of PEOU on BI. Figure 1 shows a graphical representation of this conceptual model. In the paper by Diop et al. (2019), the conceptual model contains additional variables specific to the VMS context (familiarity with the road network, information quality, and attitude toward route diversion), but we will focus

on the core TAM constructs because the simpler model suffices to illustrate the process of structural equation modeling.

Measurement Model

The constructs PEOU, PU, and BI were measured with multiple items (four items each for PEOU and PU, and three items for BI) using five-point Likert scales ranging from “extremely disagree” to “extremely agree,” with “neutral” as the mid-point anchor. The individual items are reported in Table 2. Several comments can be offered about them. The PEOU items are clearly reflective measures of the underlying construct. However, the first three PU items are probably formative indicators, because avoiding congestion, arriving at the destination on time, and making better routing and departure time choices (which is a double-barreled question) are probably contributing factors to PU. In contrast, the fourth PU item is clearly a reflective measure of the underlying construct. For both PEOU and PU, the fourth indicator is an overall assessment of perceived ease of use or perceived usefulness, respectively, whereas the other indicators refer to more specific aspects of each construct (esp. for PU). Finally, although the three measures of BI may be treated as reflective indicators, the third indicator does not measure behavioral intentions to use VMS information, but intentions to recommend the VMS system. These are different constructs. All these issues may create problems for model fit and may require an alternative measurement model specification, as discussed below.

Figure 2 displays the measurement model used for the confirmatory factor analysis (CFA), which is the model assumed by Diop et al. (2019). For identification,

Table 2 Items used to measure PEOU, PU, and BI

Construct	Item	Wording
Perceived ease of use	PEOU1	Using VMS information does not require a lot of mental effort.
	PEOU2	It is easy to learn how to use VMS information.
	PEOU3	VMS information is easy to understand.
	PEOU4	Overall, I find VMS information easy to use.
Perceived usefulness	PU1	Using VMS information helps me in avoiding congestion.
	PU2	Using VMS information helps me in arriving to my destination on time.
	PU3	Using VMS information helps me make better routing and departure time choices.
	PU4	Overall, I find VMS information useful.
Behavioral intention	BI1	I would consider using VMS information as long as it is available.
	BI2	I will very likely use VMS information if it is available.
	BI3	I would recommend others to use VMS information for their trips.

Source: Diop et al. (2019)

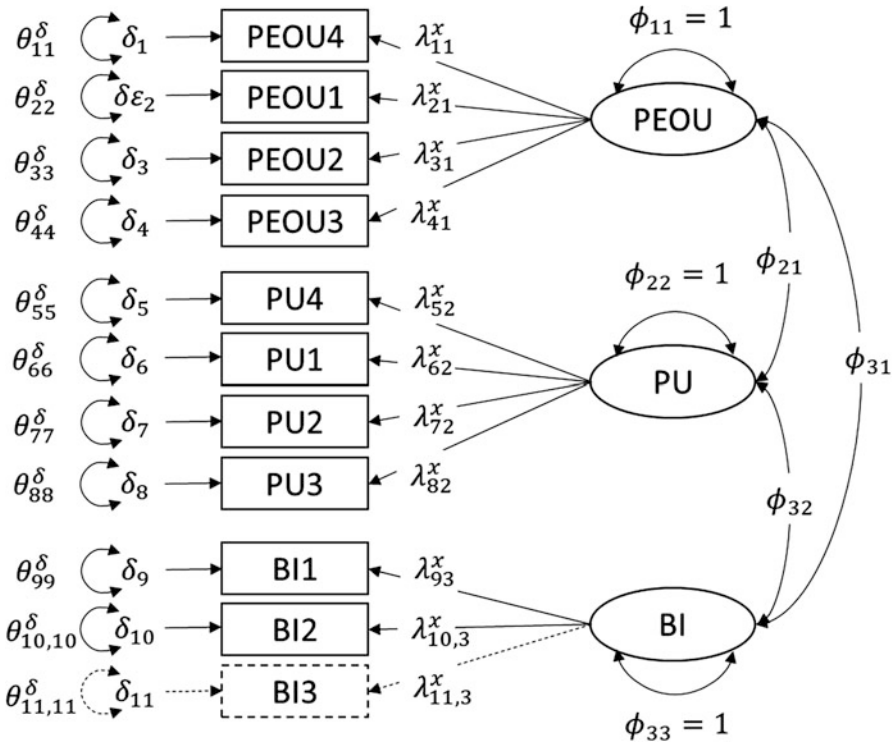


Fig. 2 Confirmatory factor analysis (measurement) model for the TAM constructs. Note: Item BI3 was eventually dropped from the model based on the measurement analysis

either one factor loading per factor needs to be fixed to one (e.g., $\lambda_{11}^x = 1, \lambda_{52}^x = 1, \lambda_{93}^x = 1$) or the factor variances have to be fixed to one ($\phi_{11} = 1, \phi_{22} = 1, \phi_{33} = 1$); as explained earlier, the latter is preferable in measurement models, as shown in Fig. 2. The null hypothesis that the model fits the data perfectly must be rejected since T_{ML} exceeds the critical value of the chi-square distribution; in particular, T_{ML} with 41 degrees of freedom is 265.25, $p < 0.0001$. The alternative fit indices are: RMSEA = 0.085 (90% CI = [0.075,0.095]); SRMR = 0.047; CFI = 0.960; and TLI = 0.947. Using conventional cutoff criteria (e.g., Hu and Bentler 1999), these indices suggest that the model shows acceptable fit (especially in terms of CFI and SRMR) or nearly acceptable fit (in terms of RMSEA and TLI) to the data. On the other hand, none of the indices meet the cutoffs suggested by www.flexiblecutoffs.org (assuming non-normality): RMSEA <0.023; SRMR <0.03; CFI > 0.982; and TLI > 0.977.

It is common in marketing research to treat five-point Likert-type item responses as continuous, interval-scaled, and normally distributed variables. Strictly speaking, this is incorrect and it may be useful to consider using an estimation and testing approach that accounts for the non-normality of the data. Using the Satorra-Bentler correction for non-normality (i.e., the MLM estimator in

lavaan) gives slightly better fit results: $T_{MLM} = 187.448$ with 41 degrees of freedom, $p < 0.001$; Robust RMSEA = 0.081 (90% CI = [0.070,0.093]), SRMR = 0.047; Robust CFI = 0.963; and Robust TLI = 0.951. However, the fit is far from perfect. Since the sample size is fairly large, the small-sample correction due to Bartlett should not have a strong effect on the result, which was indeed the case; the Bartlett correction applied to the regular chi-square statistic yielded $T_{ML}^{Bm} = 262.64$ and the Bartlett correction applied to the Satorra-Bentler chi-square statistic yielded $T_{MLM}^{Bm} = 185.85$. The small-sample correction procedure based on the F-distribution suggested by McNeish (2020) also leads to the decision that the model does not fit the data well (i.e., the p-values based on the F-distribution using either T_{ML} or T_{MLM} are essentially zero).

To diagnose the sources of misfit, one can look at the (asymptotically) standardized residuals and/or modification indices. A total of 19 residuals are significantly different from 0 at a Bonferroni-adjusted p-value of 0.0009 (the nominal alpha of 0.05 divided by the number of off-diagonal elements of 55). The four largest residuals are for PU1-PU2, PEOU2-PU1, BI1-BI2, and PEOU2-PU2. There is one very large modification index for the error covariance between PU1 and PU2 (MI = 70.88, EPC = 0.064) and three additional large modification indices for the nontarget loading of BI3 on PEOU (MI = 40.87, EPC = 0.284) and the error covariances of BI1-BI2 and PU3-PU4 (MI's of 36.93 and 33.15 and EPCs of 0.085 and 0.040, respectively).

These results suggest the following. First, the positive residual for PU1 and PU2 and the corresponding large MI show that these measures share variance that is not fully captured by the model. This is likely due to the fact that the items refer to avoiding congestion and arriving at the destination on time (which are very similar), whereas the remaining PU items are more general measures of perceived usefulness. The large residuals for PU1 and PU2 on the one hand and PEOU2 on the other hand may indicate the same problem (i.e., the closer correspondence of PU1 and PU2 relative to the other PU items). A researcher wanting to correct this problem could add a residual covariance term between PU1 and PU2, drop one of the two items (probably the item with the lower loading), or combine the two items. Alternatively, since this source of misfit is unlikely to affect the substantive findings, the problem may be ignored, which avoids overfitting and aids parsimony. After all, the items are not flagged as invalid or unreliable indicators of the construct; they merely correlate more strongly with each other than with other items measuring the same construct.

Second, the significant positive residual between BI1 and BI2 and the MI for the residual covariance between BI1 and BI2 indicate that BI1 and BI2 correlate more strongly with each other than they do with BI3. As pointed out earlier, BI3 is not a valid measure of intention to use VMS technology because it refers to intending to recommend the use of this technology to others. It may therefore be advisable to drop this item from the model, or to specify two different intention constructs, intention to use and intention to recommend the technology. The problem with the latter approach is that only a single measure of intention to recommend is available. The significant MI for the nontarget loading of BI3 on

Table 3 Composite reliability, shared variance, average variance extracted and factor correlations

	SV/AVE/CORR			
	CR	PEOU	PU	BI
PEOU	0.89	<u>0.67</u>	<i>0.47</i>	<i>0.58</i>
PU	0.90	<u>0.22</u>	0.70	<i>0.59</i>
BI	0.90	<u>0.33</u>	<u>0.35</u>	0.81

Note: *CR* Composite reliability, *SV* Shared variance (below diagonal, underlined); *AVE* Average variance extracted (diagonal, in bold), *CORR* factor correlation (above diagonal), *PEOU* Perceived ease of use, *PU* Perceived usefulness, *BI* Behavioral intention

PEOU may also hint at the fact that BI3 does not measure the same construct as BI1 and BI2.

For illustrative purposes, and because a construct that confounds intention to use and intention to recommend lacks conceptual appeal, we respecified the original model by dropping BI3 as an indicator of behavioral intention to use the VMS technology. The revised model still shows significant misfit ($T_{ML} = 201.908$ with 32 degrees of freedom, $p < 0.001$), but this should not come as a surprise since the major misspecification in the previous model (the stronger correlation between PU1 and PU2 compared to the other indicators of PU) was not corrected. The alternative fit indices show that the fit of the model has improved somewhat (RMSEA = 0.083; 90% CI = [0.073,0.095]); SRMR = 0.038; CFI = 0.967; TLI = 0.954, but particularly the RMSEA is still rather high.

In the revised CFA model, all (completely standardized) factor loadings are large (greater than 0.74) and statistically significant ($p < 0.001$). Table 3 displays the average variance extracted (AVE) and composite reliability (CR) for each of the three constructs as well as the shared variance (SV) and factor correlations between each pair of constructs. The AVE's are at least 0.67, so the indicators are quite reliable on average; the CR's are around 0.9, which indicates high internal consistency of the indicators of each construct; and the SV is well below 1 and smaller than the AVE for each pair of factors, which supports discriminant validity.

There is another measurement model specification that may be appropriate for these data. As already mentioned, the last indicator of both PEOU and PU is a global measure of each construct while the first three indicators tap into more specific aspects of ease of use and usefulness. Since all indicators are strongly related to the underlying construct, one may consider forming parceled indicators for PEOU and PU that consist of averages of the first three items. This model (see the R code for details) fits the data quite well, even though the chi-square statistic is still significant: $T_{ML} = 25.63$ with 6 degrees of freedom, $p < 0.001$; RMSEA = 0.066 (90% CI = [0.041, 0.093]); SRMR = 0.012; CFI = 0.993; and TLI = 0.983.

It is likely that many researchers would ignore the lack of fit indicated by the significant chi-square statistic in the original model (as did Diop et al.), and in some cases, this will probably not materially affect the substantive conclusions. However, the example illustrates that a detailed investigation of the sources of misfit can yield important insights into the measurement quality of different indicators, which should prove valuable in future research. The previous analysis also shows that while some

misspecifications can be corrected and justified based on conceptual considerations, others defy simple correction and ready explanation. Although a researcher should make every effort to find a model that approximates the observed covariances as well as possible, there are other considerations (e.g., the attempt to capture the full breadth of a construct) that place limits on the degree of fit that can be achieved in practice, because SEM imposes very stringent standards on model-data fit. We do not believe that researchers should restrict the domain of a construct to a single question (even if it is asked repeatedly in more or less the same way) simply to attain a good model fit.

Latent Variable Model

Since the measurement model in Fig. 2 (not including BI3 as an indicator of BI) seems reasonable, we can now consider a structural model specifying directed relationships (rather than correlations) between the constructs. As shown in Fig. 1 and Table 1, we used the fourth indicators of PEOU and PU as marker variables whose loadings were fixed to one since they are both overall measures of perceived ease of use and perceived usefulness. The structural model is saturated (i.e., there are as many path coefficients between the factors as there are factor correlations in the CFA model), so the fit of the structural equation model is identical to that of the revised CFA model. If this were not the case (i.e., if the latent variable model contained overidentifying restrictions), the fit of the structural equation model would have to be evaluated relative to the fit of the CFA model (e.g., by using a chi-square difference test). If the structural equation model were to fit the data significantly more poorly than the CFA model, the structural model would have to be revised (e.g., by relying on the modification indices for the structural paths that are fixed to zero).

Figure 3 reports standardized path coefficients (as well as estimates of indirect and total effects) with bootstrapped confidence intervals. Bootstrapped confidence intervals are preferred to assess the significance of the indirect (and total) effects, but in the present case, the confidence intervals based on MLM estimation are

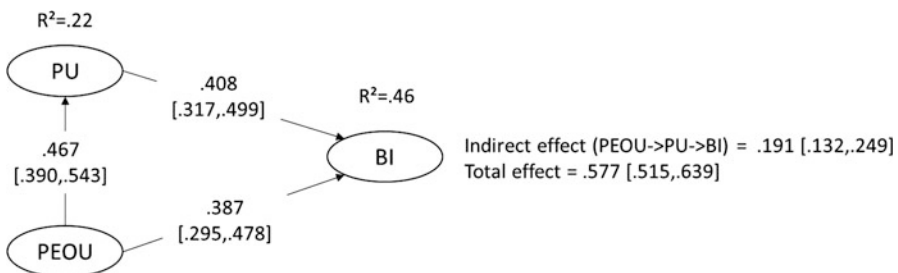


Fig. 3 Standardized estimates (with 95% confidence intervals) and R² values for the latent variable model

very similar to the bootstrapped confidence intervals for all parameters (the regular ML-based confidence intervals are somewhat too narrow). In line with the Technology Acceptance Model, BI is significantly and positively related to both PU and PEOU. In addition, PEOU has an indirect effect on BI, mediated by PU. The model explains 46% of the variance in BI and 22% of the variance in PU (the R^2 values are computed by subtracting the standardized structural residual variance of a factor from 1). One caveat should be kept in mind, however. All constructs were measured at the same time using the same questionnaire, so it is difficult to draw strong causal inferences from these results. Although it is intuitively plausible that perceived ease of use and perceived usefulness determine a driver's intention to use VMS technology, and the Technology Acceptance Model strongly supports these relationships, panel data in which PEOU and PU are measured prior to BI would provide stronger support for the predicted cause-effect relationships. The hypothesized causal effect from PEOU to PU is on particularly shaky grounds because one could certainly imagine that the two constructs simply covary or that PU affects PEOU. Moreover, the relations between the factors in the model may show upward bias due to common method variance (Baumgartner et al. 2021). Ideally, common method variance should be countered by using different methods for measuring different constructs (e.g., different types of questions, different response scales) or, if that is not possible, controlling for the presence of common method variance post hoc (e.g., by measuring potential sources of common method bias directly and including these measures as control variables; see Baumgartner and Weijters [forthcoming](#)). Finally, it should be noted that the only (substantive) sources of covariation between PEOU, PU, and BI are the direct and indirect effects of PEOU and PU on BI. It is quite unlikely that there are no other influences on the covariation between the three constructs, but since the model is saturated, it is impossible to include additional sources of covariation. Diop et al. (2019) consider other variables in their model (e.g., attitude toward route diversion is specified as an antecedent of, and thus confounding influence on the relationship between, perceived usefulness and behavior intention), but it is debatable whether the inclusion of these variables is an effective control of potential confounds.

To investigate potential endogeneity problems, we used the MIIVsem package in R to identify model-implied instrumental variables for all measurement and latent variable model equations and estimated the coefficients using 2SLS. The differences in the two types of estimates were generally small, except for the effect of PEOU on PU (the ML estimate was 0.39 whereas the 2SLS estimate was 0.45). Substantively, the results were the same. However, based on the Sargan test (even when adjusted for multiple comparisons), the null hypothesis that the model-implied instruments are uncorrelated with the equation error was rejected for every single measurement and latent variable model equation. This result is probably not too surprising since the chi-square test of model fit indicated that the model was inconsistent with the data and several large MIs suggested that some of the error correlations were highly significant. Thus, not all model-implied instruments are suitable instruments in the present case.

As an alternative to the mediation model in Fig. 1, we also estimated the interaction model in Eq. (5). The interaction was not significant, so the data provide no evidence that PEOU and PU have a multiplicative effect on BI (see the files on Github for details).

Multi-Sample Analysis

The structural model can be estimated simultaneously for multiple groups of respondents. To illustrate such an analysis, imagine that a researcher is interested in investigating the moderating effect of a driver's gender on the relations between PEOU, PU, and BI. To be able to meaningfully compare structural relationships across men and women, metric invariance has to be established first (Steenkamp and Baumgartner 1998). This requires the estimation and comparison of two two-group CFA models: a model in which the factor loadings are freely estimated in both gender groups, and a model in which corresponding factor loadings are constrained to be equal in the male versus female subsamples. The fit indices are: (a) $T_{ML} = 312.74$ with 64 degrees of freedom, RMSEA = 0.101, SRMR = 0.040, CFI = 0.953, and TLI = 0.934 for the unconstrained model, and (b) $T_{ML} = 330.66$ with 71 degrees of freedom, RMSEA = 0.098, SRMR = 0.046, CFI = 0.951, and TLI = 0.938 for the model of metric invariance. The fit of the baseline model (i.e., the unconstrained model) is marginal at best, so the results need to be interpreted with caution. The chi-square difference test comparing the unconstrained model to the metric invariance model is significant ($\Delta T_{ML}(7) = 17.927, p = 0.0123$), which implies that the model of full metric invariance fits the data significantly worse than the unconstrained model (the alternative fit indices RMSEA and TLI, which penalize less parsimonious models, actually show a slight improvement, while CFI and SRMR show a slight deterioration when metric invariance is imposed). The lack of invariance is primarily due to the loading of PEOU3 on PEOU, which has a large MI. Freeing this loading results in $T_{ML} = 316.02$ with 70 degrees of freedom, RMSEA = 0.096, SRMR = 0.041, CFI = 0.953, and TLI = 0.940. Table 4 reports the standardized structural parameter estimates for the male versus female subsamples. The results show that gender significantly moderates the effects of PEOU on PU, such that the effect is stronger for men (as compared to women). The effect of PU on BI is marginally stronger for men than women.

Concluding Comments

Structural equation modeling is used primarily in survey-based research and, particularly when applied to cross-sectional self-report data, it has encountered a fair amount of criticism because some researchers believe that it is difficult or even impossible to derive causal conclusions from structural equation models. In the early days, SEM was sometimes billed (or oversold) as causal modeling, and in complex models consisting of many exogenous and endogenous constructs, the final

Table 4 Standardized path coefficients for male versus female subsamples

Variable	IV	Males			Females			Difference		
		Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
DV	PEOU	0.568	0.036	<0.001	0.320	0.060	<0.001	0.248	0.070	0.001
PU	PU	0.471	0.045	<0.001	0.338	0.057	<0.001	0.134	0.073	0.065
BI	PEOU	0.323	0.047	<0.001	0.432	0.054	<0.001	-0.109	0.072	0.131

Note: *DV* Dependent variable, *IV* Independent variable, *PEOU* Perceived ease of use; *PU* Perceived usefulness, *BI* Behavioral intention, *Est.* Parameter estimate, *SE* Standard error

specification from which the substantive conclusions were derived often came across as ad hoc. The ambiguities associated with global goodness-of-fit tests, the problem of equivalent models (i.e., the fact that different models with very different substantive implications may fit the data equally well), the stringent assumptions imposed by multi-indicator measurement models (which have stimulated the development of very narrow measures of sometimes complex concepts), and a host of other problems have led to disillusionment about the value of SEM among (some) researchers. However, some of these issues are not unique to SEM (e.g., regression analysis faces similar problems of causality), and the ability to (a) represent the correspondence between observed measures and their presumed underlying constructs more explicitly and (b) model the relationships between constructs in a more integrative fashion are important advantages of SEM. Structural equation models can also be used in experimental contexts, in which the exogenous variables are manipulated, and particularly when the processes underlying hypothesized effects are investigated, SEM offers many benefits over regression analysis that have not been exploited by researchers. Finally, when moderators are discrete, multi-sample SEM is superior to regression-based methods, particularly when moderated mediation hypotheses are to be tested, and other approaches to modeling population heterogeneity may also be valuable.

Our discussion has focused on covariance-based SEM, but a prominent alternative (particularly in the marketing strategy and information systems literatures) is variance-based partial least squares (PLS) path modeling. Similar to Rönkko et al. (2016), we believe that PLS is mainly relevant when the emphasis is on predictive rather than explanatory modeling (see Reinartz et al. 2009). The reader is referred to Hair et al. (2017) for an introduction to PLS-SEM (see also chapter ► [“Partial Least Squares Structural Equation Modeling”](#) by Sarstedt, Ringle, and Hair, this volume).

Cross-References

- [Bayesian Models](#)
- [Challenges in Conducting International Market Research](#)
- [Choice-Based Conjoint Analysis](#)
- [Crafting Survey Research: A Systematic Process for Conducting Survey Research](#)
- [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- [Finite Mixture Models](#)
- [Mediation Analysis in Experimental Research](#)
- [Multilevel Modeling](#)
- [Panel Data Analysis: A Non-technical Introduction for Marketing Researchers](#)
- [Partial Least Squares Structural Equation Modeling](#)
- [Regression Analysis](#)

Acknowledgments Financial support from the Smeal Chair Endowment is gratefully acknowledged. The authors would like to thank two reviewers and the editors for helpful comments on a previous version of this chapter.

References

- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423.
- Bagozzi, R. P. (1980). *Causal models in marketing*. New York: Wiley.
- Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Mahwah: Erlbaum.
- Baumgartner, H., & Weijters, B. (2019). Measurement in marketing. *Foundations and Trends® in Marketing*, *12*(4), 278–400.
- Baumgartner, H., & Weijters, B. (forthcoming). Dealing with common method variance in international marketing research. *Journal of International Marketing*. in press.
- Baumgartner, H., Weijters, B., & Pieters, R. (2021). The biasing effect of common method variance: Some clarifications. *Journal of the Academy of Marketing Science*, *49*(2), 221–235. <https://doi.org/10.1007/s11747-020-00766-8>.
- Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, *35*(2), 359–372.
- Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, *38*, 37–72.
- Bollen, K. A. (2018). Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research*, *54*(1), 31–46.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken: Wiley.
- Bollen, K. A., Harden, J. J., Ray, S., & Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Structural Equation Modeling*, *21*(1), 1–19.
- Chapman, C., & Feit, E. M. D. (2019). *R for marketing research and analytics* (2nd ed.). Springer: Cham.
- Cortina, J. M., Markell-Goldstein, H. M., Green, J. P., & Chang, Y. (2021). How are we testing interactions in latent variable models? Surging forward or fighting shy? *Organizational Research Methods*, *24*(1), 26–54.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.
- Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Quarterly*, *35*(June), 335–358.
- Diamantopoulos, A., & Winklhofer, H. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(May), 269–277.
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*(12), 1203–1218.
- Diop, E. B., Zhao, S., & Duy, T. V. (2019). An extension of the technology acceptance model for understanding travelers' adoption of variable message signs. *PLoS One*, *14*(4), e0216007.
- Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14*(2), 370–388.
- Fisher, Z. F., Bollen, K. A., Gates, K., & Ronkko, M. (2020). *MIIVsem: Model implied instrumental variable (MIIV) estimation of structural equation models*. R package version 0.5.5.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Thousand Oaks: Sage.
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205–218.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30(2), 199–218.
- Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201–210.
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65(December), 457–474.
- Kline, R. B. (2013). Reverse arrow dynamics: Feedback loops and formative measurement. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 39–76). Greenwich: Information Age Publishing.
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100(1), 107–120.
- MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin*, 114, 533–541.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111(3), 490–504.
- MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: Causes, mechanisms, and procedural remedies. *Journal of Retailing*, 88(4), 542–555.
- MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*, 90(4), 710–730.
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(June), 293–334.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83–104.
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11(3), 320–341.
- Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85–110.
- McNeish, D. (2020). Should we use F-tests for model fit instead of chi-square in over-identified structural equation models? *Organizational Research Methods*, 23, 487–510.
- Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah: Erlbaum.
- Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). New York: Taylor and Francis.
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335.
- Muthén, L. K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, 9(4), 599–620.
- Niemand, T., & Mai, R. (2018). Flexible cutoff values for fit indices in the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, 46, 1148–1172.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879–903.
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539–569.

- Reinartz, W., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, 26(4), 332–344.
- Rhemtulla, M., Brousseau-Liard, A. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373.
- Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30–45.
- Rönkko, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management*, 47–48 (November), 9–27.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507–514.
- Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research*, 25(June), 78–90.
- Sterba, S. K. (2011). Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions. *Structural Equation Modeling*, 18(4), 554–577.
- Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods*, 17(4), 582–599.
- Sterba, S. K., & Rights, J. D. (2017). Effects of parceling on model selection: Parcel-allocation variability in model ranking. *Psychological Methods*, 22(1), 47–68.
- Umbach, N., Naumann, K., Brandt, H., & Kelava, A. (2017). Fitting nonlinear structural equation models in R with package nlsem. *Journal of Statistical Software*, 77(7), 1–20. <https://doi.org/10.18637/iss.v077.i07>.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(January), 4–69.
- Weijters, B., & Baumgartner, H. (2019). Analyzing Policy Capturing Data Using Structural Equation Modeling for Within-Subject Experiments (SEMWISE). *Organizational Research Methods*, 22(3), 623–648.
- Weijters, B., & Baumgartner, H. (forthcoming). On the use of balanced item parceling to counter acquiescence bias in structural equation models. *Organizational Research Methods*. in press.
- Wilcox, J. B., Howell, R. D., & Breivik, E. (2008). Questions about formative measurement. *Journal of Business Research*, 61, 1219–1228.
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6), 913–934.
- Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Boston: Cengage Learning.
- Yang, M., Jiang, G., & Yuan, K.-H. (2018). The performance of ten modified rescaled statistics as the number of variables increases. *Structural Equation Modeling*, 25(3), 414–438.



Partial Least Squares Structural Equation Modeling

Marko Sarstedt, Christian M. Ringle, and Joseph F. Hair

Contents

Introduction	588
Principles of Structural Equation Modeling	590
Path Models with Latent Variables	590
Structural Theory	591
Measurement Theory	592
Path Model Estimation with PLS-SEM	594
Background	594
The PLS-SEM Algorithm	595
Additional Considerations when Using PLS-SEM	598
Evaluation of PLS-SEM Results	601
Procedure	601
Stage 1.1: Reflective Measurement Model Assessment	603
Stage 1.2: Formative Measurement Model Assessment	606
Stage 2: Structural Model Assessment	608
Research Application	612
Corporate Reputation Model	612
Data	613

M. Sarstedt (✉)

Otto-von-Guericke University, Magdeburg, Germany

Faculty of Business and Law, University of Newcastle, Callaghan, NSW, Australia

e-mail: marko.sarstedt@ovgu.de

C. M. Ringle

Hamburg University of Technology (TUHH), Hamburg, Germany

Faculty of Business and Law, University of Newcastle, Callaghan, NSW, Australia

e-mail: c.ringle@tuhh.de

J. F. Hair

University of South Alabama, Mobile, AL, USA

e-mail: jhair@southalabama.edu

Model Estimation	615
Results Evaluation	616
Conclusions	621
Cross-References	623
References	623

Abstract

Partial least squares structural equation modeling (PLS-SEM) has become a popular method for estimating path models with latent variables and their relationships. A common goal of PLS-SEM analyses is to identify key success factors and sources of competitive advantage for important target constructs such as customer satisfaction, customer loyalty, behavioral intentions, and user behavior. Building on an introduction of the fundamentals of measurement and structural theory, this chapter explains how to specify and estimate path models using PLS-SEM. Complementing the introduction of the PLS-SEM method and the description of how to evaluate analysis results, the chapter also offers an overview of complementary analytical techniques. A PLS-SEM application of the widely recognized corporate reputation model illustrates the method.

Keywords

Partial least squares structural equation modeling · PLS-SEM · Path model analysis · Composite modeling · Results evaluation

Introduction

In the 1970s and 1980s, the Swedish econometrician Herman Wold (1975, 1982, 1985) “vigorously pursued the creation and construction of models and methods for the social sciences, where ‘soft models and soft data’ were the rule rather than the exception, and where approaches strongly oriented at prediction would be of great value” (Dijkstra 2010, p. 24). One method that emerged from Wold’s efforts was partial least squares path modeling, which later evolved to partial least squares structural equation modeling (PLS-SEM; Hair et al. 2011). PLS-SEM estimates the parameters of a set of equations in a structural equation model by combining principal component analysis with regression-based path analysis (Mateos-Aparicio 2011). Wold (1982) proposed his “soft model basic design” underlying PLS-SEM as an alternative to Jöreskog’s (1973) covariance-based SEM (chapter ► “[Structural Equation Modeling](#)”), also referred to as factor-based SEM. Covariance-based SEM has been labeled as hard modeling because of its comparably restrictive assumptions in terms of data distribution and sample size. Importantly, “it is not the concepts nor the models nor the estimation techniques which are ‘soft’, only the distributional assumptions” (Lohmöller 1989, p. 64).

A common goal of PLS-SEM analyses is to identify key success factors and sources of competitive advantage (Albers 2010; Hair et al. 2012a) for important target constructs such as customer satisfaction and customer loyalty (e.g., Fornell

et al. 1996) or behavioral intentions and user behavior (Venkatesh et al. 2003). For creating and estimating complex path models with latent variables and their relationships, PLS-SEM has achieved widespread popularity in the social sciences. Indeed, as evidenced in numerous studies that have reviewed PLS-SEM publications in a variety of disciplines, applications have increased substantially in recent years (Table 1). PLS-SEM applications have also gained prominence in other fields of scientific inquiry, such as agriculture, engineering, environmental sciences, geography, and medicine (Sarstedt 2019).

In light of the increasing maturation of the field, researchers have also started exploring the knowledge infrastructure of methodological research on PLS-SEM by analyzing the relationships between authors, countries, and co-citation networks (Hwang et al. 2020; Khan et al. 2019). As a result of these developments, a growing number of textbooks (e.g., Garson 2016; Hair et al. 2018b, 2022; Henseler 2021; Mehmetoglu and Venturini 2021; Ramayah et al. 2016; Wong 2019) and edited books on the method (e.g., Avkiran and Ringle 2018; Esposito Vinzi et al. 2010; Latan and Noonan 2017) have been published, further popularizing PLS-SEM (Ringle 2019).

A key methodological reason for PLS-SEM’s attractiveness is that the approach follows a causal-predictive paradigm, in which the aim is to test the predictive power of a model carefully developed on the grounds of theory and logic (Chin et al. 2020).

Table 1 Review articles on the use of PLS-SEM in different disciplines (Hair et al. 2022). (Reprinted by permission of the publisher (SAGE Publications))

Discipline	References
Accounting	Lee et al. (2011) Nitzl (2016)
Construction management	Zeng et al. (2021)
Entrepreneurship	Manley et al. (2020)
Family business	Sarstedt et al. (2014)
Higher education	Ghasemy et al. (2020)
Hospitality and tourism	Ali et al. (2018) do Valle and Assaker (2016) Usakli and Kucukergin (2018)
Human resource management	Ringle et al. (2020)
International business research	Richter et al. (2016)
Knowledge management	Cepeda-Carrión et al. (2019)
Management	Hair et al. (2012a)
Management information systems	Hair et al. (2017a) Ringle et al. (2012)
Marketing	Hair et al. (2012b)
Operations management	Bayonne et al. (2020) Peng and Lai (2012)
Psychology	Willaby et al. (2015)
Software engineering	Russo and Stol (2021)
Supply chain management	Kaufmann and Gaeckler (2015)

In addition, PLS-SEM enables researchers to estimate very complex models with many constructs and indicator variables, with considerably smaller sample size requirements compared to factor-based SEM methods. PLS-SEM also offers much flexibility in estimating multifaceted model relationships such as in conditional process models (Sarstedt et al. 2020a) or higher-order models (Sarstedt et al. 2019). A final reason is the accessibility of user-friendly software with a graphical interface such as ADANCO, PLS-Graph, SmartPLS, and XLSTAT, as well as the statistical computing software environment R that includes cSEM, matrixpls, SEMinR, and semPLS as complements to other programs.

The objective of this chapter is to explain the fundamentals of PLS-SEM. Building on Hair et al. (2022), this chapter first provides an introduction to the fundamentals of measurement and structural model specification as a basis for the use of the the PLS-SEM method. Next, we discuss the evaluation of results, provide an overview of complementary analytical techniques, and conclude by describing an application of the PLS-SEM method to a well-known corporate reputation model using SmartPLS 3 (Ringle et al. 2015), the most comprehensive and up-to-date software for conducting PLS-SEM analyses (Sarstedt and Cheah 2019).

Principles of Structural Equation Modeling

Path Models with Latent Variables

A path model is a diagram that displays the hypotheses and variable relationships to be estimated in a structural equation modeling analysis (Bollen 2002). Figure 1 shows an example of a path model with three latent variables (Y_1 , Y_2 , and Y_3) and their indicators.

Latent variables, also referred to as constructs, are elements in statistical models that represent conceptual variables that researchers define in their theoretical models.

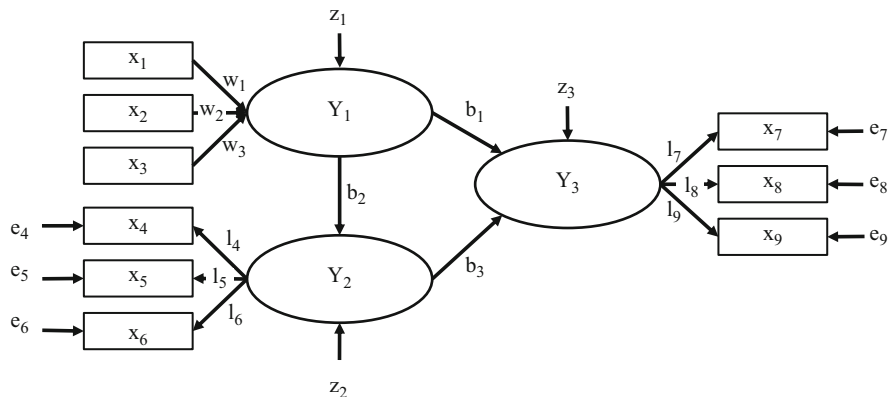


Fig. 1 Path model with latent variables

Examples of typical social sciences constructs include job satisfaction, organizational commitment, trust, and customer loyalty. Constructs are visualized as circles or ovals (Y_1 to Y_3) in path models, linked via single-headed arrows that represent causal-predictive relationships. The indicators, often also named manifest variables or items, are directly measured or observed variables that represent the raw data (e.g., respondents' answers to a questionnaire). They are represented as rectangles (x_1 to x_9) in path models and are linked to their corresponding constructs through arrows. Constructs in most instances are represented by a minimum of three or more indicators to ensure they are valid measures of the concept. Researchers sometimes include single-item constructs in their models. As construct and indicator are equivalent in this case, the relationship between construct and indicator is typically represented by a line rather than an arrow.

A path model consists of two elements. The structural model represents the causal-predictive relationships between the constructs, whereas the measurement models represent the relationships between each construct and its associated indicators. In PLS-SEM, the structural model is sometimes referred to as the inner model and the measurement models are sometimes referred to as outer models. To develop path models, researchers need to draw on both structural theory and measurement theory, which indicate the relationships between the elements of a path model.

Structural Theory

Structural theory specifies the latent variables to be considered in the analysis of a certain phenomenon and their relationships. The location and sequence of the constructs are based on theory and on the researcher's experience and accumulated knowledge (Falk and Miller 1992). When researchers develop path models, the sequence is typically from left to right. The latent variables on the left side of the path model are independent variables, and any latent variable on the right-hand side is the dependent variable (Fig. 1). However, latent variables can also serve as both independent and dependent variables in the model (Haenlein and Kaplan 2004).

When a latent variable only serves as an independent variable, it is called an exogenous latent variable (Y_1 in Fig. 1). When a latent variable only serves as a dependent variable (Y_3 in Fig. 1), or as both an independent and a dependent variable (Y_2 in Fig. 1), it is called an endogenous latent variable. Endogenous latent variables always have error terms associated with them. In Fig. 1, the endogenous latent variables Y_2 and Y_3 have one error term each (z_2 and z_3), which reflect the sources of variance not predicted by the respective antecedent construct(s) in the structural model. The exogenous latent variable Y_1 also has an error term (z_1) but in PLS-SEM, this error term is constrained to zero because of the way the method treats the (formative; i.e., arrows point from indicators to construct) measurement model of this particular construct (Diamantopoulos 2011). Therefore, this error term is typically omitted in the display of a PLS path model. In case an exogenous latent variable draws on a reflective measurement model theory (arrows point from construct to indicator), there is no error term attached to this particular construct.

The strength of the relationships between latent variables is represented by path coefficients (i.e., b_1 , b_2 , and b_3), and the coefficients are the result of regressions of each endogenous latent variable on their direct antecedent constructs. For example, b_1 and b_3 result from the regression of Y_3 on Y_1 and Y_2 .

Measurement Theory

Measurement theory specifies how to measure latent variables. Researchers can generally choose between two different types of measurement models (Diamantopoulos and Winklhofer 2001; Sarstedt et al. 2016): reflective measurement models and formative measurement models.

Reflective measurement models have direct relationships from the construct to the indicators and treat the indicators as error-prone manifestations of the underlying construct (Bollen 1989). The following equation formally illustrates the relationship between a latent variable and its observed indicators:

$$x = l \cdot Y + e, \quad (1)$$

where x is the observed indicator variable, Y is the latent variable, the loading l is a regression coefficient quantifying the strength of the relationship between x and Y , and e represents the random measurement error. This equation is a bivariate regression with x being the dependent variable and Y being the independent variable. The latent variables Y_2 and Y_3 in the path model shown in Fig. 1 have reflective measurement models with three indicators each. When using reflective indicators (also called effect indicators), the items should be a representative sample of all items of the construct's conceptual domain (Nunnally and Bernstein 1994). If the items stem from the same domain, they capture the same concept and, hence, should be highly correlated (Edwards and Bagozzi 2000).

In contrast, in a formative measurement model, a linear combination of a set of indicators forms the construct (i.e., the relationship is from the indicators to the construct). Hence, "variation in the indicators precedes variation in the latent variable" (Borsboom et al. 2003, p. 208). Indicators of formatively measured constructs do not necessarily have to correlate strongly as is the case with reflective indicators. Note, however, that strong indicator correlations can also occur in formative measurement models and do not necessarily imply that the measurement model is reflective in nature (Nitzl and Chin 2017).

When referring to formative measurement models, researchers need to distinguish two types of indicators: causal indicators and composite indicators (Bollen 2011). Constructs measured with causal indicators have an error term, which implies that the construct has not been perfectly measured by its indicators (Bollen and Bauldry 2011). More precisely, causal indicators show conceptual unity in that they correspond to the researcher's definition of the concept (Bollen and Diamantopoulos 2017). But researchers will hardly ever be able to identify all indicators relevant for adequately capturing the construct's domain (e.g., Bollen and Lennox 1991). The

error term captures all the other “causes” or explanations of the construct that the set of causal indicators do not capture (Diamantopoulos 2006). The existence of a construct’s error term in causal indicator models suggests that the construct can, in principle, be equivalent to the conceptual variable of interest, provided that the model has perfect fit (e.g., Grace and Bollen 2008). If the indicators x_1 , x_2 , and x_3 represent causal indicators, Y_j ’s error term z_j would capture these other “causes” (Fig. 1). A measurement model with causal indicators can formally be described as.

$$Y = \sum_{k=1}^K w_k \cdot x_k + z, \quad (2)$$

where w_k indicates the contribution of x_k ($k = 1, \dots, K$) to Y , and z is an error term associated with Y .

Composite indicators constitute the second type of indicators associated with formative measurement models. When measurement models are specified with composite indicators, researchers assume that the indicators define the construct in full (Sarstedt et al. 2016). Hence, the error term, which in causal indicator models represents “omitted causes,” is set to zero in formative measurement models with composite indicators ($z_j = 0$ in Fig. 1). A measurement model with composite indicators takes the following form, where Y is a linear combination of indicators x_k ($k = 1, \dots, K$), each weighted by an indicator weight w_k (Bollen 2011; McDonald 1996):

$$Y = \sum_{k=1}^K w_k \cdot x_k. \quad (3)$$

According to Henseler (2017, p. 180), measurement models with composite indicators “are a prescription of how the ingredients should be arranged to form a new entity,” which he refers to as artifacts or emergent variables (Henseler 2021). That is, composite indicators define the construct’s empirical meaning. Henseler (2017) identifies Aaker’s (1991) conceptualization of brand equity as a typical conceptual variable with composite indicators (i.e., an artifact) in advertising research, comprising the elements brand awareness, brand associations, brand quality, brand loyalty, and other proprietary assets. The use of artifacts is especially prevalent in the analysis of secondary and archival data, which typically lack a comprehensive substantiation on the grounds of measurement theory (Hair et al. 2019a; Rigdon 2013). For example, a researcher may use secondary data to form an index of a company’s communication activities, covering aspects such as online advertising, sponsoring, or product placement. Alternatively, composite indicator models can be thought of as a means to capture the essence of a conceptual variable using a limited number of indicators (Sarstedt et al. 2016). For example, a researcher may be interested in measuring the salient aspects of a company’s corporate social responsibility using a set of five (composite) indicators that capture important features relevant to the particular study.

More recent research contends that composite indicators can be used to measure any concept including attitudes, perceptions, and behavioral intentions (Nitzl and Chin 2017), as long as they operationally define the concept. But composite indicators are not a free ride for careless measurement. Instead, “as with any type of measurement conceptualization, researchers need to offer a clear construct definition and specify items that closely match this definition – that is, they must share conceptual unity” (Sarstedt et al. 2016, p. 4002). Thus, composite indicator models view construct measurement as approximation of conceptual variables, acknowledging the practical problems which arise with measuring unobservable conceptual variables that populate theoretical models (Rigdon et al. 2017, 2019).

Path Model Estimation with PLS-SEM

Background

Different from factor-based SEM (chapter ▶ “[Structural Equation Modeling](#)”), PLS-SEM explicitly calculates case values (construct scores) for the latent variables as part of the algorithm. For this purpose, the “unobservable variables are estimated as exact linear combinations of their empirical indicators” (Fornell and Bookstein 1982, p. 441) such that the resulting composites capture most of the variance of the exogenous constructs’ indicators that is useful for predicting the endogenous constructs’ indicators (e.g., McDonald 1996). PLS-SEM uses these composites to represent the constructs in a PLS path model, considering them as approximations of the conceptual variables under consideration (e.g., Hair and Sarstedt 2019; Rigdon 2012; Rigdon et al. 2017).

Since PLS-SEM-based model estimation always relies on composites, regardless of the measurement model specification, the method can process reflectively and formatively specified measurement models without identification issues (Hair et al. 2011). Identification of PLS path models only requires that each construct is linked to the nomological net of constructs (Henseler et al. 2016a). This characteristic also applies to model settings in which endogenous constructs are specified formatively as PLS-SEM relies on a multistage estimation process, which separates measurement from structural model estimation (Rigdon et al. 2014).

Three aspects are important for understanding the interplay between data, measurement, and model estimation in PLS-SEM. *First*, PLS-SEM handles all indicators of formative measurement models as composite indicators. Hence, a formatively specified construct in PLS-SEM does not have an error term as is the case with causal indicators in factor-based SEM (Diamantopoulos 2011).

Second, when the data stem from a common factor model population (i.e., the indicator covariances define the data’s nature), PLS-SEM’s parameter estimates deviate from the prespecified values. This characteristic, often incorrectly referred to as PLS-SEM bias, suggests the method overestimates the measurement model parameters and underestimates the structural model parameters (e.g., Chin et al.

2003). The degree of over- and underestimation decreases when both the number of indicators per construct and sample size increase (consistency at large; Hui and Wold 1982). The term PLS-SEM bias is a misnomer, however, as it implies that the data stem from a factor model population in which the indicator covariances define the nature of the data (e.g., Marcoulides et al. 2012; Rigdon 2016; Sarstedt et al. 2016). Numerous studies have shown that when the data stem from a composite model population where linear combinations of the indicators define the data's nature, PLS-SEM estimates are unbiased and consistent (Cho and Choi 2020; Hair et al. 2017b; Sarstedt et al. 2016). Apart from that, research has shown that the bias produced by PLS-SEM when estimating data from common factor model populations is low in absolute terms (e.g., Reinartz et al. 2009), particularly compared to the bias that common factor-based SEM produces when estimating data from composite model populations. Specifically, Sarstedt et al. (2016) find that the bias produced by factor-based SEM is, on average, 11 times higher than the bias produced by PLS-SEM when using each method on models inconsistent with what the methods assume (i.e., factor-based SEM on composite models and PLS-SEM on common factor models).

Third, PLS-SEM's use of composites not only has implications for the method's philosophy of measurement but also for its area of application. In PLS-SEM, once the weights are derived, the method always produces a single specific (i.e., determinate) score for each case per construct. This characteristic sets PLS-SEM apart from factor-based SEM, where construct scores are indeterminate, which can have considerable negative consequences for the validity of the results (Rigdon et al. 2019). Using these determinate scores as input, PLS-SEM applies a series of ordinary least squares regressions, which estimate the model parameters so they maximize the endogenous constructs' explained variance (i.e., their R^2 values). While this estimation process maximizes explanatory power, the computation of determinate construct scores makes PLS-SEM particularly well-suited for prediction where the aim is to apply model parameters estimated from a training sample to generate falsifiable predictions for other observations (hold out cases) not used in the model estimation (Hwang et al. 2020). Several studies have offered evidence of PLS-SEM's efficacy for prediction (Becker et al. 2013a; Evermann and Tate 2016; Cho et al. 2021). Hence, by using PLS-SEM, researchers simultaneously gain an understanding of the causal relationships derived from theory and logic (explanation) and also the model's predictive power, which is fundamental for establishing its practical relevance (Hair and Sarstedt 2021b; Shugan 2009).

The PLS-SEM Algorithm

Model estimation in PLS-SEM draws on a three-stage approach that belongs to the family of (alternating) least squares algorithms (Mateos-Aparicio 2011). Figure 2 illustrates the PLS-SEM algorithm as presented by Lohmöller (1989). Henseler et al. (2012) offer a graphical illustration of the SEM algorithm's stages.

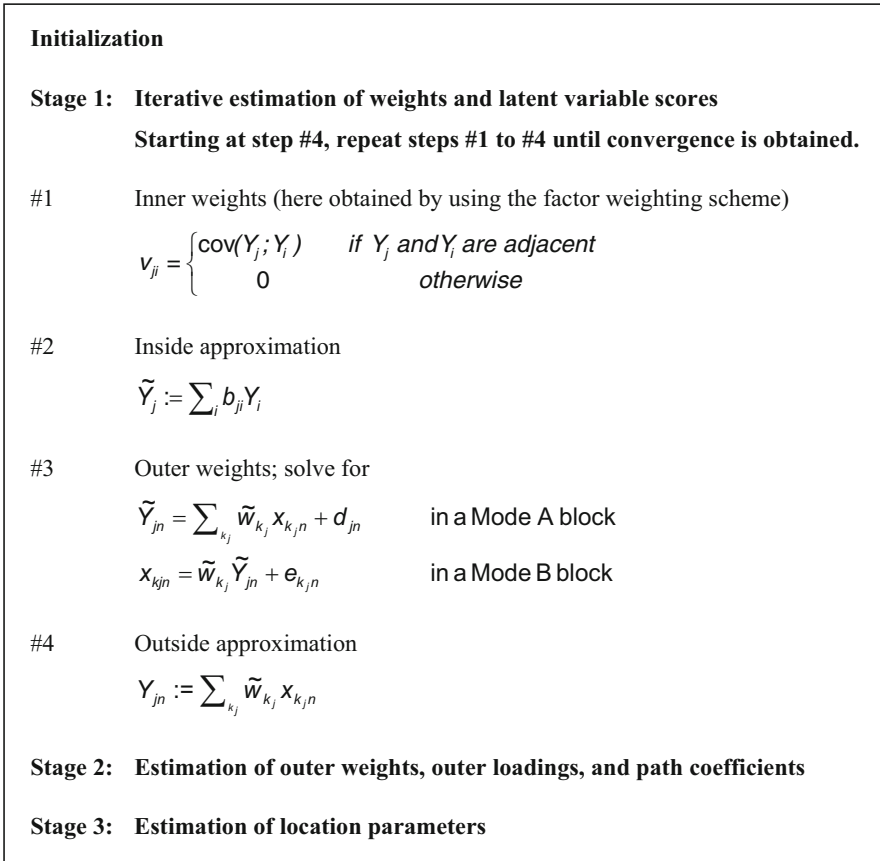


Fig. 2 The basic PLS-SEM algorithm. (Adapted from Lohmöller 1989, p. 29)

The algorithm starts with an initialization stage in which it establishes preliminary latent variable scores. To compute these scores, the algorithm typically uses unit weights (i.e., 1) for all indicators in the measurement models (Hair et al. 2022).

Stage 1 of the PLS-SEM algorithm iteratively determines the inner weights (i.e., the path coefficients) and latent variable scores by means of a four-step procedure. Step #1 uses the initial latent variable scores from the initialization of the algorithm to determine the inner weights b_{ji} between the adjacent latent variables Y_j (i.e., the dependent one) and Y_i (i.e., the independent one) in the structural model. Literature suggests three approaches to determine the inner weights (Chin 1998; Lohmöller 1989; Tenenhaus et al. 2005). In the centroid scheme, the inner weights are set to +1 if the covariance between Y_j and Y_i is positive and -1 if this covariance is negative. In case two latent variables are unconnected, the weight is set to 0. In the factor weighting scheme, the inner weight corresponds to the covariance between Y_j and Y_i and is set to zero in case the latent variables are unconnected. Finally, the path weighting scheme takes into account the direction of the inner model relationships

(Lohmöller 1989). Chin (1998, p. 309) notes that the path weighting scheme “attempts to produce a component that can both ideally be predicted (as a predictand) and at the same time be a good predictor for subsequent dependent variables.” As a result, the path weighting scheme leads to slightly higher R^2 values in the endogenous latent variables compared to the other schemes and should therefore be preferred. In most instances, however, the choice of the inner weighting scheme has very little bearing on the results (Lohmöller 1989; Noonan and Wold 1982).

Step #2, the inside approximation, computes proxies for all latent variables \tilde{Y}_j by using the weighted sum of its adjacent latent variables scores Y_i . Then, for all the indicators in the measurement models, Step #3 computes new outer weights, which indicate the strength of the relationship between each latent variable \tilde{Y}_j and its corresponding indicators. To do so, the PLS-SEM algorithm uses two different estimation modes. When using Mode A (i.e., correlation weights), the bivariate correlation between each indicator and the construct determine the outer weights. In contrast, Mode B (i.e., regression weights) computes indicator weights by regressing each construct on its associated indicators.

By default, estimation of reflectively specified constructs draws on Mode A, whereas PLS-SEM uses Mode B for formatively specified constructs. However, Cho et al. (2021) show that this reflex-like use of Mode A and Mode B is not optimal when using PLS-SEM for prediction purposes. Their simulation study shows that Mode A provides higher degrees of out-of-sample prediction in situations commonly encountered in empirical research (see also Becker et al. 2013a).

Figure 2 shows the formal representation of these two modes, where $x_{k,n}$ represents the raw data for indicator k ($k = 1, \dots, K$) of latent variable j ($j = 1, \dots, J$) and observations n ($n = 1, \dots, N$); \tilde{Y}_{jn} are the latent variable scores from the inside approximation in Step #2, \tilde{w}_{kj} are the outer weights from Step #3, d_{jn} is the error term from a bivariate regression, and $e_{k,j,n}$ is the error term from a multiple regression. The updated weights from Step #3 (i.e., \tilde{w}_{kj}) and the indicators (i.e., $x_{k,n}$) are linearly combined to update the latent variables scores (i.e., Y_{jn}) in Step #4 (outside approximation). Note that the PLS-SEM algorithm uses standardized data as input and always standardizes the generated latent variable scores in Step #2 and Step #4. After Step #4, a new iteration starts. The algorithm terminates when the weights obtained from Step #3 change marginally from one iteration to the next (typically $1 \cdot 10^{-7}$), or when the maximum number of iterations is achieved (typically 300).

Stages 2 and 3 use the final latent variable scores from Stage 1 as input for a series of ordinary least squares regressions. These regressions compute the final outer loadings, outer weights, and path coefficients as well as related elements such as indirect, and total effects, R^2 values of the endogenous latent variables, and the indicator and latent variable correlations (Lohmöller 1989).

Research has proposed several variations of the original PLS-SEM algorithm. Lohmöller's (1989) extended PLS-SEM algorithm, for example, allows assigning more than one latent variable to a block of indicators and imposing orthogonality restrictions among constructs in the structural model. Becker and Ismail (2016) developed a modified version of the original PLS-SEM algorithm that uses sampling

(post-stratification) weights to correct for sampling error. Their weighted PLS-SEM approach considers a weights vector defined by the researcher in order to ensure correspondence between sample and population structure (Cheah et al. 2020). Furthermore, Bentler and Huang's (2014) PLSe algorithm as well as Dijkstra and Henseler's (2015a, b) consistent PLS (PLSc) approach both represent modified versions of Lohmöller's (1989) original PLS-SEM algorithm that produce unbiased and consistent estimates of common factor models. That is, PLSe and PLSc both follow a composite modeling logic, but introduce a correction factor to produce results that mimic those of factor-based SEM. That is, PLSe and PLSc assume the data stem from common factor model population. But in fact, PLS-SEM does not produce biased estimates per se; the only exception is when the method is used to estimate common factor models, similar to when factor-based SEM produces biased estimates when used to estimate composite models (Sarstedt et al. 2016). In light of this concern, Hair et al. (2017a, p. 443) note: "It is unclear why researchers would use these alternative approaches to PLS-SEM when they could easily apply the much more widely recognized and validated CB-SEM [i.e., factor-based SEM] method."

Additional Considerations when Using PLS-SEM

Research has witnessed a considerable debate about situations that favor or hinder the use of PLS-SEM (e.g., Goodhue et al. 2012; Hair et al. 2019b; Marcoulides et al. 2012; Marcoulides and Saunders 2006; Henseler et al. 2014). In the following sections, we complement our previous discussion of the method's treatment of latent variables and the consequences for measurement model specification and estimation by introducing further relevant aspects to consider when using PLS-SEM, which have been discussed in the literature (e.g., Hair et al. 2013, 2019a). Where necessary, we refer to differences between factor-based SEM and PLS-SEM even though such comparisons should not be made indiscriminately (e.g., Marcoulides and Chin 2013; Rigdon 2016; Rigdon et al. 2017; Hair et al. 2017b).

Distributional Assumptions

Many researchers indicate they prefer the non-parametric PLS-SEM approach because their data's distribution does not meet the rigorous requirements of the parametric factor-based SEM approach (e.g., Hair et al. 2012b; Nitzl 2016; do Valle and Assaker 2016). However, this line of reasoning does not consider that maximum likelihood estimation in factor-based SEM is fairly robust against violations of normality (e.g., Chou et al. 1991; Olsson et al. 2000) and comes with a variety of estimators that are robust against nonnormality (Lei and Wu 2012). Thus, justifying the use of PLS-SEM solely on the grounds of data distribution is not sufficient.

Statistical Power

When using PLS-SEM, researchers benefit from the method's greater statistical power compared to factor-based SEM, even when estimating data generated from a common factor model population. Because of its greater statistical power, the

PLS-SEM method is more likely to identify an effect as significant when it is indeed present in the population.

The characteristic of higher statistical power makes PLS-SEM particularly suitable for exploratory research settings where theory is less developed. As Wold (1980, p. 70) notes, “the arrow scheme is usually tentative since the model construction is an evolutionary process. The empirical content of the model is extracted from the data, and the model is improved by interactions through the estimation procedure between the model and the data and the reactions of the researcher.”

Model Complexity and Sample Size

PLS-SEM works efficiently with small sample sizes when models are complex (e.g., Hair et al. 2017b; Sarstedt et al. 2016; Willaby et al. 2015). Prior reviews of SEM applications show that the average number of constructs per model is clearly higher in PLS-SEM (approximately eight constructs; e.g., Hair et al. 2017a; Kaufmann and Gaeckler 2015; Ringle et al. 2012) compared to factor-based SEM (approximately five constructs; e.g., Shah and Goldstein 2006; Baumgartner and Homburg 1996). Similarly, the number of indicators per construct is typically higher in PLS-SEM compared to factor-based SEM, which is not surprising considering the negative effect of more indicators on χ^2 -based fit measures in factor-based SEM. Different from factor-based SEM, the PLS-SEM algorithm does not simultaneously compute all the model relationships, but instead uses separate ordinary least squares regressions to estimate the model’s partial regression relationships – as implied by its name. As a result, the overall number of model parameters can be extremely high in relation to the sample size as long as each partial regression relationship draws on a sufficient number of observations. Reinartz et al. (2009), Henseler et al. (2014), and Sarstedt et al. (2016) show that PLS-SEM provides solutions when other methods do not converge, or develop inadmissible solutions, regardless of whether using common factor or composite model data. However, as Hair et al. (2013, p. 2) note, “some researchers abuse this advantage by relying on extremely small samples relative to the underlying population” and that “PLS-SEM has an erroneous reputation for offering special sampling capabilities that no other multivariate analysis tool has.” PLS-SEM can be applied with smaller samples in many instances when other methods fail, but the legitimacy of such analyses depends on the size and the nature of the population (e.g., in terms of its heterogeneity). No statistical method – including PLS-SEM – can offset a badly designed sample. To determine the necessary sample size, researchers should run power analyses that take into account the model structure expected effect sizes and the significance level (e.g., Marcoulides and Chin 2013) and provide power tables for a range of path model constellations. In addition, Kock and Hadaya (2018) proposed the inverse square root method, which considers the probability that the ratio of a path coefficient and its standard error will be greater than the critical value of a test statistic for a specific significance level – see Hair et al. (2022) for illustrations of the method.

While much focus has been devoted to PLS-SEM’s small sample size capabilities (e.g., Goodhue et al. 2012), discussions often overlook the method’s suitability for analyzing large datasets, such as those generated by Internet research, social media,

and social networks (e.g., Akter et al. 2017; Hair and Sarstedt 2021a). Analyses of social media data typically focus on prediction, rely on complex models with little theoretical substantiation (Stieglitz et al. 2014), and often lack a comprehensive substantiation on the grounds of measurement theory (Hair et al. 2019a; Rigdon 2013). PLS-SEM's non-parametric nature, its ability to handle complex models with many (e.g., say eight or considerably more) constructs and indicators along with its high statistical power, make it a valuable method for social media analytics and the analysis of other types of large-scale data.

Goodness-of-Fit and Prediction

PLS-SEM does not have an established goodness-of-fit measure. As a consequence, some researchers conclude that PLS-SEM's use for theory testing and confirmation is limited (e.g., Westland 2019). Recent research has, however, started reexamining goodness-of-fit measures proposed in the early days of PLS-SEM (Lohmöller 1989) or suggesting new ones, thereby broadening the method's applicability (e.g., Dijkstra and Henseler 2015a). One of the earliest proposed measures is the goodness-of-fit index (GoF), proposed by Tenenhaus et al. (2005, p. 173) as "an operational solution to this problem as it may be meant as an index for validating the PLS model globally." Henseler and Sarstedt (2013) challenged the usefulness of the GoF both conceptually and empirically, showing that the metric does not represent a goodness-of-fit criterion for PLS-SEM. Other measures include the standardized root mean square residual (SRMR), the root mean square residual covariance (RMS_{theta}), and the exact fit test (Dijkstra and Henseler 2015a; Lohmöller 1989; Henseler et al. 2014). But, while simulation studies sought to demonstrate their efficacy for PLS-SEM-based model fit testing (Schuberth et al. 2018), Hair et al. (2022) note that these measures have proven ineffective in detecting model misspecifications in settings commonly encountered in applied research.

In addition, literature casts doubt on whether measured fit – as understood in a factor-based SEM context – is a relevant concept for PLS-SEM (Hair et al. 2022; Lohmöller 1989; Rigdon 2012). Factor-based SEM follows an explanatory modeling perspective in that the algorithm estimates all the model parameters based on the objective of minimizing the divergence between the empirical covariance matrix and the model-implied covariance matrix. In contrast, the PLS-SEM algorithm follows a causal-prediction modeling perspective in that the method aims to maximize the amount of explained variance of the endogenous latent variables. Explanation and prediction are two distinct concepts of statistical modeling and estimation (e.g., Hair et al. 2019b). "In explanatory modeling the focus is on minimizing bias to obtain the most accurate representation of the underlying theory. In contrast, predictive modeling seeks to minimize the combination of bias and estimation variance, occasionally sacrificing theoretical accuracy for improved empirical precision" (Shmueli 2010, p. 293). Correspondingly, a grossly misspecified model can yield superior predictions whereas a correctly specified model can perform extremely poor in terms of prediction – see the Appendix in Shmueli (2010) for an illustration.

Researchers using PLS-SEM overcome this seeming dichotomy between explanatory and predictive modeling since they expect their model to have high predictive

Table 2 Reasons for using PLS-SEM

Reasons for using PLS-SEM
<ul style="list-style-type: none"> • The goal is to predict and explain a key target construct and/or to identify its relevant antecedent constructs. • The path model is relatively complex as evidenced in many constructs per model (six or more) and indicators per construct (more than four indicators), • The path model includes formatively measured constructs. • The sample size is limited (e.g., in business-to-business research) and also when it is large. • The research is based on secondary or archival data, which lack a comprehensive substantiation on the grounds of measurement theory. • The objective is to use latent variable scores in subsequent analyses..

accuracy, while also being grounded in well-developed causal explanations. Gregor (2006, p. 626) refers to this interplay as explanation and prediction theory, noting that this approach “implies both understanding of underlying causes and prediction, as well as description of theoretical constructs and the relationships among them.” This perspective corresponds to Jöreskog and Wold’s (1982, p. 270) understanding of PLS-SEM in which they labeled the method as a “causal-predictive” technique, meaning that when structural theory is strong, path relationships can be interpreted as causal. Hence, validation using goodness-of-fit measures is also relevant in a PLS-SEM context but less so compared to factor-based SEM. Instead, researchers should primarily focus on the assessment of their model’s predictive performance (e.g., Rigdon 2012), for example, on the grounds of Shmueli et al.’s (2016) PLS_{predict} procedure and Liengaard et al.’s (2021) cross-validated predictive ability test (CVPAT).

Table 2 summarizes the rules of thumb researchers should consider when determining whether PLS-SEM is the appropriate statistical tool for their research.

Evaluation of PLS-SEM Results

Procedure

Evaluating PLS-SEM results involves completing two stages, as illustrated in Fig. 3. Stage 1 addresses the examination of reflective measurement models (Stage 1.1), formative measurement models (Stage 1.2), or both. If the evaluation provides support for the measurement quality, the researcher continues with the structural model evaluation in Stage 2 (Hair et al. 2022). In brief, Stage 1 examines the measurement theory, while Stage 2 covers the structural theory that addresses the relationships among the latent variables, representing the proposed hypotheses.

Researchers have developed numerous guidelines for assessing PLS-SEM results (Chin 2010; Hair et al. 2019a, 2022; Roldán and Sánchez-Franco 2012), which may be summarized under the general term confirmatory composite analysis (CCA; Hair et al. 2018a, 2020). While the following illustrations draw on Hair et al. (2020), there

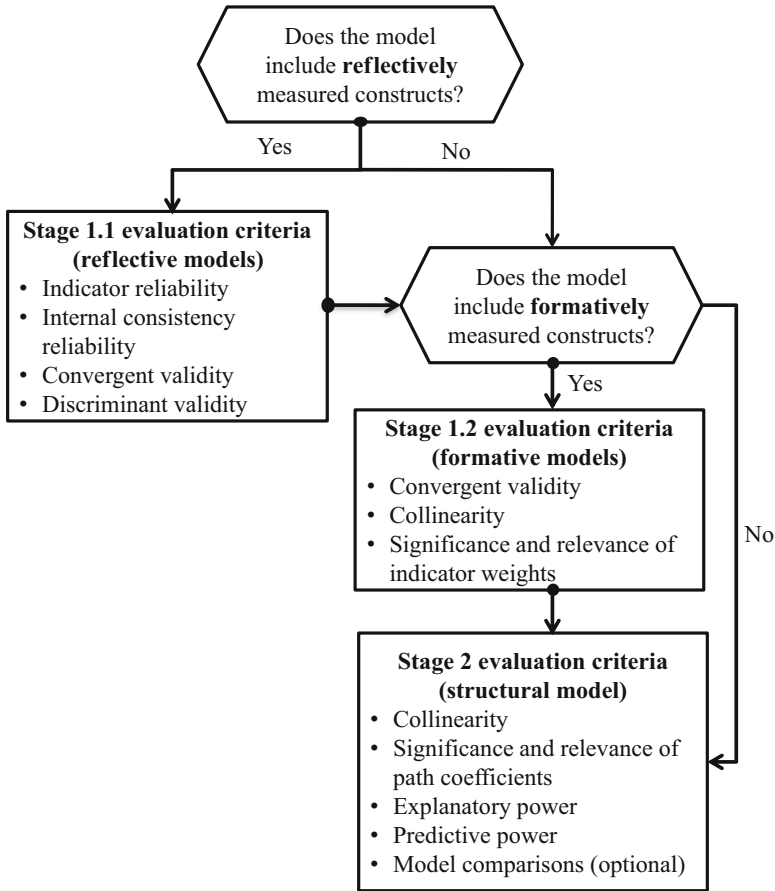


Fig. 3 PLS-SEM model evaluation. (Adapted from Sarstedt et al. 2014)

is disagreement as to which analysis steps define a confirmatory composite analysis (Henseler 2021; Henseler and Schuberth 2020; Schuberth et al. 2018). Hair et al.'s (2020) approach does not emphasize fit, but focuses on the assessment of the model's predictive power using Shmueli et al.'s (2016) PLS_{predict} procedure. In contrast, Schuberth et al.'s (2018) approach requires fit but does not refer to out-of-sample prediction – see also Henseler (2021) and Henseler and Schuberth (2020). In this chapter, we follow the CCA steps recommended by Hair et al. (2020).

Starting with the measurement model assessment and continuing with the structural model assessment, the following guidelines offer rules of thumb for interpreting the adequacy of the results. Note that a rule of thumb is a broadly applicable and easily applied guideline for decision-making that should not be strictly interpreted for every situation. Therefore, the threshold for a rule of thumb may vary depending on the research context.

Stage 1.1: Reflective Measurement Model Assessment

In the case of reflectively specified constructs, a researcher begins Stage 1 by examining the indicator loadings. Loadings above 0.708 indicate the construct explains more than 50% of the indicator's variance, demonstrating that the indicator exhibits a satisfactory degree of item reliability.

The next step involves the assessment of the constructs' internal consistency reliability. When using PLS-SEM, internal consistency reliability is generally evaluated using Jöreskog's (1971) composite reliability ρ_c , which is defined as follows (for standardized data):

$$\rho_c = \frac{\left(\sum_{k=1}^K l_k\right)^2}{\left(\sum_{k=1}^K l_k\right)^2 + \sum_{k=1}^K \text{var}(e_k)}, \quad (4)$$

where l_k symbolizes the standardized outer loading of the indicator variable k of a specific construct measured with K indicators, e_k is the measurement error of indicator variable k , and $\text{var}(e_k)$ denotes the variance of the measurement error, which is defined as $1 - l_k^2$.

For the composite reliability criterion, higher values indicate higher levels of reliability. For instance, researchers can consider values between 0.60 and 0.70 as acceptable in exploratory research, whereas results between 0.70 and 0.95 represent satisfactory to good reliability levels (Hair et al. 2022). However, values that are too high (e.g., higher than 0.95) are problematic, as they suggest that the items are almost identical and redundant. The reason may be (almost) the same item questions in a survey or undesirable response patterns such as straight lining (Diamantopoulos et al. 2012).

Cronbach's alpha is another measure of internal consistency reliability that assumes the same thresholds but yields lower values than the composite reliability (ρ_c). This statistic is defined in its standardized form as follows, where K represents the construct's number of indicators and \bar{r} the average non-redundant indicator correlation coefficient (i.e., the mean of the lower or upper triangular correlation matrix):

$$\text{Cronbach's } \alpha = \frac{K \cdot \bar{r}}{[1 + (K - 1) \cdot \bar{r}]}. \quad (5)$$

Generally, in PLS-SEM Cronbach's alpha is considered the lower bound, while ρ_c defines the upper bound of internal consistency reliability when estimating reflective measurement models with PLS-SEM. Hence, the actual reliability of a construct likely falls between Cronbach's alpha and the composite reliability ρ_c .

As an alternative and building on Dijkstra (2010), subsequent research has proposed the exact (or consistent) reliability coefficient ρ_A (Dijkstra 2014; Dijkstra and Henseler 2015b), which is defined as

$$\rho_A := (\widehat{w}'\widehat{w})^2 \cdot \frac{\widehat{w}'(S - \text{diag}(S))\widehat{w}}{\widehat{w}'(\widehat{w}\widehat{w}' - \text{diag}(\widehat{w}\widehat{w}'))\widehat{w}} \tag{6}$$

where \widehat{w} represents the indicator weights estimates, diag indicates the diagonal of the corresponding matrix, and S the sample covariance matrix. The ρ_A reliability metric usually lies between Cronbach's α and the composite reliability ρ_c , and is therefore considered a good compromise between these other two measures (Hair et al. 2019a).

The next step in assessing reflective measurement models addresses convergent validity, which is the extent to which a construct converges in its indicators by explaining the items' variance. Convergent validity is assessed by the average variance extracted (AVE) across all items associated with a particular reflectively measured construct and is also referred to as communality. The AVE is calculated as the mean of the squared loadings of each indicator associated with a construct (for standardized data):

$$\text{AVE} = \frac{\left(\sum_{k=1}^K l_k^2\right)}{K}, \tag{7}$$

where l_k and K are defined as explained above. An acceptable threshold for AVE is 0.50 or higher. This level or higher indicates that, on average, the construct explains (more than) 50% of the variance of its items.

Once the reliability and the convergent validity of reflectively measured constructs have been successfully established, the final step is to assess their discriminant validity. This analysis reveals to which extent a construct is empirically distinct from other constructs both in terms of how much it correlates with other constructs and how distinctly the indicators represent only this single construct. Discriminant validity assessment in PLS-SEM involves analyzing Henseler et al.'s (2015) heterotrait-monotrait ratio (HTMT) of correlations. The HTMT criterion is defined as the mean value of the indicator correlations across constructs relative to the (geometric) mean of the average correlations of indicators measuring the same construct. The HTMT of the constructs Y_i and Y_j with, respectively, K_i and K_j indicators is defined as follows:

$$\text{HTMT}_{ij} = \underbrace{\frac{1}{K_i K_j} \sum_{g=1}^{K_i} \sum_{h=1}^{K_j} r_{ig,jh}}_{\text{average heterotrait-heteromethod correlation}} \div \underbrace{\left(\frac{2}{K_i(K_i - 1)} \cdot \sum_{g=1}^{K_i-1} \sum_{h=g+1}^{K_i} r_{ig,ih} \cdot \frac{2}{K_j(K_j - 1)} \cdot \sum_{g=1}^{K_j-1} \sum_{h=g+1}^{K_j} r_{jg,jh} \right)^{\frac{1}{2}}}_{\text{geometric mean of the average monotrait-heteromethod correlation of construct } Y_i \text{ and the average monotrait-heteromethod correlation of construct } Y_j}, \tag{8}$$

where $r_{ig,jh}$ represents the correlations of the indicators (i.e., within and across the measurement models of latent variables Y_i and Y_j). Figure 4 shows the correlation matrix of the six indicators used in the reflective measurement models of constructs Y_2 and Y_3 from Fig. 1.

Therefore, high HTMT values indicate discriminant validity problems. Based on prior research and their simulation study results, Henseler et al. (2015) suggest a threshold value of 0.90 if the path model includes constructs that are conceptually very similar (e.g., affective satisfaction, cognitive satisfaction, and loyalty); that is, an HTMT value above 0.90 depicts a lack of discriminant validity. However, when the constructs in the path model are conceptually more distinct, researchers should consider 0.85 as threshold for HTMT (Henseler et al. 2015).

In addition, researchers can (and should) use bootstrap confidence intervals (see next section for a discussion of the bootstrapping concept) to test if the HTMT is significantly lower than 1.00 (Henseler et al. 2015) or another threshold value such as 0.90 or 0.85. The concrete threshold should be defined based on the study context (Franke and Sarstedt 2019). For example, assuming a threshold of 0.85 and assuming a significance level of 5%, researchers need to assess whether the upper boundary of the one-sided 95% bootstrap confidence interval (i.e., UB_{95}) is lower than 0.85. This upper boundary can also be inferred from a two-sided 90% bootstrap confidence interval. In order to obtain the bootstrap confidence intervals, in line with Aguirre-Urreta and Rönkkö (2018), researchers should generally use the percentile method. However, when the reliability coefficient’s bootstrap distribution is skewed, the bias-corrected and accelerated (BCa) method should be preferred to obtain bootstrap confidence intervals. The recommended number of bootstrap samples researchers should use is 10,000 (Streukens and Leroi-Werelds 2016). We discuss the different bootstrap confidence interval types and parameter settings in greater detail in the next section.

Trait		Y_2			Y_3		
Trait	Method	x_4	x_5	x_6	x_7	x_8	x_9
Y_2	x_4	1					
	x_5	$r_{4,5}$	1				
	x_6	$r_{4,6}$	$r_{5,6}$	1			
Y_3	x_7	$r_{4,7}$	$r_{5,7}$	$r_{6,7}$	1		
	x_8	$r_{4,8}$	$r_{5,8}$	$r_{6,8}$	$r_{7,8}$	1	
	x_9	$r_{4,9}$	$r_{5,9}$	$r_{6,9}$	$r_{7,9}$	$r_{8,9}$	1

Fig. 4 Correlation matrix example

Stage 1.2: Formative Measurement Model Assessment

Formatively specified constructs are evaluated differently from reflectively measured constructs. Their evaluation involves the examination of (1) the convergent validity, (2) indicator collinearity, and (3) statistical significance and relevance of the indicator weights – see Fig. 3.

In formative measurement model evaluation, convergent validity refers to the degree to which the formatively specified construct correlates with an alternative measure of the same concept. Originally proposed by Chin (1998), the procedure is referred to as redundancy analysis. To execute this procedure for determining convergent validity, researchers must plan ahead in the research design stage by including an alternative measure of the formatively measured construct in their questionnaire. Cheah et al. (2018) show that a single item, which captures the essence of the construct under consideration, is generally sufficient as an alternative measure – despite limitations with regard to criterion validity (Diamantopoulos et al. 2012). When the model is based on secondary data, an available variable measuring a similar concept would be used (Houston 2004). Hair et al. (2022) suggest the correlation of the formatively measured construct with the reflectively measured item(s) should be 0.708 or higher, which implies that the construct explains (more than) 50% of the alternative measure's variance (Carlson and Herdman 2012).

Collinearity assessment involves computing each item's variance inflation factor (VIF) by running a multiple regression of each indicator in the measurement model of the formatively measured construct on all the other items of the same construct. The R^2 values of the k -th regression facilitates the computation of the VIF for the k -th indicator, using the following formula:

$$VIF_k = \frac{1}{1 - R_k^2} \quad (9)$$

Higher R^2 values in the k -th regression imply that the variance of the k -th item can be explained by the other items in the same measurement model, which indicates collinearity issues. Likewise, the higher the VIF , the greater the level of collinearity. As a rule of thumb, VIF values above 3 are indicative of collinearity among the indicators. However, collinearity issues can also occur at lower VIF values of 3 (e.g., Mason and Perreault 1991). Hence, when the analysis produces unexpected sign changes in the indicator weights, researchers should reconsider the model set-up in an effort to reduce the collinearity.

The third step in assessing formatively measured constructs is examining the statistical significance and relevance (i.e., the size) of the indicator weights. In contrast to regression analysis, PLS-SEM does not make any distributional assumptions regarding the error terms that would facilitate the immediate testing of the weights' significance based on the normal distribution. Instead, the researcher must run bootstrapping, a procedure that draws a large number of subsamples (typically 10,000) from the original data. The model is then estimated for each of the subsamples, yielding a high number of estimates for each model parameter.

Using the subsamples from bootstrapping, the researcher can construct a distribution of the parameter under consideration and compute bootstrap standard errors, which allow for determining the statistical significance of the original indicator weights. More precisely, bootstrap standard errors allow for computing t -values (and corresponding p -values). When interpreting the results, reviewers and editors should be aware that bootstrapping is a random process, which yields different results every time it is initiated. While the results from one bootstrapping run to the next generally do not differ fundamentally when using a large number of bootstrap samples such as 10,000 (Streukens and Leroi-Werelds 2016), bootstrapping-based p -values slightly lower than a predefined cut-off level should give rise to concern. In such a case, researchers may have repeatedly applied bootstrapping until a certain parameter has become significant, a practice referred to as p -hacking.

As an alternative, researchers can use the bootstrapping results to construct different types of confidence intervals. Aguirre-Urreta and Rönkkö (2018) show that the percentile method performs very well in a PLS-SEM context in terms of coverage (i.e., the proportion of times the population value of the parameter is included in the $1-\alpha\%$ confidence interval in repeated samples) and balance (i.e., how $\alpha\%$ of cases fall to the right or to the left of the interval). If a weight's confidence interval includes zero, this provides evidence that the weight is not statistically significant, making the indicator a candidate for removal from the measurement model. However, instead of mechanically deleting the indicator, researchers should first consider its loading, which represents the indicator's absolute contribution to the construct. While an indicator might not have a strong relative contribution (e.g., because of the large number of indicators in the formative measurement model), its absolute contribution can still be substantial and meaningful (Cenfetelli and Bassellier 2009). Based on these considerations, the following rules of thumb apply (Hair et al. 2022):

- If the weight is statistically significant, the indicator is retained.
- If the weight is nonsignificant, but the indicator's loading is 0.50 or higher, the indicator is still retained if theory and expert judgment support its inclusion.
- If the weight is nonsignificant and the loading is low (i.e., below 0.50), the indicator should be deleted from the measurement model.

Researchers must be cautious when deleting formative indicators based on statistical outcomes for at least the following two reasons. First, the indicator weight is a function of the number of indicators used to measure a construct: The higher the number of indicators, the lower their average weight. In other words, formative measurement models have an inherent limit to the number of indicators that can retain a statistically significant weight (e.g., Cenfetelli and Bassellier 2009). Second, as formative indicators define the construct's empirical meaning, indicator deletion should be considered with caution and should generally be the exception. Content validity considerations are imperative before deleting formative indicators (e.g., Diamantopoulos and Winklhofer 2001).

Having assessed the formative indicator weights' statistical significance, the final step is to examine each indicator's relevance for shaping the construct. In terms of relevance, indicator weights are standardized to values that are usually between -1 and $+1$, with weights closer to $+1$ (or -1) representing strong positive (or negative) relationships, and weights closer to 0 indicating weak relationships. Note that values below -1 and above $+1$ may technically occur, for instance, when collinearity is at critical levels.

Stage 2: Structural Model Assessment

Provided the measurement model assessment indicates satisfactory quality, the researcher moves to the assessment of the structural model in Stage 2 of the PLS-SEM evaluation process (Fig. 3). After checking for potential collinearity issues among the constructs, this stage considers the significance and relevance of the structural model relationships (i.e., the path coefficients) as well as the model's explanatory and predictive power. Some research situations call for the computation and comparison of alternative models, which can emerge from different theories or contexts. PLS-SEM facilitates the comparison of alternative models using criteria that are well known from the regression literature, as well as more recent out-of-sample prediction metrics. As model comparisons are not relevant for every PLS-SEM analysis, this assessment is optional.

Computation of the path coefficients linking the constructs is based on a series of regression analyses. Therefore, the researcher must first ascertain that collinearity issues do not bias or distort the regression results. This step is analogous to the formative measurement model assessment, with the difference that the scores of the exogenous latent variables serve as input for the *VIF* assessments. *VIF* values above 3 are indicative of collinearity among sets of predictor constructs. However, as indicated in the context of formative measurement model assessment, collinearity can also occur at lower *VIF* values.

Subsequently, the strength and significance of the path coefficients is evaluated regarding the relationships (structural paths) hypothesized between the constructs. Similar to the assessment of formative indicator weights, the significance assessment builds on bootstrapping standard errors as a basis for calculating *t*-values and *p*-values of path coefficients, or – as recommended in the literature – their percentile confidence intervals (Aguirre-Urreta and Rönkkö 2018). A path coefficient is significant at the 5% probability of error level if zero does not fall into the 95% percentile confidence interval. For example, a path coefficient of 0.15 with 0.1 and 0.2 as lower and upper bounds of the 95% percentile confidence interval would be considered significant since zero does not fall into this confidence interval. On the contrary, with a lower bound of -0.05 and an upper bound of 0.35, we would consider this coefficient as not significant.

In terms of relevance, path coefficients are usually between -1 and $+1$, with coefficients closer to $+1$ representing strong positive relationships, and those closer to -1 indicating strong negative relationships (note that values below -1 and above $+1$

may technically occur, for instance, when collinearity is at critical levels). A path coefficient of say 0.5 implies that if the independent construct increases by one standard deviation unit, the dependent construct will increase by 0.5 standard deviation units when keeping all other independent constructs constant. Determining whether the size of the coefficient is meaningful should be decided within the research context. When examining the structural model results, researchers should also interpret total effects. The total effect corresponds to the sum of the direct effect and all the indirect effects between two constructs in the path model. With regard to the path model shown in Fig. 1, Y_1 has a direct effect (b_1) and an indirect effect ($b_2 \cdot b_3$) via Y_2 on the endogenous construct Y_3 . Hence, the total effect of Y_1 on Y_3 is $b_1 + b_2 \cdot b_3$. The examination of total effects between constructs, including all their indirect effects, provides a more comprehensive picture of the structural model relationships (Nitzl et al. 2016).

The next step involves reviewing the coefficient of determination (R^2). The R^2 measures the variance explained in each of the endogenous constructs and is therefore a measure of the model's explanatory power (Shmueli and Koppius 2011), also referred to as in-sample predictive power (Rigdon 2012). The R^2 ranges from 0 to 1, with higher levels indicating a higher degree of explanatory power. As a rough rule of thumb, the R^2 values of 0.75, 0.50, and 0.25 can be considered substantial, moderate, and weak (Henseler et al. 2009; Hair et al. 2011). Acceptable R^2 values are based on the context. In some disciplines an R^2 value as low as 0.10 is considered satisfactory, for example, when predicting stock returns (Raithel et al. 2012). In other contexts, scientists usually expect higher R^2 values above 0.65. An example is the customer satisfaction construct in American Customer Satisfaction Index model applications (Fornell et al. 1996; chapter ► [“Measuring Customer Satisfaction and Customer Loyalty”](#)).

More importantly, the R^2 is a function of the number of predictor constructs – the greater the number of predictor constructs, the higher the R^2 . Therefore, the R^2 should always be interpreted relative to the context of the study based on the R^2 values from related studies and models of similar complexity. R^2 values can also be too high when the model overfits the data. Model overfit is present when the partial regression model is too complex, which results in fitting the random noise inherent in the sample rather than reflecting the overall population. The same model would likely not fit as well on another sample drawn from the same population (Sharma et al. 2018). When measuring a concept that is inherently predictable, such as physical processes, R^2 values of 0.90 might be plausible. Similar R^2 value levels in a model that predicts human attitudes, perceptions and intentions likely indicate model overfit (Hair et al. 2019a).

In addition to evaluating the R^2 values of all endogenous constructs, the change in the R^2 value when a specified exogenous construct is omitted from the model can be used to evaluate whether the omitted construct has a substantive impact on the endogenous constructs. This measure is referred to as the f^2 effect size and can be calculated as

$$f^2 = \frac{R^2_{included} - R^2_{excluded}}{1 - R^2_{included}} \quad (10)$$

where R^2_{included} and R^2_{excluded} are the R^2 values of the endogenous latent variable when a selected exogenous latent variable is included in or excluded from the model. Technically, the change in the R^2 values is calculated by estimating a specific partial regression in the structural model twice (i.e., with the same latent variable scores). First, the model is estimated with all exogenous latent variables included (yielding R^2_{included}) and, second, with a selected exogenous latent variable excluded (yielding R^2_{excluded}). As a guideline, f^2 values of 0.02, 0.15, and 0.35, respectively, represent small, medium, and large effects (Cohen 1988) of an exogenous latent variable. Effect size values of less than 0.02 indicate that there is no effect.

To assess a PLS path model's predictive power, also referred to as out-of-sample predictive power, researchers can draw on Shmueli et al.'s (2016) PLS_{predict} procedure. PLS_{predict} executes k -fold cross-validation by randomly partitioning the dataset into k subsets (folds). In the following, PLS_{predict} then combines $k - 1$ subsets into a single analysis sample that is used to predict the indicator values of a specific target constructs in the remaining data subset (i.e., the holdout sample). This process is repeated k times such that each subset serves as holdout sample once. Shmueli et al. (2019) recommend setting $k = 10$, but researchers need to make sure the analysis sample for each subset (fold) meets minimum sample size guidelines. PLS_{predict} can also be run repeatedly to alleviate the impact of potentially extreme samples resulting from the random partitioning of the data into k folds. As a rule of thumb, researchers should generally run PLS_{predict} with ten repetitions.

To quantify the degree of prediction error, researchers can draw on several prediction statistics. The default statistic is the root mean squared error (RMSE), which weights large prediction errors more strongly than small errors. When the prediction error distribution is highly nonsymmetric, researchers may use the mean absolute error (MAE), which measures the average magnitude of the errors in a set of predictions without considering their direction (over or under). Both RMSE and MAE cannot be interpreted absolutely as their values depend on the measurement scale of the indicators under consideration. For example, an indicator measured on a scale from 0 to 100 can cover a much greater range of prediction errors than a 7-point Likert scale.

Hence, researchers need to compare the RMSE (or MAE) values with a linear regression model (LM) benchmark to generate predictions for the manifest variables by running a linear regression of each of the dependent construct's indicators on the indicators of the exogenous constructs in the PLS path model (Danks and Ray 2018). In comparing the RMSE (or MAE) values with the LM values, the following guidelines apply (Shmueli et al. 2019):

1. If *all* indicators in the PLS-SEM analysis have lower RMSE (or MAE) values compared to the naïve LM benchmark, the model has high predictive power.
2. If the *majority* (or the same number) of indicators in the PLS-SEM analysis yields smaller prediction errors compared to the LM, this indicates medium predictive power.
3. If the *minority* of the dependent construct's indicators produces lower PLS-SEM prediction errors compared to the naïve LM benchmark, this indicates that the model has low predictive power.

4. If the PLS-SEM analysis (compared to the LM) yields lower prediction errors in terms of the RMSE (or the MAE) for *none* of the indicators, this indicates that the model lacks predictive power.

Researchers can also assess the $Q^2_{predict}$ statistic, which indicates whether the PLS-SEM-based predictions outperform the most naïve benchmark, defined as the indicator mean from the holdout samples. A $Q^2_{predict}$ larger than zero indicates the PLS path model outperforms this most naïve benchmark. Importantly, when interpreting PLS_{predict} results, researchers should focus on the model's key endogenous construct rather than examining the prediction errors for the indicators of all endogenous constructs. Shmueli et al. (2019) present a systematic application of the PLS_{predict} procedure including the $Q^2_{predict}$ criterion and the LM benchmark.

In a final, optional step, researchers may be interested in comparing different model configurations resulting from different theories or research contexts. Sharma et al. (2018) and Danks et al. (2020) compared the efficacy of various metrics for model comparison tasks and found that Schwarz's (1978) Bayesian information criterion (BIC) and Geweke and Meese's (1981) criterion (GM) achieve a sound tradeoff between model fit and predictive power in the estimation of PLS path models. These (Information Theoretic) model selection criteria facilitate the comparison of models in terms of model fit and predictive power without having to use a holdout sample, which is particularly useful for PLS-SEM analyses that often draw on small sample sizes. In applying these metrics, researchers should estimate each model separately and select the model that minimizes the value in BIC or GM for a certain target construct. While BIC and GM exhibit practically the same performance in model selection tasks, BIC is easier to compute. Hence, focusing on this criterion is sufficient in most model comparison tasks. The BIC for a certain model i is defined as follows:

$$BIC_i = n \left[\log\left(\frac{SSE_i}{n}\right) + \frac{p_j \cdot \log(n)}{n} \right], \quad (11)$$

where SSE_i is the sum of squared errors for the i -th model in a set of alternative models, n is the sample size, and p_j is the number of predictors of the construct of interest plus 1.

One issue in the application of the BIC is that – in its simple form (i.e., raw values) – the criterion does not offer any insights regarding the relative weights of evidence in favor of models under consideration (Burnham and Anderson 2002). More precisely, while the differences in BIC values are useful in ranking and selecting models, such differences can often be small in practice, leading to model selection uncertainty. To resolve this issue, researchers can use the BIC values to compute Akaike weights, which indicate a model's relative likelihood, given the data and a set of competing models (Danks et al. 2020) – see Wagenmakers and Farrell (2004) for an application.

A further advancement in the field of prediction-oriented model comparisons in PLS-SEM is Liengaard et al.'s (2021) CVPAT, which proves valuable for developing

and validating theories from a prediction standpoint (Hair et al. 2022). Future extensions of CVPAT will allow researchers to test the predictive power of their models on a standalone basis (Hair et al. 2020).

Research Application

Corporate Reputation Model

The empirical application builds on the corporate reputation model and data that Hair et al. (2022) use in their book *Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, and that Hair et al. (2018b) also employ in their *Advanced Issues in Partial Least Squares Structural Equation Modeling* book. The PLS path model creation and estimation was executed using the SmartPLS 3 software (Ringle et al. 2015). The model files, datasets and software used in this market research application can be downloaded at <https://www.smartpls.com>.

Figure 5 shows the corporate reputation model as displayed in SmartPLS 3. Originally presented by Eberl (2010), the goal of this model is to explain the effects of corporate reputation on customer satisfaction (CUSA) and, ultimately, customer loyalty (CUSL). Corporate reputation represents a company’s overall evaluation by its stakeholder (Helm et al. 2010), which comprises two dimensions (Schwaiger

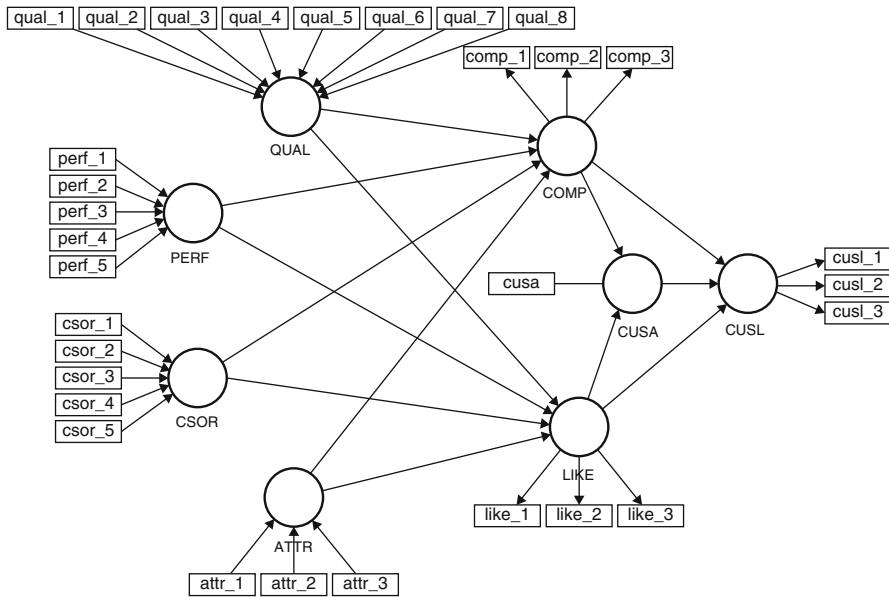


Fig. 5 Corporate reputation model in SmartPLS 3

2004). The first dimension captures cognitive evaluations of the company, and the construct is the company's competence (*COMP*). The second dimension captures affective judgments, which determine the company's likeability (*LIKE*). This two-dimensional reputation measurement has been validated in different countries and applied in various research studies (e.g., Eberl and Schwaiger 2005; Raithel and Schwaiger 2015; Schloderer et al. 2014). Research has shown that the approach performs favorably (in terms of convergent validity and predictive validity) compared with alternative reputation measures (e.g., Sarstedt et al. 2013). Schwaiger (2004) also identified four exogenous constructs that represent the key sources of the two corporate reputation dimensions: (1) the quality of a company's products and services, as well as the quality of its customer orientation (*QUAL*); (2) the company's economic and managerial performance (*PERF*); (3) the company's corporate social responsibility (*CSOR*); and (4) the company's attractiveness (*ATTR*).

In terms of construct measurement, *COMP*, *LIKE*, and *CUSL* have reflectively specified measurement models with three items. *CUSA* draws – for illustrative purposes – on a single-item measure. The four exogenous latent variables *QUAL*, *PERF*, *CSOR*, and *ATTR* have formative measurement models. Table 3 provides an overview of all items' wordings.

Data

The model estimation draws on data from four German mobile telecommunications providers. A total of 344 respondents rated the questions related to the items on a 7-point Likert scale, whereby a value of seven always represents the best possible judgment and a value of one the opposite. The most complex partial regression in the PLS path model has eight independent variables (i.e., the formative measurement model of *QUAL*). Hence, following Cohen's (1992) recommendations for multiple ordinary least squares regression analysis or running a power analysis, one would need only 54 observations to detect R^2 values of around 0.25, assuming a significance level of 5% and a statistical power of 80%. When considering the more conservative inverse square root method suggested by Kock and Hadaya (2018), the minimum sample size requirement is approximately 275, assuming a minimum path coefficient of 0.15 at a 5% probability of error level.

The dataset has only 11 missing values, which are coded with the value –99. The maximum number of missing data points per item is 4 of 334 (1.16%) in *cusl_2*. Since the relative number of missing values is very small, we continue the analysis by using the mean value replacement of missing data option. Box plots diagnostic by means of IBM SPSS Statistics (Sarstedt and Mooi 2019) reveals influential observations, but no outliers. Finally, the skewness and excess kurtosis values, as provided by the SmartPLS 3 data view, show that all the indicators are within the –2 and +2 acceptable range (George and Mallery 2019).

Table 3 Item wordings (Hair et al. 2022)

Attractiveness (ATTR) - formative	
<i>attr_1</i>	[the company] is successful in attracting high-quality employees.
<i>attr_2</i>	I could see myself working at [the company].
<i>attr_3</i>	I like the physical appearance of [the company] (company, buildings, shops, etc.).
Competence (COMP) - reflective	
<i>comp_1</i>	[the company] is a top competitor in its market.
<i>comp_2</i>	As far as I know, [the company] is recognized worldwide.
<i>comp_3</i>	I believe that [the company] performs at a premium level.
Corporate Social Responsibility (CSOR) - formative	
<i>csor_1</i>	[the company] behaves in a socially conscious way.
<i>csor_2</i>	[the company] is forthright in giving information to the public.
<i>csor_3</i>	[the company] has a fair attitude toward competitors.
<i>csor_4</i>	[the company] is concerned about the preservation of the environment.
<i>csor_5</i>	[the company] is not only concerned about profits.
Customer loyalty (CUSL) - reflective	
<i>cusl_1</i>	I would recommend [company] to friends and relatives.
<i>cusl_2</i>	If I had to choose again, I would choose [company] as my mobile phone services provider.
<i>cusl_3</i>	I will remain a customer of [company] in the future.
Customer satisfaction (CUSA) - single item	
<i>Cusa</i>	If you consider your experiences with [company], how satisfied are you with [company]?
Likeability (LIKE) – Reflective	
<i>like_1</i>	[the company] is a company that I can better identify with than other companies.
<i>like_2</i>	[the company] is a company that I would regret more not having if it no longer existed than I would other companies.
<i>like_3</i>	I regard [the company] as a likeable company.
Quality (QUAL) – Formative	
<i>qual_1</i>	The products/services offered by [the company] are of high quality.
<i>qual_2</i>	[the company] is an innovator, rather than an imitator with respect to [industry].
<i>qual_3</i>	[the company]'s products/services offer good value for money.
<i>qual_4</i>	The services [the company] offers are good.
<i>qual_5</i>	Customer concerns are held in high regard at [the company].
<i>qual_6</i>	[the company] is a reliable partner for customers.
<i>qual_7</i>	[the company] is a trustworthy company.
<i>qual_8</i>	I have a lot of respect for [the company].
Performance (PERF) - formative	
<i>perf_1</i>	[the company] is a very well-managed company.
<i>perf_2</i>	[the company] is an economically stable company.
<i>perf_3</i>	The business risk for [the company] is modest compared to its competitors.
<i>perf_4</i>	[the company] has growth potential.
<i>perf_5</i>	[the company] has a clear vision about the future of the company.

Model Estimation

The model estimation uses the basic PLS-SEM algorithm by Lohmöller (1989), the path weighting scheme, a maximum of 300 iterations, a stop criterion of 0.0000001 (or $1 \cdot 10^{-7}$), and equal indicator weights for the initialization (default settings in the SmartPLS 3 software). After running the algorithm, it is important to ascertain that the algorithm converged (i.e., the stop criterion has been reached) and did not reach the maximum number of iterations. However, with sufficiently high numbers of maximum iterations (e.g., 300 and higher), the PLS-SEM algorithm practically always converges in empirical studies, even in very complex market research applications.

Figure 6 shows the PLS-SEM results. The numbers on the path relationships represent the standardized regression coefficients while the numbers displayed in the circles of the endogenous latent variables are the R^2 values. An initial assessment shows that *CUSA* has the strongest effect (0.505) on *CUSL*, followed by *LIKE* (0.344) and *COMP* (0.006). These three constructs explain 56.2% (i.e., the R^2 value) of the variance of the endogenous construct *CUSL*. Similarly, we can interpret the relationships between the exogenous latent variables *ATTR*, *CSOR*, *PERF*, and *QUAL*, as well as the two corporate reputation dimensions *COMP* and *LIKE*. But before we address the interpretation of these results, we must assess the constructs' reflective and formative measurement models.

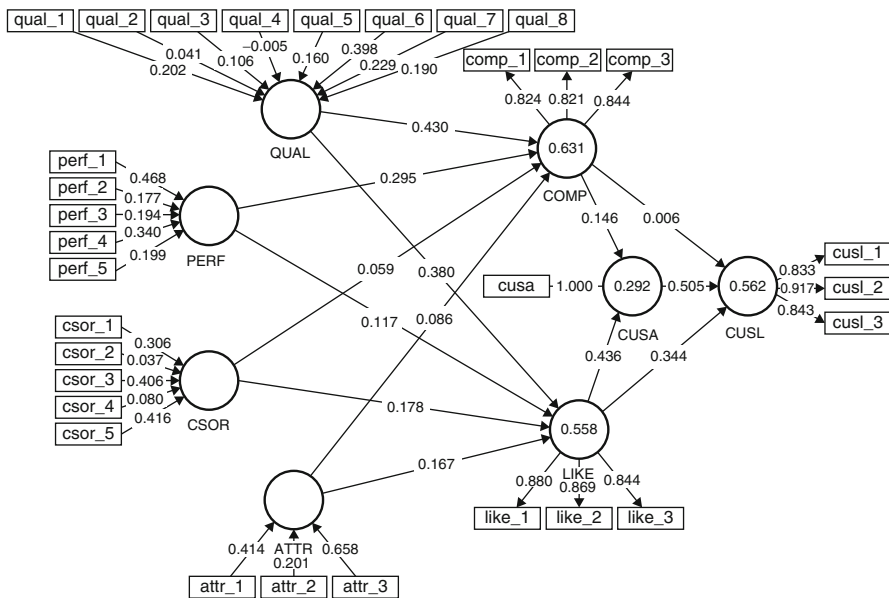


Fig. 6 Corporate reputation model and PLS-SEM results

Results Evaluation

Reflective Measurement Model Assessment

The evaluation of the PLS-SEM results begins with an assessment of the reflective measurement models (i.e., *COMP*, *CUSL*, and *LIKE*). Table 4 shows the results and evaluation criteria outcomes. We find that all three reflective measurement models meet the relevant assessment criteria. More specifically, all the outer loadings are above 0.708, indicating that all indicators exhibit a sufficient level of reliability. Furthermore, all AVE values are above 0.50, providing support for the measures' convergent validity. The composite reliability ρ_c has values of 0.869 and higher, which is clearly above the expected minimum level of 0.70. Moreover, the Cronbach's alpha values range between 0.776 and 0.831, which is acceptable. Finally, all composite reliability ρ_A values meet the 0.70 threshold. These results suggest that the construct measures of *COMP*, *CUSL*, and *LIKE* exhibit high levels of internal consistency reliability.

Finally, we assess the discriminant validity by using the HTMT criterion. All the results are clearly below the conservative threshold of 0.85 (Table 5). Next, we run the bootstrapping procedure with 10,000 samples with percentile bootstrap confidence intervals, and one-tailed testing at the 0.05 significance level (which corresponds to a two-sided 90% confidence interval). The results provide the HTMT confidence interval's upper bounds (i.e., UB_{95}) are below 0.85, suggesting that all the HTMT values are significantly different from this conservative threshold (Table 5). This even holds for *CUSA* and *CUSL* as well as *COMP* and *LIKE*, which are conceptually similar. We thus conclude that discriminant validity has been established.

The *CUSA* construct is not included in the reflective (and subsequent formative) measurement model assessment, because it is a single-item construct. For this construct, indicator data and latent variable scores are identical. Consequently,

Table 4 PLS-SEM assessment results of reflective measurement models

Latent variable	Indicators	Convergent validity			Internal consistency reliability		
		Loadings > 0.70	Indicator reliability > 0.50	AVE > 0.50	Cronbach's alpha 0.70–0.90	Reliability ρ_A > 0.70	Composite reliability ρ_c > 0.70
<i>COMP</i>	<i>comp_1</i>	0.824	0.679	0.688	0.776	0.786	0.869
	<i>comp_2</i>	0.821	0.674				
	<i>comp_3</i>	0.844	0.712				
<i>CUSL</i>	<i>cusl_1</i>	0.833	0.694	0.748	0.831	0.839	0.899
	<i>cusl_2</i>	0.917	0.841				
	<i>cusl_3</i>	0.843	0.711				
<i>LIKE</i>	<i>like_1</i>	0.880	0.774	0.747	0.831	0.836	0.899
	<i>like_2</i>	0.869	0.755				
	<i>like_3</i>	0.844	0.712				

Table 5 HTMT values

	COMP	CUSA	CUSL	LIKE
COMP				
CUSA	0.465 (UB ₉₅ : 0.552)			
CUSL	0.532 (UB ₉₅ : 0.618)	0.755 (UB ₉₅ : 0.809)		
LIKE	0.780 (UB ₉₅ : 0.843)	0.577 (UB ₉₅ : 0.640)	0.737 (UB ₉₅ : 0.803)	

Note: UB₉₅: represents the upper bounds of the 95% confidence interval

CUSA does not have a measurement model, which can be assessed using the standard evaluation criteria.

Formative Measurement Model Assessment

The formative measurement model assessment initially focuses on the constructs’ convergent validity by conducting a redundancy analysis of each construct (i.e., *ATTR*, *CSOR*, *PERF*, and *QUAL*). The redundancy analysis draws on global single items, which summarize the essence each formatively measured construct purports to measure. These single items have been included in the original questionnaire. For example, respondents had to answer the statement, “Please assess to which degree [the company] acts in socially conscious ways,” measured on a scale of 1 (not at all) to 7 (extremely). This question can be used as an endogenous single-item construct to validate the formative measurement of corporate social responsibility (*CSOR*). For this purpose, we need to create a new PLS path model for each formatively measured construct that explains the global measure as an endogenous single-item construct as shown in Fig. 7. All the path relationships between the formatively measured construct and its global single-item measure (i.e., 0.874, 0.857, 0.811, and 0.805) are above the critical value of 0.70. We thus conclude that convergent validity of the formatively measured constructs has been established.

Next, we assess whether critical levels of collinearity substantially affect the formative indicator weight estimates. We find that the highest *VIF* value (i.e., 2.269 for the formative indicator *qual_3*) is clearly below the more conservative threshold value of 3, suggesting that collinearity is not at a critical level.

Testing the indicator weights’ significance draws on the bootstrapping procedure (10,000 samples, percentile bootstrap confidence intervals, two-tailed testing at the 0.05 significance level). Table 6 shows the resulting 95% percentile confidence intervals. The results show that most of the indicator weights are significant, with the exception of *csor_2*, *csor_4*, *qual_2*, *qual_3*, and *qual_4*, whose indicator weight confidence intervals include the value 0. However, these indicators exhibit statistically significant loadings above the 0.50 threshold, providing support for their absolute contribution to the constructs. In addition, prior research has substantiated the relevance of these indicators for the measurement of the *CSOR* and *QUAL* constructs (Eberl 2010; Sarstedt et al. 2013; Schwaiger 2004). Therefore, we retain the nonsignificant, but relevant, indicators in the formative measurement models.

To summarize, the results of the reflective and formative measurement model assessment suggest that all construct measures exhibit satisfactory levels of

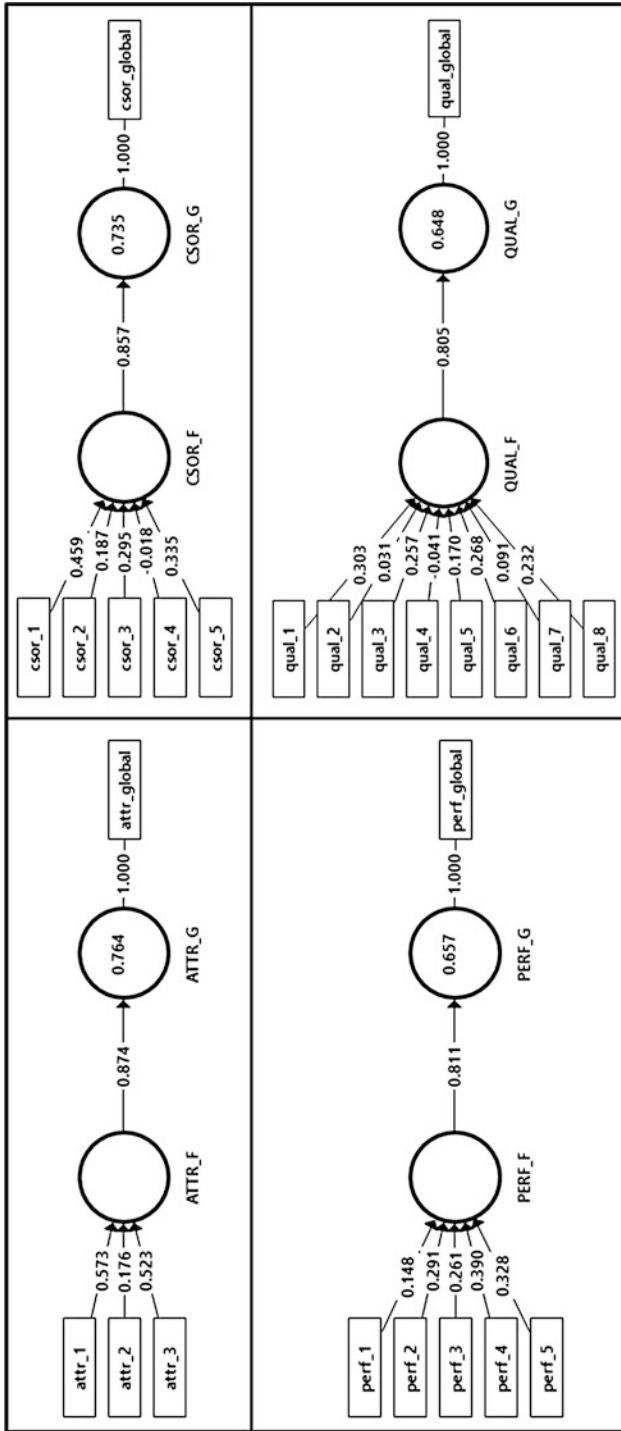


Fig. 7 Redundancy analysis

Table 6 Formative indicator weights and significance testing results

Formative constructs	Formative indicators	Outer weights (outer loadings)	95% confidence interval	Significant ($p < 0.05$)?
ATTR	<i>attr_1</i>	0.414 (0.755)	[0.273, 0.542]	Yes
	<i>attr_2</i>	0.201 (0.506)	[0.067, 0.320]	Yes
	<i>attr_3</i>	0.658 (0.891)	[0.541, 0.777]	Yes
CSOR	<i>csor_1</i>	0.306 (0.771)	[0.126, 0.461]	Yes
	<i>csor_2</i>	0.037 (0.571)	[-0.094, 0.187]	No
	<i>csor_3</i>	0.406 (0.838)	[0.244, 0.552]	Yes
	<i>csor_4</i>	0.080 (0.617)	[-0.070, 0.225]	No
	<i>csor_5</i>	0.416 (0.848)	[0.222, 0.6583]	Yes
PERF	<i>perf_1</i>	0.468 (0.846)	[0.329, 0.594]	Yes
	<i>perf_2</i>	0.177 (0.690)	[0.036, 0.305]	Yes
	<i>perf_3</i>	0.194 (0.573)	[0.092, 0.299]	Yes
	<i>perf_4</i>	0.340 (0.717)	[0.209, 0.475]	Yes
	<i>perf_5</i>	0.199 (0.638)	[0.075, 0.338]	Yes
QUAL	<i>qual_1</i>	0.202 (0.741)	[0.086, 0.309]	Yes
	<i>qual_2</i>	0.041 (0.570)	[-0.051, 0.143]	No
	<i>qual_3</i>	0.106 (0.749)	[-0.009, 0.217]	No
	<i>qual_4</i>	-0.005 (0.664)	[-0.106, 0.112]	No
	<i>qual_5</i>	0.160 (0.787)	[0.053, 0.270]	Yes
	<i>qual_6</i>	0.398 (0.856)	[0.268, 0.509]	Yes
	<i>qual_7</i>	0.229 (0.722)	[0.111, 0.334]	Yes
	<i>qual_8</i>	0.190 (0.627)	[0.066, 0.304]	Yes

reliability and validity. We can therefore proceed with the assessment of the structural model.

Structural Model Assessment

Following the structural model assessment procedure (Fig. 3), we first need to check the structural model for collinearity issues by examining the *VIF* values of all sets of predictor constructs in the model. Most *VIF* values are below the conservative threshold of 3, except for *QUAL* in the regressions of *COMP* and *LIKE* on the four formative predictor constructs. However, as the *VIF* value (3.487) is very close to 3, we conclude that collinearity among the predictor constructs is not a critical issue in the structural model.

When analyzing the path coefficient estimates of the structural model (Table 7), we start with the key target construct *CUSL* on the right-hand side of the PLS path model (Fig. 6). The construct *CUSA* (0.505) has the strongest effect on *CUSL*, followed by *LIKE* (0.344), while the effect of *COMP* (0.006) is very close to zero. Bootstrapping results substantiate that the effects of *CUSA* and *LIKE* on *CUSL* are significant, while *COMP* does not have a significant effect at the 5% probability of error level. Moreover, *COMP* has a significant, but relatively small effect on *CUSA* (0.146), while the effect of *LIKE* is relatively strong (0.436). We further find that the model explains 56.2% of *CUSL*'s variance (i.e., $R^2 = 0.562$), which is relatively high

Table 7 Path coefficients of the structural model and significance testing results

	Path coefficient	95% confidence interval	Significant ($p < 0.05$)?	f^2 effect size
<i>ATTR</i> → <i>COMP</i>	0.086	[-0.015, 0.190]	No	0.009
<i>ATTR</i> → <i>LIKE</i>	0.167	[0.034, 0.297]	Yes	0.030
<i>COMP</i> → <i>CUSA</i>	0.146	[0.008, 0.270]	Yes	0.018
<i>COMP</i> → <i>CUSL</i>	0.006	[-0.104, 0.112]	No	<0.001
<i>CSOR</i> → <i>COMP</i>	0.059	[-0.051, 0.169]	No	0.005
<i>CSOR</i> → <i>LIKE</i>	0.178	[0.070, 0.278]	Yes	0.035
<i>CUSA</i> → <i>CUSL</i>	0.505	[0.414, 0.584]	Yes	0.412
<i>LIKE</i> → <i>CUSA</i>	0.436	[0.321, 0.557]	Yes	0.159
<i>LIKE</i> → <i>CUSL</i>	0.344	[0.232, 0.457]	Yes	0.138
<i>PERF</i> → <i>COMP</i>	0.295	[0.167, 0.417]	Yes	0.082
<i>PERF</i> → <i>LIKE</i>	0.117	[-0.027, 0.250]	No	0.011
<i>QUAL</i> → <i>COMP</i>	0.430	[0.291, 0.550]	Yes	0.143
<i>QUAL</i> → <i>LIKE</i>	0.380	[0.272, 0.513]	Yes	0.094

taking into account that the model only considers the effects of customer satisfaction and the rather abstract concept of corporate reputation as predictors of customer loyalty. With a value of 0.292, R^2 of *CUSA* is clearly lower but still satisfactory, considering that only *LIKE* and *COMP* explain customer satisfaction in this model.

When analyzing the key predictors of *LIKE*, which has a substantial R^2 value of 0.558, we find that *QUAL* has the strongest significant effect (0.380), followed by *CSOR* (0.178), and *ATTR* (0.167). *PERF* (0.117) has the weakest effect on *LIKE*, which is not significant at the 5% level (Table 7). Corporate reputation's cognitive dimension *COMP* also has a substantial R^2 value of 0.631. Analyzing this construct's predictors shows that *QUAL* (0.430) and *PERF* (0.295) have the strongest significant effects. On the contrary, the effects of *ATTR* (0.086) and *CSOR* (0.059) on *COMP* are not significant at the 5% level. Analyzing the exogenous constructs' total effects on *CUSL* shows that *QUAL* has the strongest total effect (0.248), followed by *CSOR* (0.105), *ATTR* (0.101), and *PERF* (0.089). These results suggest that companies should focus on marketing activities that positively influence the customers' perception of the quality of their products and services.

Table 7 also shows the f^2 effect sizes. Relatively high f^2 effect sizes occur for the relationships *CUSA* → *CUSA* (0.412), *LIKE* → *CUSA* (0.159), *QUAL* → *COMP* (0.143) and *LIKE* → *CUSL* (0.138). These relationships also have particularly strong path coefficients of 0.30 and higher. Interestingly, the relationship between *QUAL* and *LIKE* has a strong path coefficient of 0.380, but only a weak f^2 effect size of 0.094. All the other f^2 effect sizes in the structural model are weak and, if below 0.02, negligible.

The next step is to assess the model's predictive power by running the PLS_{predict} procedure with ten folds and ten repetitions. The focus is on the model's key target construct *CUSL* and its three indicators *cusl_1*, *cusl_2*, and *cusl_3*. The results in Table 8 show that all three indicators achieve $Q^2_{predict}$ larger than zero, indicating that

Table 8 PLS_{predict} results

	$Q^2_{predict}$	RMSE	
		PLS-SEM	LM
<i>cust_1</i>	0.260	1.299	1.312
<i>cust_2</i>	0.234	1.522	1.538
<i>cust_3</i>	0.142	1.530	1.567

the model outperforms the naïve benchmark (i.e., the training sample means). Analyzing the prediction errors produced by the PLS path model shows their distribution is not highly unsymmetric. Hence, the following analyses focus on the RMSE statistic. The analysis shows that the RMSE values produced by the PLS path model are consistently lower than those of the LM benchmark. For example, while the PLS-SEM analysis produces an RMSE value of 1.299 for *cust_1*, the LM benchmark’s RMSE value is 1.312.

The final step involves comparing the original reputation model (Fig. 6) with an alternative, more complex model in which *ATTR*, *CSOR*, *PERF*, and *QUAL* additionally relate to *CUSA* and *CUSL*. As in the PLS_{predict} analysis, the focus is on the key target construct *CUSL*. Computing the BIC for these two models yields a value of -261.602 for the original model and -245.038 for the alternative, more complex model. This result provides empirical support for the original model. Similarly, Liengaard et al. (2021) support the established corporate reputation model using CVPAT when comparing it to an alternative version of this model.

Conclusions

Prior research discussing the benefits and limitations of PLS-SEM or analyzing its performance (e.g., in terms of parameter estimation) has usually not acknowledged that the method takes on a fundamentally different philosophy of measurement compared to factor-based SEM (e.g., Rhemtulla et al. 2020). Rather than assuming a common factor model structure, PLS-SEM draws on composite model logic to represent reflective and formative measurement models. The method linearly combines sets of indicators to form composites that represent the conceptual variables of interest (Lohmöller 1989; Wold 1982). Different from factor-based SEM, which equates constructs and the conceptual variables that they represent (Rigdon et al. 2019), PLS-SEM is an approximation method that inherently recognizes that constructs and conceptual variables are not identical (Rigdon et al. 2017). As Rigdon (2016, p. 19) notes, “common factor proxies cannot be assumed to carry greater significance than composite proxies in regard to the existence or nature of conceptual variables.”

PLS-SEM offers a good approximation of common factor models in situations where factor-based SEM (chapter ▶ “Structural Equation Modeling”) cannot deliver results due to its methodological limitations in terms of, for example, model complexity, sample size requirements, or inclusion of composite variables in the model (Reinartz et al. 2009; Sarstedt et al. 2016; Willaby et al. 2015). Bentler and Huang’s

(2014) PLSc as well as Dijkstra and Henseler's (2015b) PLSc algorithm allow researchers to mimic factor-based SEM results while benefiting from the original PLS-SEM method's flexibility in terms of model specification. Such an analysis assumes, however, that factor-based SEM is the correct estimator that delivers the true results as a benchmark for SEM (Hair et al. 2019a).

Most importantly, PLS-SEM constitutes a causal-predictive approach to SEM, which focuses on establishing the predictive power of a model, whose structure has been derived from theory and logic. PLS-SEM strikes a balance between factor-based SEM, which follows a confirmatory paradigm, and modern machine learning methods, which focus on prediction (Hair and Sarstedt 2021a) by providing a "cognitive path to predictions" (Douglas 2009, p. 454). As Hair and Sarstedt (2021a) note, "we live in a noisy, probabilistic world in which we can at best make imperfect predictions. In such a world, causal explanation reduces the complexity of the world to make it more manageable and understandable." At the same time, solely following the confirmation-only paradigm limits the practical usefulness of research as the 'correct' model does not necessarily exhibit high levels of predictive power.

While standard PLS-SEM analyses provide important insights into the strength and significance of the hypothesized model relationships, more advanced modeling and estimating techniques shed further light on the proposed relationships. Research has brought forward a variety of complementary analysis techniques and procedures, which extend the methodological toolbox of researchers working with the method (e.g., to conduct robustness checks; Sarstedt et al. 2020b). Examples of these methods include the confirmatory tetrad analysis (CTA-PLS), which enables researchers to statistically test if the measurement model operationalization should rather build on effect or composite indicators (Gudergan et al. 2008), and latent class techniques, which allow assessing if unobserved heterogeneity affects the model estimates. Prominent examples of latent class techniques for PLS-SEM include finite mixture partial least squares (Hahn et al. 2002; Sarstedt et al. 2011), PLS genetic algorithm segmentation (Ringle et al. 2014; Ringle et al. 2013), prediction-oriented segmentation (Becker et al. 2013b), iterative reweighted regressions (Schlittgen et al. 2016), and a modified *k*-means clustering approach (Fordellone and Vichi 2020). Further methods to account for heterogeneity in the structural model include the analysis of moderating effects (Memon et al. 2019), and the multigroup analysis (Matthews 2017), including testing for measurement invariance (Henseler et al. 2016b).

Approaches for combing PLS-SEM with the necessary condition analysis (NCA; Richter et al. 2020) and the fuzzy-set qualitative comparative analysis (fsQCA; e.g., Leischning et al. 2016; Rasoolimanesh et al. 2021), testing nonlinear effects (Hair et al. 2018b), higher-order constructs (Sarstedt et al. 2019), mediation effect (Nitzl et al. 2016), conditional process models (Sarstedt et al. 2020a), and model comparison using CVPAT (Liengaard et al. 2021) and its extensions for predictive model assessment and comparison (Sharma et al. 2021) complement the set of advanced PLS-SEM procedures. A further complementary method, the importance-performance map analysis (IPMA), facilitates richer outcome discussions in that it

extends the analysis of total effects in the model by adding a second results dimension to the analysis which incorporates the average values of the latent variables (Ringle and Sarstedt 2016). Finally, Hult et al. (2018) introduced a procedure for handling endogeneity in PLS path models, which occurs when a construct's error term is correlated with the scores of one or more explanatory variables in a partial regression relationship (chapter ▶ “Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”). In such a situation, path coefficient estimates become causally uninterpretable, which proves problematic in PLS-SEM analyses that have a strict confirmatory focus. Hair et al. (2018b) provide a more detailed overview and introduction to these complementary techniques for more advanced PLS-SEM analyses. Sarstedt et al. (2020b) discuss a series of robustness tests that draw on advanced modeling and model evaluation techniques.

Cross-References

- ▶ Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers
- ▶ Measuring Customer Satisfaction and Customer Loyalty
- ▶ Structural Equation Modeling

Acknowledgments This chapter uses the statistical software SmartPLS 3 (<https://www.smartpls.com>). Ringle acknowledges a financial interest in SmartPLS.

References

- Aaker, D. A. (1991). *Managing brand equity: Capitalizing on the value of a brand name*. New York: Free Press.
- Aguirre-Urreta, M. I., & Rönkkö, M. (2018). Statistical inference with PLS using bootstrap confidence intervals. *MIS Quarterly*, 42(3), 1001–1020.
- Akter, S., Fosso Wamba, S., & Dewan, S. (2017). Why PLS-SEM is suitable for complex modeling? An empirical illustration in big data analytics quality. *Production Planning & Control*, 28(11–12), 1011–1021.
- Albers, S. (2010). PLS and success factor studies in marketing. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (Springer handbooks of computational statistics series) (Vol. II, pp. 409–425). Berlin/Heidelberg: Springer.
- Ali, F., Rasoolimanesh, S. M., Sarstedt, M., Ringle, C. M., & Ryu, K. (2018). An assessment of the use of partial least squares structural equation modeling (PLS-SEM) in hospitality research. *The International Journal of Contemporary Hospitality Management*, 30(1), 514–538.
- Avkiran, N. K., & Ringle, C. M. (Eds.). (2018). *Partial least squares structural equation modeling: Recent advances in banking and finance*. Cham: Springer.
- Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 13(2), 139–161.
- Bayonne, E., Marin-Garcia, J. A., & Alfalla-Luque, R. (2020). Partial least squares (PLS) in operations management research: Insights from a systematic literature review. *Journal of Industrial Engineering and Management*, 13(3), 565–597.

- Becker, J.-M., & Ismail, I. R. (2016). Accounting for sampling weights in PLS path modeling: Simulations and empirical examples. *European Management Journal*, 34(6), 606–617.
- Becker, J.-M., Rai, A., & Rigdon, E. E. (2013a). Predictive validity and formative measurement in structural equation modeling: Embracing practical relevance. In *2013 Proceedings of the International Conference on Information Systems, Milan*.
- Becker, J.-M., Rai, A., Ringle, C. M., & Völckner, F. (2013b). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS Quarterly*, 37(3), 665–694.
- Bentler, P. M., & Huang, W. (2014). On components, latent variables, PLS and simple methods: Reactions to Rigdon's rethinking of PLS. *Long Range Planning*, 47(3), 138–145.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53(1), 605–634.
- Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, 35(2), 359–372.
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, 16(3), 265–284.
- Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal–formative indicators: A minority report. *Psychological Methods*, 22(3), 581–596.
- Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305–314.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203–219.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Heidelberg: Springer.
- Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, 15(1), 17–32.
- Centefelli, R. T., & Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *MIS Quarterly*, 33(4), 689–708.
- Cepeda Carrión, G., Cegarra-Navarro, J.-G., & Cillo, V. (2019). Tips to use partial least squares structural equation modelling (PLS-SEM) in knowledge management. *Journal of Knowledge Management*, 23(1), 67–89.
- Cheah, J.-H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM. *International Journal of Contemporary Hospitality Management*, 30(11), 3192–3210.
- Cheah, J.-H., Roldán, J. L., Ciavolino, E., Ting, H., & Ramayah, T. (2020). Sampling weight adjustments in partial least squares structural equation modeling: Guidelines and illustrations. *Total Quality Management & Business Excellence*, forthcoming.
- Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Mahwah: Lawrence Erlbaum.
- Chin, W. W. (2010). How to write up and report PLS analyses. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (Springer handbooks of computational statistics series) (Vol. II, pp. 655–690). Heidelberg: Springer.
- Chin, W. W., Marcolin, B. L., & Newsted, P. R. (2003). A partial least squares latent variable modeling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study. *Information Systems Research*, 14(2), 189–217.
- Chin, W. W., Cheah, J.-H., Liu, Y., Ting, H., Lim, X.-J., & Cham, T. H. (2020). Demystifying the role of causal-predictive modeling using partial least squares structural equation modeling in information systems research. *Industrial Management & Data Systems*, 120(12), 2161–2209.
- Cho, G., & Choi, J. Y. (2020). An empirical comparison of generalized structured component analysis and partial least squares path modeling under variance-based structural equation models. *Behaviormetrika*, 47, 243–272.

- Cho, G., Hwang, H., Kim, S., Lee, J., Sarstedt, M., & Ringle, C. M. (2021). A comparative study of the predictive power of component-based approaches to structural equation modeling. *Working Paper*.
- Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-Normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, 44(2), 347–357.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah: Lawrence Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Danks, N., & Ray, S. (2018). Predictions from partial least squares models. In F. Ali, S. M. Rasoolimanesh, & C. Cobanoglu (Eds.), *Applying partial least squares in tourism and hospitality research* (pp. 35–52). Bingley: Emerald.
- Danks, N. P., Sharma, P. N., & Sarstedt, M. (2020). Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLS-SEM). *Journal of Business Research*, 113, 13–24.
- Diamantopoulos, A. (2006). The error term in formative measurement models: Interpretation and modeling implications. *Journal of Modelling in Management*, 1(1), 7–17.
- Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Quarterly*, 35(2), 335–358.
- Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2), 269–277.
- Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, 40(3), 434–449.
- Dijkstra, T. K. (2010). Latent variables and indices: Herman Wold's basic design and partial least squares. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (Springer handbooks of computational statistics series) (Vol. II, pp. 23–46). Berlin/Heidelberg: Springer.
- Dijkstra, T. K. (2014). PLS' Janus face – Response to professor Rigdon's 'rethinking partial least squares modeling: In praise of simple methods'. *Long Range Planning*, 47(3), 146–153.
- Dijkstra, T. K., & Henseler, J. (2015a). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics & Data Analysis*, 81, 10–23.
- Dijkstra, T. K., & Henseler, J. (2015b). Consistent partial least squares path modeling. *MIS Quarterly*, 39(2), 297–316.
- do Valle, P. O., & Assaker, G. (2016). Using partial least squares structural equation modeling in tourism research: A review of past research and recommendations for future applications. *Journal of Travel Research*, 55(6), 695–708.
- Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, 76(4), 444–463.
- Eberl, M. (2010). An application of PLS in multi-group analysis: The need for differentiated corporate-level Marketing in the Mobile Communications Industry. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (Springer handbooks of computational statistics series) (Vol. II, pp. 487–514). Berlin/Heidelberg: Springer.
- Eberl, M., & Schwaiger, M. (2005). Corporate reputation: Disentangling the effects on financial performance. *European Journal of Marketing*, 39(7/8), 838–854.
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5(2), 155–174.
- Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (Eds.). (2010). *Handbook of partial least squares: Concepts, methods and applications* (Springer handbooks of computational statistics series) (Vol. II). Heidelberg: Springer.
- Evermann, J., & Tate, M. (2016). Assessing the predictive performance of structural equation model estimators. *Journal of Business Research*, 69(10), 4565–4582.

- Falk, R. F., & Miller, N. B. (1992). *A primer for soft modeling*. Akron: University of Akron Press.
- Fordellone, M., & Vichi, M. (2020). Finding groups in structural equation modeling through the partial least squares algorithm. *Computational Statistics & Data Analysis*, *147*, 106957.
- Fornell, C. G., & Bookstein, F. L. (1982). Two structural equation models: LISREL and PLS applied to consumer exit-voice theory. *Journal of Marketing Research*, *19*(4), 440–452.
- Fornell, C. G., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: Nature, purpose, and findings. *Journal of Marketing*, *60*(4), 7–18.
- Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: A comparison of four procedures. *Internet Research*, *29*(3), 430–447.
- Garson, G. D. (2016). *Partial least squares regression and structural equation models*. Asheboro: Statistical Associates.
- George, D., & Mallery, P. (2019). *IBM SPSS statistics 25 step by step: A simple guide and reference* (15th ed.). New York: Routledge.
- Geweke, J., & Meese, R. (1981). Estimating regression models of finite but unknown order. *International Economic Review*, *22*(1), 55–70.
- Ghasemy, M., Teeroovengadam, V., Becker, J.-M., & Ringle, C. M. (2020). This fast car can move faster: A review of PLS-SEM application in higher education research. *Higher Education*, *80*, 1121–1152.
- Goodhue, D. L., Lewis, W., & Thompson, R. (2012). Does PLS have advantages for small sample size or non-Normal data? *MIS Quarterly*, *36*(3), 981–1001.
- Grace, J. B., & Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: The role of composite variables. *Environmental and Ecological Statistics*, *15*(2), 191–213.
- Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611–642.
- Gudergan, S. P., Ringle, C. M., Wende, S., & Will, A. (2008). Confirmatory tetrad analysis in PLS path modeling. *Journal of Business Research*, *61*(12), 1238–1249.
- Haenlein, M., & Kaplan, A. M. (2004). A Beginner's guide to partial least squares analysis. *Understanding Statistics*, *3*(4), 283–297.
- Hahn, C., Johnson, M. D., Herrmann, A., & Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review*, *54*(3), 243–269.
- Hair, J. F. (2021). Next-generation prediction metrics for composite-based PLS-SEM. *Industrial Management & Data Systems*, *121*(1), 5–11.
- Hair, J. F., & Sarstedt, M. (2019). Composites vs. factors: Implications for choosing the right SEM method. *Project Management Journal*, *50*(6), 1–6.
- Hair, J. F., & Sarstedt, M. (2021a). Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing. *Journal of Marketing Theory & Practice*, *29*(1), 65–77.
- Hair, J. F., & Sarstedt, M. (2021b). Explanation plus prediction – The logical focus of project management research. *Project Management Journal*, forthcoming.
- Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–151.
- Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012a). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, *45*(5-6), 320–340.
- Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012b). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, *40*(3), 414–433.
- Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better results and higher acceptance. *Long Range Planning*, *46*(1-2), 1–12.
- Hair, J. F., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017a). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management & Data Systems*, *117*(3), 442–458.

- Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., & Thiele, K. O. (2017b). Mirror, mirror on the wall: A comparative evaluation of composite-based structural equation modeling methods. *Journal of the Academy of Marketing Science*, 45(5), 616–632.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2018a). *Multivariate data analysis* (8th ed.). Mason: Cengage.
- Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2018b). *Advanced issues in partial least squares structural equation modeling (PLS-SEM)*. Thousand Oaks: Sage.
- Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019a). When to use and how to report the results of PLS-SEM. *European Business Review*, 31(1), 2–24.
- Hair, J. F., Sarstedt, M., & Ringle, C. M. (2019b). Rethinking some of the rethinking of partial least squares. *European Journal of Marketing*, 53(4), 566–584.
- Hair, J. F., Howard, M. C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research*, 109, 101–110.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Thousand Oaks: Sage.
- Helm, S., Eggert, A., & Garnefeld, I. (2010). Modelling the impact of corporate reputation on customer satisfaction and loyalty using PLS. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of partial least squares: Concepts, methods and applications* (Springer handbooks of computational statistics series) (Vol. II, pp. 515–534). Heidelberg: Springer.
- Henseler, J. (2017). Using variance-based structural equation modeling for empirical advertising research at the Interface of design and behavioral research. *Journal of Advertising*, 46(1), 178–192.
- Henseler, J. (2021). *Composite-based structural equation modeling: Analyzing latent and emergent variables*. New York: Guilford Press.
- Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics*, 28(2), 565–580.
- Henseler, J., & Schuberth, F. (2020). Using confirmatory composite analysis to assess emergent variables in business research. *Journal of Business Research*, 120, 147–156.
- Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. In R. R. Sinkovics & P. N. Ghauri (Eds.), *Advances in international marketing* (Vol. 20, pp. 277–320). Bingley: Emerald.
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2012). Using partial least squares path modeling in international advertising research: Basic concepts and recent issues. In S. Okazaki (Ed.), *Handbook of research in international advertising* (pp. 252–276). Cheltenham: Edward Elgar Publishing.
- Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., Ketchen, D. J., Hair, J. F., Hult, G. T. M., & Calantone, R. J. (2014). Common beliefs and reality about partial least squares: Comments on Rönkkö & Evermann (2013). *Organizational Research Methods*, 17(2), 182–209.
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135.
- Henseler, J., Hubona, G. S., & Ray, P. A. (2016a). Using PLS path modeling in new technology research: Updated guidelines. *Industrial Management & Data Systems*, 116(1), 2–20.
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2016b). Testing measurement invariance of composites using partial least squares. *International Marketing Review*, 33(3), 405–431.
- Houston, M. B. (2004). Assessing the validity of secondary data proxies for marketing constructs. *Journal of Business Research*, 57(2), 154–161.
- Hui, B. S., & Wold, H. (1982). Consistency and consistency at large of partial least squares estimates. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation, part II* (pp. 119–130). Amsterdam: North-Holland.

- Hult, G. T. M., Hair, J. F., Dorian, P., Ringle, C. M., Sarstedt, M., & Pinkwart, A. (2018). Addressing endogeneity in marketing applications of partial least squares structural equation modeling. *Journal of International Marketing*, 26(3), 1–21.
- Hwang, H., Sarstedt, M., Cheah, J.-H., & Ringle, C. M. (2020). A concept analysis of methodological research on composite-based structural equation modeling: Bridging PLSPM and GSCA. *Behaviormetrika*, 47(1), 219–241.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409–426.
- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 255–284). New York: Seminar Press.
- Jöreskog, K. G., & Wold, H. (1982). The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. In H. Wold & K. G. Jöreskog (Eds.), *Systems under indirect observation, part I* (pp. 263–270). Amsterdam: North-Holland.
- Kaufmann, L., & Gaeckler, J. (2015). A structured review of partial least squares in supply chain management research. *Journal of Purchasing and Supply Management*, 21(4), 259–272.
- Khan, G., Sarstedt, M., Shiao, W.-L., Hair, J. F., Ringle, C. M., & Fritze, M. (2019). Methodological research on partial least squares structural equation modeling (PLS-SEM): A social network analysis. *Internet Research*, 29(3), 407–429.
- Kock, N., & Hadaya, P. (2018). Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Information Systems Journal*, 28(1), 227–261.
- Latan, H., & Noonan, R. (Eds.). (2017). *Partial least squares structural equation modeling: Basic concepts, methodological issues and applications*. Berlin/Heidelberg: Springer.
- Lee, L., Petter, S., Fayard, D., & Robinson, S. (2011). On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems*, 12(4), 305–328.
- Lei, P.-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–179). New York: Guilford Press.
- Lienggaard, B. D., Sharma, P. N., Hult, G. T. M., Jensen, M. B., Sarstedt, M., Hair, J. F., & Ringle, C. M. (2021). Prediction: Coveted, yet forsaken? Introducing a cross-validated predictive ability test in partial least squares path modeling. *Decision Sciences*, 52(2), 362–292.
- Leischnig, A., Henneberg, S. C., & Thornton, S. C. (2016). Net versus combinatory effects of firm and industry antecedents of sales growth. *Journal of Business Research*, 69(9), 3576–3583.
- Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg: Physica.
- Manley, S. C., Hair, J. F., Williams, R. I., & McDowell, W. C. (2020). Essential new PLS-SEM analysis methods for your entrepreneurship analytical toolbox. *International Entrepreneurship and Management Journal*, forthcoming.
- Marcoulides, G. A., & Chin, W. W. (2013). You write, but others read: Common methodological misunderstandings in PLS and related methods. In H. Abdi, W. W. Chin, V. Esposito Vinzi, G. Russolillo, & L. Trinchera (Eds.), *New perspectives in partial least squares and related methods* (Springer proceedings in Mathematics & Statistics) (Vol. 56, pp. 31–64). New York: Springer.
- Marcoulides, G. A., & Saunders, C. (2006). Editor's comments: PLS: A silver bullet? *MIS Quarterly*, 30(2), iii–ix.
- Marcoulides, G. A., Chin, W. W., & Saunders, C. (2012). When imprecise statistical statements become problematic: A response to Goodhue, Lewis, and Thompson. *MIS Quarterly*, 36(3), 717–728.
- Mason, C. H., & Perreault, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28(3), 268–280.
- Mateos-Aparicio, G. (2011). Partial least squares (PLS) methods: Origins, evolution, and application to social sciences. *Communications in Statistics - Theory and Methods*, 40(13), 2305–2317.

- Matthews, L. (2017). Applying multigroup analysis in PLS-SEM: A step-by-step process. In H. Latan & R. Noonan (Eds.), *Partial least squares path modeling: Basic concepts, methodological issues and applications* (pp. 219–243). Cham: Springer.
- McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, 31(2), 239–270.
- Mehmetoglu, M., & Venturini, S. (2021). *Structural equation modelling with partial least squares using Stata and R*. Boca Raton: CRC Press.
- Memon, M. A., Cheah, J. H., Ramayah, H. T., Chuah, F., & Cham, T. H. (2019). Moderation analysis: Issues and guidelines. *Journal of Applied Structural Equation Modeling*, 3(1), i–xi.
- Nitzl, C. (2016). The use of partial least squares structural equation modelling (PLS-SEM) in management accounting research: Directions for future theory development. *Journal of Accounting Literature*, 37, 19–35.
- Nitzl, C., & Chin, W. W. (2017). The case of partial least squares (PLS) path modeling in managerial accounting. *Journal of Management Control*, 28(2), 137–156.
- Nitzl, C., Roldán, J. L., & Cepeda Carrión, G. (2016). Mediation analysis in partial least squares path modeling: Helping researchers discuss more sophisticated models. *Industrial Management & Data Systems*, 119(9), 1849–1864.
- Noonan, R., & Wold, H. (1982). PLS path modeling with indirectly observed variables: A comparison of alternative estimates for the latent variable. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observations: Part II* (pp. 75–94). Amsterdam: North-Holland.
- Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4), 557–595.
- Peng, D. X., & Lai, F. (2012). Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management*, 30(6), 467–480.
- Raithel, S., & Schwaiger, M. (2015). The effects of corporate reputation perceptions of the general public on shareholder value. *Strategic Management Journal*, 36(6), 945–956.
- Raithel, S., Sarstedt, M., Scharf, S., & Schwaiger, M. (2012). On the value relevance of customer satisfaction: Multiple drivers and multiple markets. *Journal of the Academy of Marketing Science*, 40(4), 509–525.
- Ramayah, T., Cheah, J., Chuah, F., Ting, H., & Memon, M. A. (2016). *Partial least squares structural equation modeling (PLS-SEM) using SmartPLS 3.0: An updated and practical guide to statistical analysis*. Kuala Lumpur: Pearson.
- Rasoolimanesh, S. M., Ringle, C. M., Sarstedt, M., & Olya, H. (2021). The combined use of symmetric and asymmetric approaches: Partial least squares-structural equation modeling and fuzzy-set qualitative comparative analysis. *International Journal of Contemporary Hospitality Management*, forthcoming.
- Reinartz, W. J., Haenlein, M., & Henseler, J. (2009). An empirical comparison of the efficacy of covariance-based and variance-based SEM. *International Journal of Research in Marketing*, 26(4), 332–344.
- Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30–45.
- Richter, N. F., Sinkovics, R. R., Ringle, C. M., & Schlägel, C. (2016). A critical look at the use of SEM in international business research. *International Marketing Review*, 33(3), 376–404.
- Richter, N. F., Schubring, S., Hauff, S., Ringle, C. M., & Sarstedt, M. (2020). When predictors of outcomes are necessary: Guidelines for the combined use of PLS-SEM and NCA. *Industrial Management & Data Systems*, 120(12), 2243–2267.
- Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, 45(5–6), 341–358.

- Rigdon, E. E. (2013). Partial least squares path modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling. A second course* (2nd ed., pp. 81–116). Charlotte: Information Age Publishing.
- Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal*, 34(6), 598–605.
- Rigdon, E. E., Becker, J.-M., Rai, A., Ringle, C. M., Diamantopoulos, A., Karahanna, E., Straub, D., & Dijkstra, T. K. (2014). Conflating antecedents and formative indicators: A comment on Aguirre-Urreta and Marakas. *Information Systems Research*, 25(4), 780–784.
- Rigdon, E. E., Sarstedt, M., & Ringle, C. M. (2017). On comparing Results from CB-SEM and PLS-SEM. Five perspectives and five recommendations. *Marketing ZFP—Journal of Research and Management*, 39(3), 4–16.
- Rigdon, E. E., Becker, J. M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research*, 54(3), 429–443.
- Ringle, C. M. (2019). What makes a great textbook? Lessons learned from Joe Hair. In B. J. Babin & M. Sarstedt (Eds.), *The great facilitator: Reflections on the contributions of Joseph F. Hair, Jr. to marketing and business research* (pp. 131–150). Cham: Springer.
- Ringle, C. M., & Sarstedt, M. (2016). Gain more insight from your PLS-SEM results: The importance-performance map analysis. *Industrial Management & Data Systems*, 116(9), 1865–1886.
- Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: A critical look at the use of PLS-SEM in MIS quarterly. *MIS Quarterly*, 36(1), iii–xiv.
- Ringle, C. M., Sarstedt, M., Schlittgen, R., & Taylor, C. R. (2013). PLS path modeling and evolutionary segmentation. *Journal of Business Research*, 66(9), 1318–1324.
- Ringle, C. M., Sarstedt, M., & Schlittgen, R. (2014). Genetic algorithm segmentation in partial least squares structural equation modeling. *OR Spectrum*, 36(1), 251–276.
- Ringle, C. M., Wende, S., & Becker, J.-M. (2015). *SmartPLS 3 [computer software]*. Bönningstedt: SmartPLS. Retrieved from <https://www.smartpls.com>.
- Ringle, C. M., Sarstedt, M., Mitchell, R., & Gudergan, S. P. (2020). Partial least squares structural equation modeling in HRM research. *International Journal of Human Resource Management*, 31(12), 1617–1643.
- Roldán, J. L., & Sánchez-Franco, M. J. (2012). Variance-based structural equation modeling: Guidelines for using partial least squares in information systems research. In M. Mora, O. Gelman, A. L. Steenkamp, & M. Raisinghani (Eds.), *Research methodologies, innovations and philosophies in software systems engineering and information systems* (pp. 193–221). Hershey: IGI Global.
- Russo, D., & Stol, K. J. (2021). PLS-SEM for software engineering research: An introduction and survey. *ACM Computing Surveys*, 54(4), 1–38.
- Sarstedt, M. (2019). Der Knacks and a Silver Bullet. In B. J. Babin & M. Sarstedt (Eds.), *The great facilitator: Reflections on the contributions of Joseph F. Hair, Jr. to marketing and business research* (pp. 155–164). Cham: Springer.
- Sarstedt, M., & Cheah, J.-H. (2019). Partial least squares structural equation modeling using SmartPLS: A software review. *Journal of Marketing Analytics*, 7(3), 196–202.
- Sarstedt, M., & Mooi, E. (2019). *A concise guide to market research: The process, data, and methods using IBM SPSS statistics* (3rd ed.). Berlin/Heidelberg: Springer.
- Sarstedt, M., Becker, J.-M., Ringle, C. M., & Schwaiger, M. (2011). Uncovering and treating unobserved heterogeneity with FIMIX-PLS: Which model selection criterion provides an appropriate number of segments? *Schmalenbach Business Review*, 63(1), 34–62.
- Sarstedt, M., Wilczynski, P., & Melewar, T. C. (2013). Measuring reputation in global markets – A comparison of reputation measures' convergent and criterion validities. *Journal of World Business*, 48(3), 329–339.
- Sarstedt, M., Ringle, C. M., Smith, D., Reams, R., & Hair, J. F. (2014). Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers. *Journal of Family Business Strategy*, 5(1), 105–115.

- Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, 69(10), 3998–4010.
- Sarstedt, M., Hair, J. F., Cheah, J.-H., Becker, J.-M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order models. *Australasian Marketing Journal*, 27(3), 197–211.
- Sarstedt, M., Hair, J. F., Nitzl, C., Ringle, C. M., & Howard, M. C. (2020a). Beyond a tandem analysis of SEM and PROCESS: Use PLS-SEM for mediation analyses! *International Journal of Market Research*, 62(3), 288–299.
- Sarstedt, M., Ringle, C. M., Cheah, J. H., Ting, H., Moisescu, O. I., & Radomir, L. (2020b). Structural model robustness checks in PLS-SEM. *Tourism Economics*, 26(4), 531–554.
- Schlittgen, R., Ringle, C. M., Sarstedt, M., & Becker, J.-M. (2016). Segmentation of PLS path models by iterative reweighted regressions. *Journal of Business Research*, 69(10), 4583–4592.
- Schloderer, M. P., Sarstedt, M., & Ringle, C. M. (2014). The relevance of reputation in the nonprofit sector: The moderating effect of socio-demographic characteristics. *International Journal of Nonprofit and Voluntary Sector Marketing*, 19(2), 110–126.
- Schuberth, F., Henseler, J., & Dijkstra, T. K. (2018). Confirmatory composite analysis. *Frontiers in Psychology*, 9, 2541.
- Schwaiger, M. (2004). Components and parameters of corporate reputation: An empirical study. *Schmalenbach Business Review*, 56(1), 46–71.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, 24(2), 148–169.
- Sharma, P. N., Shmueli, G., Sarstedt, M., Danks, N., & Ray S. (2018). Prediction-oriented model selection in partial least squares path modeling. *Decision Sciences*, forthcoming.
- Sharma, P. N., Liengard, B. D., Hair, J. F., Sarstedt, M., & Ringle C. M. (2021). Predictive model assessment and selection in composite-based modeling using PLS-SEM: Extensions and guidelines for using CVPAT. *Working Paper*.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553–572.
- Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. B. (2016). The elephant in the room: Evaluating the predictive performance of PLS models. *Journal of Business Research*, 69(10), 4552–4564.
- Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J.-H., Ting, H., & Ringle, C. M. (2019). Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. *European Journal of Marketing*, 53(11), 2322–2347.
- Shugan, S. (2009). Relevancy is robust prediction, not alleged realism. *Marketing Science*, 28(5), 991–998.
- Stieglitz, S., Linh, D.-X., Bruns, A., & Neuberger, C. (2014). Social media analytics. An interdisciplinary approach and its implications for information systems. *Business and Information Systems Engineering*, 6, 89–96.
- Streukens, S., & Leroi-Werelds, S. (2016). Bootstrapping and PLS-SEM: A step-by-step guide to get more out of your bootstrap results. *European Management Journal*, 34(6), 618–632.
- Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1), 159–205.
- Usakli, A., & Kucukergin, K. G. (2018). Using partial least squares structural equation modeling in hospitality and tourism: Do researchers follow practical guidelines? *International Journal of Contemporary Hospitality Management*, 30(11), 3462–3512.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478.
- Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192–196.
- Westland, J. C. (2019). Partial least squares path analysis. In *Structural equation models: From paths to networks* (2nd ed., pp. 17–38). Cham: Springer.

- Willaby, H. W., Costa, D. S. J., Burns, B. D., MacCann, C., & Roberts, R. D. (2015). Testing complex models with small sample sizes: A historical overview and empirical demonstration of what partial least squares (PLS) can offer differential psychology. *Personality and Individual Differences, 84*, 73–78.
- Wold, H. (1975). Path models with latent variables: The NIPALS approach. In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capocchi (Eds.), *Quantitative sociology: International perspectives on mathematical and statistical modeling* (pp. 307–357). New York: Academic.
- Wold, H. (1980). Model construction and evaluation when theoretical knowledge is scarce: Theory and application of PLS. In J. Kmenta & J. B. Ramsey (Eds.), *Evaluation of econometric models* (pp. 47–74). New York: Academic.
- Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observations: Part II* (pp. 1–54). Amsterdam: North-Holland.
- Wold, H. (1985). Partial least squares. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 581–591). New York: Wiley.
- Wong, K. K. K. (2019). *Mastering partial least squares structural equation modeling (PLS-SEM) with SmartPLS in 38 hours*. Bloomington: iUniverse.
- Zeng, N., Liu, Y., Gong, P, Hertogh, M., & König, M. (2021). Do right PLS and do PLS right: A critical review of the application on PLS in construction management reserarch. *Frontiers of Engineering Management*, forthcoming.



Automated Text Analysis

Ashlee Humphreys

Contents

Introduction	634
Foundations of Text Analysis	635
History	635
Approaches to Text Analysis	635
Dictionary-Based Methods	636
Classification Methods	638
Topic Modeling	638
Market Research Applications of Text Analysis	639
Sentiment Analysis	640
Studying Word of Mouth Through Text Analysis	641
Topic Discovery and Creating Positioning Maps from Online Text	642
Measurement of the Organization and Firm Environment	642
Issues in Working with Textual Data	643
Extended Example: Word-Of-Mouth Differences Between Experts and Nonexperts to a Product Launch	644
Purpose	644
Stage 1: Develop a Research Question	645
Stage 2: Data Collection	646
Stage 3: Construct Definition	647
Stage 4: Operationalization	648
Stage 5: Interpretation and Analysis	650
Stage 6: Validation	657
Conclusion and Future Directions	659
Cross-References	659
References	659

A. Humphreys (✉)

Integrated Marketing Communications, Medill School of Journalism, Media, and Integrated
Marketing Communications, Northwestern University, Evanston, IL, USA
e-mail: a-humphreys@northwestern.edu

Abstract

The amount of text available for analysis by marketing researchers has grown exponentially in the last two decades. Consumer reviews, message board forums, and social media feeds are just a few sources of data about consumer thought, interaction, and culture. However, written language is filled with complex meaning, ambiguity, and nuance. How can marketing researchers possibly transform this rich linguistic representation into quantifiable data for statistical analysis and modeling? This chapter provides an introduction to text analysis, covering approaches that range from top-down deductive methods to bottom-up inductive methods for text mining. After covering some foundational aspects of text analysis, applications to marketing research such as sentiment analysis, topic modeling, and studying organizational communication are summarized and explored, including a case study of word-of-mouth response to a product launch.

Keywords

Text analysis · computer-assisted text analysis · automated content analysis · content analysis · topic modeling · sentiment analysis · LDA · word-of-mouth

Introduction

Automated or computer-assisted text analysis describes a family of methods for parsing, classifying, and then quantifying textual data for further statistical analysis. Although automated text analysis using computers dates to the 1960s, the rise of digital technology for communicating has created a deluge of textual data for analysis and increased managerial desire to gain insights from text produced by consumers. Platforms like Twitter and Facebook provide a space for consumer-to-consumer discussion of products, brands, and services. Retail sites like Amazon, Best Buy, and Zappos and review sites like CNET and Yelp! host consumer reviews on a nearly endless array of products and services. Particular brand sites like Sephora, Gap, and Brooks Brothers offer social shopping capabilities such as consumer reviews represented by stars and extensive product reviews that detail fit, material, and quality (Stephen and Toubia 2010). This text from consumers, firms, and the media can provide insight into consumer needs and wants, sentiment, market structure, and transmission of word-of-mouth communication.

This chapter presents a high-level overview of methods for conducting text analysis in market research and provides resources for further investigating the methodological details depending on the approach one takes to text analysis.

Foundations of Text Analysis

History

To understand the implementation of automatic analysis, it will help to first review its relation to and its emergence from traditional content analysis. Content analysis is a method used in the social sciences to systematically assess and analyze the content of a message, usually in the form of text. Although traditions of content analysis go as far back as sixteenth-century monastic life, modern content analysis was first proposed by Max Weber (1924) to study the press. Since then, scholars in sociology and communications have used human-coded content analysis to investigate differences in media content, describe trends in communications over time, reveal patterns of organizational or individual attention, and examine attitudes, interests, intentions, or values of an individual or a group (e.g., Berelson 1971; Gamson and Modigliani 1989).

Traditional content analysis was first introduced to consumer behavior with Kassarijian's (1977) outline of the method and was then updated by Kolbe and Burnett (1991) in an attempt to improve reliability and objectivity, focusing primarily on standards for calculating inter-coder agreement (see also Grayson and Rust 2001). In consumer research and marketing, traditional content analysis has been used to analyze trends in magazine advertisements (Belk and Pollay 1985), direct mail (Stevenson and Swayne 1999), newspaper articles (Garrett 1987), and word-of-mouth communication (Moore 2015; Phelps et al. 2004) to name a few. Although automated text analysis can improve the efficiency and reliability of traditional content analysis, it also has limitations. For instance, computerized text analysis can miss subtleties in the text and cannot code finer shades of meaning. While dealing with negation is possible (Jia et al. 2009; Villarroel Ordenes et al. 2017), it remains somewhat analytically onerous.

Automated text analysis is not radically new, but it has become easier to implement since the widespread of adoption of the personal computer. The General Inquirer (Stone 1966) was one of the first computer content-analytic tools used in consumer research (Kranz 1970). Since then, vast strides have been made in automated text analysis. Kranz's (1970) early three-page treatment of computer-assisted content analysis in marketing deals with dictionary creation, but does not address category creation, validity, or measurement decisions. Since then, a variety of approaches have emerged.

Approaches to Text Analysis

In current practice, there are essentially two orientations toward automated text analysis: top-down vs. bottom-up approaches (Boyd and Pennebaker 2015a; Mehl and Gill 2008). The top-down approach counts concepts of interest, identified either

through a list of words or through a set of rules. Top-down, also called dictionary-based, methods are deductively or theoretically driven in the sense that researchers use them to look for hypothesized patterns in text from a known set of concepts. Bottom-up approaches, on the other hand, code all concepts present in the text and then look for patterns (Rayson 2009). These approaches can range considerably from methods of supervised learning, where researchers define some preliminary categories and then train the computer to sort documents based on latent differences, to discovery-oriented approaches such as calculating then flagging statistically significant differences between groups of texts (Rayson 2009), or fully automated processes where a computer identifies topics based on word co-occurrence (Lee and Bradlow 2011). In this way, bottom-up approaches to text analysis become similar to data mining approaches. That is, first the researcher looks at all differences in the data and builds conclusions from those differences.

Top-down, dictionary-based methods have been used extensively in social sciences like consumer research (Humphreys and Wang 2018), psychology (Chung and Pennebaker 2013; Mehl and Gill 2008; Pennebaker and King 1999), sociology (Van de Rijdt et al. 2013), and political science (Grimmer and Stewart 2013; Lasswell and Leites 1949) due to their ability to translate theoretical constructs into text and the transparency in reporting results and reliabilities. Bottom-up methods, on the other hand, have been used more extensively in engineering, computer science, and marketing science. Marketing strategy has drawn from both approaches, although dictionary-based approaches appear to be more common (Ertimur and Coskuner-Balli 2015; Humphreys 2010; Ludwig et al. 2013; Packard et al. 2014). This chapter briefly covers the fundamentals of each approach before moving to their application in marketing.

Dictionary-Based Methods

Dictionary-based methods for text analysis are based on a predeveloped word list, or dictionary, for counting the occurrence of words in a text. Standardized dictionaries are available for many constructs such as sentiment (e.g., Hutto and Gilbert 2014), marketing-related constructs like authenticity and brand personality (Kovács et al. 2013; Opoku et al. 2006), as well as many standard concepts in psychology (Pennebaker et al. 2001; Snefjella and Kuperman 2015) and other fields like political science (Dunphy et al. 1974; Stone 1966). In addition to using a standard dictionary, many researchers choose to create their own dictionary to fit the specific context, although this should be done only if a standard dictionary is not available.

There are several methods for dictionary creation ranging from inductive to deductive. The most inductive method of dictionary creation is to work from a concordance, or all words in the document listed in terms of frequency and group words according to relevant categories for the research question and hypothesis (Chung and Pennebaker 2013). If the researcher does not know what categories are relevant a priori, qualitative methods of reading and coding the text prior to dictionary development can be used to create a set of relevant concepts and a list of words

for their operationalization in text (Humphreys 2010). For example, to study institutional logics pertaining to the Yoga industry in newspaper articles, Ertimur and Coskuner-Balli (2015) first open and then axially code a dataset of newspaper articles and other historical texts. Generally, a random sample of 10–20% of the dataset is sufficient for coding (Humphreys and Wang 2018), but researchers should be mindful of unevenness in data quantity according to category or time period and stratify accordingly (Humphreys 2010). The most deductive method for dictionary creation is to create a wordlist from theoretical concepts or categories. However, one should be mindful of the tendency for researchers and writers to pick more abstract words than are generally present in textual data (Palmquist et al. 2009). For this reason, careful postmeasurement validation is necessary to ensure construct validity. After text is cleaned and stored and the dictionary has been created, researchers use a program like Diction, LIWC, WordStat, or R to execute counts. Data can then be saved and analyzed using a traditional statistical package or, for some packages like Wordstat and R, analyzed within the same package.

After calculating word frequencies, postmeasurement validation should be performed, and for this there are a variety of methods ranging from methods that are iterative with dictionary development to stand-alone calculations of inter-rater reliability. Weber (2005) suggests a saturation procedure whereby researchers pull a sample of 10 or 20 instances of a concept and have a research assistant code them as accurately representing the category (or not). If the rate is below 80%, the dictionary category should be revised until the threshold is met. Pennebaker et al. (2001) recommend a method of validating the dictionary, but not the resulting measurements. Here, three research assistants count a word as being representative of the category or not, and words are retained if two of the three coders agree. If they do not, the word should be dropped from the dictionary. Percentage agreements on dictionary categories can then be calculated and reported, and the general threshold is similar to that for Krippendorff's alpha, above 75%. A final option is to compare the computer-coded results with an extensive set of human-coded results from two or more coders. To do this, one selects a random sample from the dataset (the amount may vary depending on the size of the dataset) and human coders code the text according to the category descriptions, calculating reliability as one would in a traditional content analysis. This can then be compared to the additional "coder" of the computer to produce a similarity score. Although this final method has the advantage of comparison with traditional content analysis, it is not always necessary and in some cases can produce misguided results. Human coders pick up on subtle meanings that computers cannot and likewise computers are able to code concepts consistently and evenly over an entire dataset without omission or bias. For this reason, comparing human to computer coding can in some cases be like comparing apples to oranges.

Dictionary-based analyses have studied a wide range of theoretical concepts such as emotion (Berger and Milkman 2012), construal level (Sneffjella and Kuperman 2015), institutional logics (Ertimur and Coskuner-Balli 2015), risk (Humphreys and Thompson 2014), speech acts (Ludwig et al. 2016; Villarroel Ordenes et al. 2017), and framing (Fiss and Hirsch 2005; Humphreys and Latour 2013; Jurafsky et al.

2014). A wide variety of contexts can be explored through dictionary-based analysis such as product and restaurant reviews (Barasch and Berger 2014, Jurafsky et al. 2014; Kovács et al. 2013), tweets (Mogilner et al. 2010), customer service calls (Packard et al. 2014), blogs (Arsel and Bean 2013), and news articles (Humphreys 2010; Humphreys and Thompson 2014).

Classification Methods

Bottom up methods include classification and topic modeling. Classification methods of text analysis are based on categorizing documents into different “types” and then further describing what textual elements best predict the likelihood of being a “type.” For example, Tirunillai and Tellis (2012) use classification to train a model to recognize positive versus negative reviews based on star rating. Using a training data set, they use both a Naïve Bayes and a support vector machine (SVM) classifier to find which words predict star rating and then use this information to categorize the entire set of reviews, achieving a precision – meaning their algorithm predicts true positives – 68–85% of the time, depending on the product category. Villarroel Ordenes et al. (2017) further refine measures of sentiment by using both explicit and implicit indicators of emotion to measure sentiment and sentiment strength, also testing their framework on a set of starred reviews from Tripadvisor, Amazon, and Barnes and Noble. Classification models vary in sophistication; accuracy of these approaches varies from 55% to 96% for sentiment, for example (Hutto and Gilbert 2014). In general, considerations for model selection are based on the underlying frequency of occurrence of words that one wants to use to make predictions and the clarity of categories one wants to produce. For instance, SVM classification provides clear, mutually-exclusive categories, while LDA produces probabilistic groupings where it is possible for categories to overlap.

Classification models have been used to study reviews (Tirunillai and Tellis 2012; Van Laer et al. 2017), online forums (Homburg et al. 2015), email (Ludwig et al. 2016), and literary texts (Boyd and Pennebaker 2015b; Plaisant et al. 2006). For example, to measure sentiment of message board posts, Homburg et al. (2015) classify a training dataset of unambiguously positive and negative posts. They then use sentiment as a dependent measure to understand how much firm engagement actually increases positive consumer sentiment, finding that there are diminishing returns to engagement.

Topic Modeling

Topic modeling is an approach that begins by parsing text into discrete words, and then finding recurring patterns in co-occurrence that are statistically unlikely if one assumes that word occurrence is independent. In this way, the analysis identifies categories that may be latently represented by the manifest presence of words, and these word groupings are then labeled to represent meaningful concepts or traits in

the data as one would in factor analysis. For example, in a study of hotel reviews, Mankad et al. (2016) use latent Dirichlet allocation (LDA) to identify five topics that occur in users' TripAdvisor comments, identifying amenities, location, transactions, value, and experience as key topics mentioned by reviewers. Latent semantic analysis (LSA), k-means clustering (Lee and Bradlow 2011), probabilistic latent semantic analysis (PLSA), and LDA (Blei et al. 2003) are all methods for topic modeling, with LDA being the most recent and common analytical methods for topic modeling.

LSA is based on the relatively straightforward process of generating a matrix that represents word occurrence (0 for nonoccurrence and 1 for occurrence) and then generating a vector of similarity that represents either the similarity between *documents* (the dot product of the rows) or the similarity between two or more *words* (the dot product of the columns). These vectors can then be reduced using singular value decomposition (SVD) to represent the "topics" that tend to occur across documents. PLSA is a similar process; topics are treated as word distributions based on probability.

LDA is a hierarchical Bayesian model for determining the mixture of topics present in a given document. Like PLSA, it assumes topics are probabilistic distributions of words, except it uses a Dirichlet prior for estimation, which reduces over-fitting. For LDA, one sets the number of topics prior to running the analysis (other methods such as hierarchical Dirichlet Process do not need this assumption). Using assumptions that there is a certain probability distribution for the choice of topic, and a certain distribution within that for choice of words to represent that topic, LDA produces a final list of topics (as represented by a list of words in that topic) and probabilities that a given topic is in the document. Although most approaches are word or phrase based, Büschken and Allenby (2016) conduct an LDA analysis using sentences as the unit of analysis and find that this produces results more predictive of rating than word-based LDA. A sentence-based model assumes that all words in the sentence are part of the same topic, which is reasonable, given Grice's maxims of relation and manner (Grice 1975). Büschken and Allenby (2016) use this model to identify topics for Italian restaurants and hotels from reviews on Expedia and we8there.com.

LDA has been used in a wide range of applications (Büschken and Allenby 2016; Tirunillai and Tellis 2014). As with dictionary approaches, postmeasurement validation, in this case using a hold-out sample or other predictive technique (e.g., external DV) is highly advisable. Machines will only read literal meaning, and therefore homonyms and other colloquialisms including sarcasm can be problematic, as they are overly general and overly specific words. Further, careful cleaning and preparation of the text can reduce errors, as textual markers can sometimes be added during data collection (e.g., headers, footers, etc.).

Market Research Applications of Text Analysis

This section discusses ways that text analysis has been incorporated into marketing research. Although potentially useful for many types of sources and research questions, text analysis has been particularly fruitful for representing consumer sentiment,

studying word-of-mouth communication, and creating positioning maps from online text, among other uses.

Sentiment Analysis

Many text analytic programs and practitioners claim to measure sentiment, but it is not always clear what goes into this key metric. Before discussing the text analysis of sentiment, it might first to help to discuss what sentiment is and what it is trying to capture. In most marketing contexts, researchers and practitioners are interested in consumer attitude toward a brand, product, or service. Yet attitudes are complex mental structures composed not only of emotion, but also cognitive beliefs and intentions (Fishbein and Ajzen 1972). Further, the importance an attitude for any given product for ultimate purchase and future behavior like loyalty depends to a large degree on context and involvement (Petty and Cacioppo 1979). Further, people may articulate attitudes online that do not fully reflect their underlying attitude, there may be selection bias in the attitudes they choose to articulate, and they may behave differently than the attitudes they espouse. Nonetheless, discourse online, as expressed in sentiment, can reflect some underlying attitude about a brand, product, or service, and importantly can affect the social consensus shared among other consumers. Sentiment has been shown to predict movie sales (Krauss et al. 2008; Mestyán et al. 2013) and stock market returns (Bollen et al. 2011; De Choudhury et al. 2008; Tirunillai and Tellis 2012), although there may be natural biases in nonreporting of null results. Structurally, most approaches seek to classify or measure text as having positive, negative, or sometimes neutral sentiment, and some approaches transform this into net sentiment, subtracting negative words from positive words (e.g., Ludwig et al. 2013; Homburg et al. 2015). Top-down approaches do this using a dictionary or lexicon of words, while bottom-up approaches use some underlying external classification like human coding of a training set or customer ratings to identify the set of words that indicate sentiment.

In addition to valence, sentiment can also have strength and certainty. Previous research has used both explicit, semantic indicators of emotion along with implicit, more pragmatic indicators of emotion such as speech acts (commission, assertion, and direction) to successfully measure strength of sentiment (Villarreal Ordenes et al. 2017). Work has further shown that other types of speech such as demonstratives (Potts and Schwarz 2010) and other pragmatic markers can indicate expressive content, commonly expressed in product reviews (Constant et al. 2009).

Using predeveloped, standardized dictionaries is one of the most reliable ways to measure sentiment across contexts, as these wordlists have been developed and tested on a wide range of textual data, and some have themselves been developed through bottom-up approaches.

VADAR, for example, uses a dictionary with a rule-based approach for measuring sentiment. Specifically, Hutto and Gilbert (2014) use a combination of dictionaries based on previous standardized dictionaries like LIWC and General Inquirer but then also develop five rules that take into account syntax and grammar to measure intensity

as well. Bottom-up approaches to measure sentiment produce accuracies ranging from 55% to 96%, depending on the context (Hutto and Gilbert 2014). For example, Tirunillai and Tellis (2012) use star rating to create a classification system for sentiment, with an accuracy rate of 68–85%.

Studying Word of Mouth Through Text Analysis

The primary use of text analysis in marketing research to date has been to study online word-of-mouth communication. Consumers have always shared product information through interpersonal communication (Arndt 1967), and this communication has been shown to be more effective than commercial messages (Brown and Reingen 1987; see also Godes and Mayzlin 2004; Money et al. 1998). And yet while word-of-mouth communication was previously communicated face to face or over the telephone, it is now visible and archived on social shopping sites (Stephen and Toubia 2010), social media (Humphreys 2015), and third-party review sites and platforms. Product reviews on Amazon, hotel reviews on TripAdvisor, and restaurant reviews on Yelp! have all provided marketing insights to better understand the relationship of ratings to sales and stock price (Moe and Schweidel 2014; Schweidel and Moe 2014; Moe and Trusov 2011). For example, Moe and Trusov (2011) find that positive reviews have a direct effect on sales, but this effect is somewhat short-lived because of downward convergence as people post more ratings (i.e., the social dynamics of posts result in reviews becoming relatively more negative over time). Further, positivity can vary depending on platform (Schweidel and Moe 2014; Villarroel Ordenes et al. 2017).

Word of mouth online can be represented by measuring valence, volume, and variance (Godes and Mayzlin 2004). Volume and variance are relatively compatible with existing modeling measures, as volume can be aggregated and variance can be measured through star ratings or other user input. Valence, while partially captured by star measures, is perhaps best measured by sentiment, which requires text analysis as a method for converting the unstructured data of linguistic description into data that can be incorporated into quantitative models. There is also, it should be noted, a wide range of linguistic properties and semantic content beyond valence that usefully informs marketing research (Humphreys and Wang 2018). For instance, Kovács et al. (2013) show that restaurants have higher ratings if reviewers mention authenticity in their reviews, even when controlling for restaurant quality.

The role of emotion in the spread of word of mouth is one key topic. In a study of sharing news articles, Berger and Milkman (2012) find that positive emotion increases virality, but so too does the presence of intense negative emotion like anger or anxiety in the article. Effects of the sender and speech context have also been investigated through text analysis using pronouns. Using a standard dictionary for first-person personal pronouns (“I,” “me”), Packard and Wooten (2013) find that consumers self-enhance more in word of mouth to signal knowledge about a particular domain. Consumers have also been shown to engage in self-presentation by sharing fewer negative emotions when broadcasting to a large audience versus

narrowcasting to a smaller one (Barasch and Berger 2014). When evaluating a product like a movie, consumers are more likely to use pronouns referring to themselves when expressing views about taste vs. their views about quality (Spiller and Belogolova 2016).

Topic Discovery and Creating Positioning Maps from Online Text

Text analysis can be used to create positioning maps for brands, companies, or products and to visualize market structure based on attributes within a particular category. Bottom-up methods such as LDA, LSA, and similar methods like k-means clustering are used to group words in a text (like reviews) into attributes or brands based on common co-occurrence. For example, to create a visualization of market structure for cameras from a set of reviews on [Epinions.com](#), Lee and Bradlow (2011) first extract phrases related to particular attributes (e.g., battery life, photo quality) and then use k-means clustering to group phrases based on their similarity (calculated as cosine similarity between vectors of words). They then go on to show that this kind of analysis reveals attributes mentioned by and important to consumers, but absent from expert reviews such as size, design, and screen brightness. Similarly, using text data from diabetes forums, Netzer et al. (2012) find several side effects commonly mentioned on the forum, but absent from a site like WebMD (e.g., weight gain, kidney problems).

Topic-based models are compatible with psychological theories such as spreading activation in semantic memory (Collins and Loftus 1975). For instance, based on the idea that people talk about brands together that are related in semantic memory, Netzer et al. (2012) produce a perceptual map for car brands using reviews from [Edmunds.com](#) and compare that to results from perceptual maps based on more typical survey and brand-switching based on sales approaches. In doing so, they find several notable differences between the results based on text analysis versus those based on sales or survey data. For instance, based on the sales data, Korean brands of cars are not associated with the Japanese brands. However, based on the textual data, these brands are grouped together. This suggests that while text analysis can capture cognitive associations, these may not necessarily translate into behavior such as brand switching (Table 1).

Measurement of the Organization and Firm Environment

Finally, text analysis can be used to measure organizational attention through the analysis of shareholder reports, press releases, and other marketing communication. These studies are primarily based on dictionary-based analysis, and often create dictionaries rather than using standardized dictionaries to fit the industry or original context and research question. For example, scholars have developed dictionaries to study the changes in CSR language over time to reveal differences in developing countries (Gandolfo et al. 2016). In an analysis of annual reports, Lee et al. (2004) find that companies that issued internal reasons for negative events had higher stock

Table 1 Types of text analysis

Type of text analysis	Materials	Theoretical areas	Software/methods	Relevant examples
Dictionary-based	Reviews, tweets, online forums, news articles, press releases, annual reports	Sentiment/emotion, psychological mindset (e.g., construal level), brand attention and brand value, legitimacy/corporate image, customer service	LIWC, WordStat, Diction	Humphreys (2010), Berger and Milkman (2012), Packard et al. (2018)
Classification	Reviews, online forums, literary texts, tweets, email	Sentiment, deception, product attributes, market structure	SVM, Naïve Bayes, k-nearest neighbor, neural networks, WordStat	Homburg et al. (2015), Van Laer et al. (2018), Tirunillai and Tellis (2012)
Topic modeling	Product or service reviews, online forums	Product attributes, positioning, market structure, customer needs	LDA, LSA, PLSA, K-means clustering, R, WordStat	Netzer et al. (2012), Lee and Bradlow (2006), Buschken and Allenby (2016)

prices a year after the event, suggesting that organizations who attribute blame to firm-controlled factors appear more in control than those who do not and therefore have more favorable impressions from investors. Interactions between firm employees or agents can also be better understood. For example, Ludwig et al. (2016) develop a method for detecting deception in sales emails. They find that deceivers are more likely to use elaborate, superfluous descriptions, and less self-referencing, quickly taking on the linguistic style of their intralocular.

Firm environment can also be captured through measuring media such as newspapers, magazines, and trade publications. For example, Humphreys (2010) shows that changes in the institutional and cultural environment enabled the legitimization of the casino gambling industry in the United States. Humphreys and Thompson (2014) study the environment of risk perceptions following two crises – the Exxon and BP oils spills – and find that the media narratives serve to contain risk perceptions following these disasters. Ertimur and Coskuner-Balli (Ertimur and Coskuner-Balli 2015) trace how the Yoga industry shifted over time, developing distinct institutional logics that impacted branding and positioning within the industry.

Issues in Working with Textual Data

Although language provides a window into many areas of consumer thought and market strategy, there are several issues to consider when analyzing text. Language rarely, if ever, follows patterns of normal distribution (Zipf 1932). For instance,

functional words like “a,” “he,” and “there” make up about 40% of all language in normal usage. Common words like nouns and verbs make up another 59%, and only a small fraction of those common words will usually be relevant to the research question. Textual data are often left-skewed (lots of zeros), documents often contain different numbers of words, and the words of interest are often too infrequently or too frequently occurring to make meaningful comparisons. For these reasons, after word frequency has been calculated, researchers will often transform the data prior to statistical analysis. Further, many test such as ANOVA would not be appropriate due to the non-normal distribution of the data.

Text is therefore almost always represented as a percentage of words in the document (e.g., Ludwig et al. 2013), and log transformation to account for skewedness is often commonly employed (Netzer et al. 2012), although there are several possible transformations used (Manning et al. 2008). Tf*idf is a measure often used to account for the term frequency, standardized by the overall frequency of a word in the dataset as a whole (see Salton and McGill 1983 for details in calculating tf*idf, with attendant options for transformation).

Traditional methods for measuring co-occurrence such as Pearson correlation can be problematic due to the large number of zeros in a dataset (Netzer et al. 2012). For this reason, researchers will often use cosine similarity or Jaccard distance to compare words and documents. A series of robustness checks using multiple methods to calculate co-occurrence is often necessary to ensure that results do not occur simply due to infrequently or too-frequently occurring words (Monroe et al. 2009; Netzer et al. 2012). For example, if a word like “him” is very common, it is likely to co-occur with more words than an infrequent word like “airbag.” And yet, the word “airbag” may be more diagnostic of the concept safety than a personal pronoun like “him” even though detecting the co-occurrence will be more likely. Because data are not normally distributed, statistical tests such as the Mann-Whitney test, which tests for significance in rankings rather than absolute number, can serve as a replacement for ANOVA.

Extended Example: Word-Of-Mouth Differences Between Experts and Nonexperts to a Product Launch

Purpose

This section presents a sample text analysis as an illustration of top-down, dictionary-based methods according to the six stages (Table 2) (Reprinted from the Web Appendix to Humphreys and Wang (2018), Automated Text Analysis for Consumer Research, *Journal of Consumer Research*, 44(6), 1 (April), 1274–1306, with permission from Oxford University Press.). Automated text analysis is appropriate for tracking systematic trends in language over time and making comparisons between groups of texts. To illustrate a top-down approach to text analysis, this section presents a short study of consumer response to the product launch of an mp3 player/wireless device, the Apple iTouch. This case has been selected because it

Table 2 Stages of automated content analysis

Stages of automated content analysis (dictionary-based analysis)	
<i>Stage</i>	<i>Elements of stage</i>
1. Identify a research question	Select a research topic and a question within that topic
2. Data collection	Identify sources of information Online databases or newspapers Digital converters for printed text Web scraping for internet data Archival materials Field interviews
2a. Data cleaning	Organize the file structure Spell check, if applicable Eliminate problematic characters or words
3. Construct definition	Qualitatively analyze a subsample of the data Create a word list for each concept Have human coders check and refine dictionary Preliminarily implement dictionary to check for false positives and false negatives
4. Operationalization	Conduct computer analysis to compute the raw data Make measurement decisions based on the research question: Percent of all words Percent of words within the time period or category Percent of all coded words Binary (“about” or “not about” a topic)
5. Interpretation and analysis	Make unit of analysis decisions: By article, year, decade Comparison by genre, speaker, etc. Choose the appropriate statistical method for the research question: Analysis of variance (ANOVA) Regression analysis Multidimensional scaling Correlational analysis
6. Validation	Pull a subsample and have coded by a research assistant or researcher Calculate Krippendorff’s alpha or a hit/miss rate

can be used to illustrate both comparison between groups and change over time and because it is relatively agnostic regarding theoretical framework. One could study word-of-mouth communication from a psychological, sociological, anthropological, or marketing strategy point of view (c.f. Godes and Mayzlin 2004; Kozinets 2010; Phelps et al. 2004; Winer 2009).

Stage 1: Develop a Research Question

This study proposes a specific, strategic research question: After a product launch, do experts respond differently from nonexperts? Further, how does word-of-mouth response change in expert versus nonexpert groups as the product diffuses? Word of mouth from experts can be particularly influential in product adoption, so it is

important to know how their views may change over time and in comparison with nonexpert groups. The context chosen for this study, the launch of the Apple iPod, is a good case to study because both the product category and the criteria for evaluating the product were ambiguous at the time of launch.

Stage 2: Data Collection

Data. Data were collected from two websites, Amazon.com and CNET.com. Consumer comments from Amazon were used to reflect a nonexpert or mixed consumer response, while user comments from CNET were used to measure expert response. Amazon is a website that sells everything from books to toys and has a broad audience. CNET, on the other hand, is a website dedicated exclusively to technology and is likely to have posters with greater expertise. Archival data also suggests that there are differences among visitors to the two sites.

According to Quantcast estimates (Quantcast 2010a, CNET Monthly Traffic (Estimated)) (www.quantcast.com/cnet.com), users to CNET.com are predominantly male and likely to visit websites like majorgeeks.com and read PC World. Amazon users, on the other hand, represent a broader demographic. They are more evenly divided between men and women (48/52), are more likely to have kids, and, visit websites like buy.com (Quantcast 2010b, Amazon monthly traffic (estimated)) (www.quantcast.com/amazon.com). Data were collected on November 2009.

Data were collected with the help of a research assistant from Amazon.com and CNET.com from September 5, 2007 to November 6, 2009. Keyword search for “iPod Touch” was used to gather all customer reviews available for the product at the time of analysis. Reviews for multiple versions of the device (first and second generation) were included and segmented in the analysis according to release date. The first-generation iPod Touch was released on September 5, 2007, and the second-generation was released on September 9, 2008.

Data were scraped from the internet, stored in a spreadsheet, and segmented by post. The comment date, poster name, rating, location of the poster, and the text of the comment itself were all stored as separate variables. Two levels of analysis were chosen. The most basic level of analysis is at the comment level. Each comment was coded for its content so that correlations between the content of that post and the date, poster experience, and location could be assessed. The second level of analysis is the group level, between Amazon and CNET. Comparisons can thus be made between expert and nonexpert groups based on the assumption that Amazon posters are nonexperts or a mix of experts and nonexperts, while dedicated members of the CNET community have more expertise. Lastly, because the time variable exists in the dataset, it will also be possible to periodize the data. This may be relevant in assessing the effects of different product launches (e.g., first- vs. second-generation iPods) on the textual content of posts. About 204 posts were collected from Amazon and 269 posts were collected from CNET, yielding a sample size high enough to make statistical comparisons between groups.

After a file structure was created, data were cleaned by running a spell check on all entries. Slang words (e.g., “kinda”) were replaced with their proper counterparts. Text was scanned for problematic words. For example, “touch” appeared with greater frequency than usual because it was used to refer to the product, not to the sense. For that reason, “touch” was replaced with a noncodable character like “TTT” so that it would not be counted in the haptic category used in the standard dictionary.

Stage 3: Construct Definition

Work in information processing suggests that experts process information differently from novices (Alba and Hutchinson 1987). In general, experts view products more cognitively, evaluating product attributes over benefits or uses (Maheswaran and Sternthal 1990; Maheswaran et al. 1996; Sujan 1985). While novices use only stereotypical information, experts use both attribute information and stereotypical cues (Maheswaran 1994). Experts are able to assimilate categorical ambiguity, which means one would expect for them to adjust to an ambiguous product more quickly than nonexperts (Meyers-Levy and Tybout 1989). They also tend to approach judgment in an abstract, higher level construal than nonexperts (Hong and Sternthal 2010).

From previous research, several working hypotheses can be developed. The strategic comparison we wish to make is about how experts versus nonexperts evaluate the product and whether or not this changes over time. First, one might expect that experts would use more cognitive language and that they would more critically evaluate the device.

H1: Experts will use more cognitive language than novices.

Secondly, one would also expect that experts would attend to features of the device, but nonexperts would attend more to uses of the device (Maheswaran et al. 1996). Note that this is based on the necessary assumption that users discuss or verbally elaborate on what draws their mental attention, which is reasonable according to previous research (Carley 1997).

H2: Experts will discuss features more than nonexperts.

H3: Nonexperts will discuss benefits and uses more than experts.

Thirdly, over time, one might predict that experts would be able to assimilate ambiguous product attributes while nonexperts would not. Because experts can more easily process ambiguous category information and because they have a higher construal level, one would predict that they would like this ambiguous product more than novices and would learn to assimilate the ambiguous information. For example, in this case, the capacity of the device makes it hard to categorize (cell phone vs. mp3 player). One would expect that experts would more quickly understand this ambiguity and that over time their elaboration on this feature would decrease.

H4: Experts will talk about ambiguous attributes (e.g., capacity) less over time, while nonexperts will continue to discuss ambiguous attributes. Lastly, previous

research suggests that these differences in focus, experts on features and nonexperts on benefits, would differentially influence product ratings. That is, ratings for non-experts will depend on evaluation of benefits such as entertainment, but expert ratings would be influenced more by features.

H5: Ratings will be driven by benefits for nonexperts.

H6: Ratings will be driven by features by experts.

These are only a few of the many potential hypotheses that could be explored in an analysis of online word-of-mouth communication. One could equally explore the cultural framing of new technologies (Giesler 2008) or the co-production of brand communications by seeding product reviews with bloggers (Kozinets 2010). The question posed here – do experts respond differently to new products than non-experts over time? – is meant to be illustrative of what can be done with automated text analysis rather than a rigorous test of the psychological properties of expertise.

In this illustrative example, the key constructs in examining H1 through H6 are known: expert and nonexperts, cognitive expressions, affect, product features, and benefits. We therefore proceed with a top-down approach. Operationalization for some of the constructs – cognitive and affective language – is available through a standardized measure (LIWC; Pennebaker et al. 2001), and we can therefore use a standardized dictionary for their operationalization. However, some constructs such as features and benefits are context-specific, and a custom dictionary will be necessary for operationalization. In addition, there may be other characteristics that distinguish experts from nonexperts. We will therefore also perform a bottom-up approach of classification.

Stage 4: Operationalization

For this analysis, the standard LIWC dictionary developed by Pennebaker et al. (2001) was used in addition to a custom dictionary. Table 3 presents the categories used from both the standardized and the custom dictionaries. The standard dictionary includes categories for personal pronouns such as “I,” parts of speech such as adjectives, psychometrically pretested categories such as positive and negative emotion, and content-related categories such as leisure, family, and friend-related language.

A custom dictionary was also developed to identify categories specific to the product word-of-mouth data analyzed here. Ten comments from each website were selected and open coded, with the researcher blind to the site from which they came. Then, ten more comments from each website were selected and codes were added until saturation was reached (Weber 2005). In all, the subsample required to develop the custom dictionary was 60 comments, 30 from each website, about 11% of all comments. Fourteen categories were created, each containing six words on average.

The qualitative analysis of comments revealed posters tended to talk about the product in terms of features or aesthetics. Dictionary categories were therefore created for words associated with features (e.g., GPS, camera, hard drive, battery) and for aesthetics (e.g., sharp, clean, sexy, sleek). Posters also had recurring concerns

about the capacity of the device, the cost of the product, and reported problems they experienced using the product. Categories were created for each of these concerns. Because there might be some researcher-driven interest in product uses and because posters frequently mentioned entertainment and work-related uses, categories were created for each type of use. Categories of “big” versus “small” were included because previous theorization in sociology has suggested that the success of the iPod comes from its offerings of excess – large screen, excess capacity, etc. (Sennett 2006). Two categories were created to count when competitive products were mentioned, either within the Apple brand or outside of it.

The dictionary categories were validated by three coders who suggested words for inclusion and exclusion. Percent agreements between coders on each dictionary category can be found in Table 3. Average agreement was 90%. Text files were run

Table 3 Standard and custom dictionaries

Category	Abbv	Words	No. of words	Alpha*
Social processes	Social	Mate, talk, they, child	455	97%
Affective processes	Affect	Happy, cried, abandon	915	97%
Positive emotion	Posemo	Love, nice, sweet	406	97%
Negative emotion	Negemo	Hurt, ugly, nasty	499	97%
Cognitive processes	Cogmech	Cause, know, ought	730	97%
Past tense	Past	Went, ran, had	145	94%
Present tense	Present	Is, does, hear	169	91%
Future tense	Future	Will, gonna	48	75%
Discrepancy	Discrep	Should, would, could	76	80%
Exclusive	Excl	But, without, exclude	17	67%
Perceptual processes	Percept	Observing, heard, feeling	273	96%
Relativity	Relativ	Area, bend, exit, stop	638	98%
Space	Space	Down, in, thin	220	96%
Time	Time	End, until, season	239	94%
Work	Work	Job, majors, xerox	327	91%
Aesthetics	Aesth	Sleek, cool, shiny, perfect	9	83%
Capacity	Cap	Capacity, space, storage	7	93%
Cost	Cost	Price, cost, dollars	6	100%
Big	Big	Large, huge, full	5	83%
Problems	Prob	Bugs, crash, freeze	7	100%
Competitors	Comp	Zune, Microsoft, Archos	4	67%
Apple	Apple	Nano, iPod, iPhone	4	100%
Entertainment	Ent	Music, video, fun	9	85%
Job	Job	Work, commute, conference	9	100%
Connectability	Connect	Wifi, internet, web	9	95%
Features	Feat	GPS, camera, battery	5	87%
Love	Love	Amazing, best, love	7	100%
Small	Small	Empty, small, tiny	4	100%
Expertise	Expert	Jailbreak, jailbroke, keynote	4	67%

*Alpha is the percent agreement of three coders on dictionary words in the category

through the LIWC program, first using the standard dictionary, then using the custom dictionary. A spreadsheet was created from three sets of data: (1) the comment data collected directly from the website (e.g., date of post, rating of product), (2) the computer output from the standard dictionary, and (3) the output from the custom dictionary.

Validation. Once rough findings were gleaned, the coding was validated. Twenty instances from each category were pulled from the dataset and categorized. “Hits” and “false hits” were then calculated. This yielded an average hit rate of 85% and a “false hit” rate of 15%. The least accurate category was aesthetics, with a hit rate of 70% and a false hit rate of 30%. The most accurate category was “small,” which had a hit rate of 95% and a false hit rate of 5%.

Stage 5: Interpretation and Analysis

Overall, the findings indicate that there are systematic differences between the way experts and nonexperts interpret the new device. As with most textual data, there are many potential variables and measures of interest. The standard LIWC dictionary contains 61 categories, and in the dataset studied here, 28 of these categories were significantly different among text from the three websites. We will report some of the most notable differences, including those needed to test the hypotheses.

Comparison between groups. First, we assessed differences among the two groups of comments. This was done by comparing differences in the percent of words coded in each category between groups using the Mann-Whitney test due to the skewed distribution of the data. Tables 4 and 5 show the differences by category. With the standard dictionary, several important differences between the word of mouth of nonexperts and experts can be discerned.

First, experts use more cognitive words ($M_{\text{cog}|CNET} = 16.57$, $M_{\text{cog}|Amazon} = 15.64$, Mann-Whitney $U = 30,562$, $z = 2.12$, $p < 0.05$) than nonexperts, but they also use more affective (both positive and negative) language ($M_{\text{affect}|CNET} = 7.3$ vs. $M_{\text{affect}|Amazon} = 6.53$, $U = 30,581$, $z = 2.14$, $p < 0.05$) as well. The finding that experts evaluate the product cognitively is congruent with previous research (Maheswaran et al. 1996), and the highly affective tone indicates that they are likely more involved in product evaluation (Keltling and Duhacheck 2009). However, CNET posters use more negation ($M_{\text{neg}|CNET} = 2.47$, $M_{\text{neg}|Amazon} = 1.74$, $U = 34,487$, $z = 4.81$, $p < 0.001$). Together with the presence of cognitive language, this indicates that they may be doing more critical evaluation. The first hypothesis was therefore supported.

Secondly, nonexperts focus on distal rather than proximate uses, while experts focus on device-related issues like features. Nonexperts on Amazon use more distal social, time-, family-related language (e.g., $M_{\text{social}|Amazon} = 5.55$ vs. $M_{\text{social}|NET} = 4.23$, $U = 22,259.5$, $z = -3.52$, $p < 0.001$ and $M_{\text{time}|Amazon} = 5.65$, $M_{\text{time}|CNET} = 3.89$, $U = 18,527$, $z = -6.01$, $p < 0.001$). Experts on CNET, on the other hand, focus on features ($M_{\text{features}|CNET} = 0.61$ vs. $M_{\text{features}|Amazon} = 0.41$, $U = 30,012.5$, $z = 2.10$, $p < 0.05$) and capacity ($M_{\text{connect}|CNET} = 1.08$ vs. $M_{\text{connect}|Amazon} = 0.756$, $U = 35,819$, $z = 6.14$, $p < 0.001$), but also on aesthetics

Table 4 Amazon vs. CNET differences in means, standard dictionary

	Amazon	CNET
WC	160.99	149.11
Social***	5.55	4.23
Affect [†]	6.53	7.20
Posemo	5.50	5.94
Negemo	1.10	1.31
Cogmech*	15.64	16.57
Past***	3.58	2.13
Present	8.91	9.22
Future*	0.76	1.01
Certain	1.66	1.87
Excl**	2.68	3.20
Percept***	3.34	4.86
Relativ***	11.26	9.53
Space*	4.06	4.64
Time***	5.65	3.89
Work	2.08	1.92
Achieve	2.24	2.58
Leisure [†]	3.28	3.80

[†]p < 0.10

*p < 0.05

**p < 0.01

***p < 0.001

Table 5 Differences in means, custom dictionary

	Amazon	CNET
Aesthetics***	0.168	0.833
Capacity***	0.538	1.408
Cost*	0.384	0.641
Big**	0.070	0.178
Problems [†]	0.286	0.165
Competitors	0.080	0.104
Apple*	1.461	1.927
Entertainment**	1.377	1.838
Job [†]	0.164	0.087
Connect*	0.756	1.075
Features [†]	0.413	0.606
Love***	0.746	1.470
Small*	0.054	0.135
Expert*	0.009	0.028

[†]p < 0.10

*p < 0.05

**p < 0.01

***p < 0.001

($M_{\text{aesth|CNET}} = 0.833$ vs. $M_{\text{aesth|Amazon}} = 0.168$, $U = 33,518$, $z = 5.02$, $p < 0.001$). Experts discussed aesthetics about eight times more than the mixed group on Amazon. These differences indicate that, in general, experts focus on the device itself while nonexperts focus on uses. This lends convergent evidence to support to H2 and H3.

One other finding not specified by the hypotheses is notable. Nonexperts use more past-oriented language ($M_{\text{past|Amazon}} = 3.58$ vs. $M_{\text{past|CNET}} = 2.13$, $U = 21,289$, $z = -4.20$, $p < 0.001$), while expert posters use more future-oriented language ($M_{\text{future|CNET}} = 1.01$, $M_{\text{future|Amazon}} = 0.76$, $U = 31,446$, $z = 2.83$, $p < 0.01$). This suggests that experts might frame the innovation in the future while nonexperts focus on the past. Recent research suggests experts and novices differ in temporal construal (Hong and Sternthal 2010). Experts focus on the far future while novices focus on the near future. The results here provide convergent evidence that supports previous research and suggests a further hypothesis – that novices focus on past-related information – for future experimental research (Table 6).

In an extended analysis, adding a third group could help the researcher draw more rigorous conclusions through techniques of analytic induction (Mahoney 2003; Mill 1843). That is, if an alternative explanation is possible, the researcher could include a comparison set to rule out the alternative explanation. For example, one might propose that the difference in “cost” discourse is because Amazon.com users make less money than CNET users, on average, and are therefore more concerned about price. One could then include an expert website where the users are known to have a lower income than the posters on Amazon to address this explanation. If the same results are found, this would rule out the alternative hypothesis.

Trends over time. Because the product studied here is an innovation, the change of comments over time as the product diffuses is of interest. Time was analyzed first as a continuous variable in a correlation analysis and then as a discrete variable in ordinary least squares regression analyses, where the release of the first and second generation of iTouch marked each period.

A correlation analysis was used to analyze time as a continuous variable (Table 7). We find that affect increases over time in the expert group, which indicates that group becomes more involved ($r_{\text{(affect, Date|CNET)}} = 0.144$, $p < 0.01$). Experts become less concerned with capacity ($r_{\text{(capacity, Date|CNET)}} = -0.203$ $p < 0.01$) while Amazon users do not change in their concern for capacity. This indicates that experts learn something about the product category: the limited capacity was initially a shock to reviewers, as it was unorthodox for an mp3 player. But, over time, experts learned that this new category segment – mp3 wireless devices – did not offer as much memory. This supports Hypothesis 4 (Fig. 1).

Besides the correlation analysis, we also did ordinary least square linear regression analyses to analyze whether reviewers’ expressions changed over time (Table 8). We created a binary variable, which is set to “1” if the review is posted after the second generation of iTouch is released, and “0” if the review is for the first generation of iTouch. To account for asymmetry in their distributions due to non-normality, we log-transformed the term frequency measurements of affect and capacity, our variables of interest. The results from the OLS analyses are congruent

Table 6 Correlation table, Amazon vs. CNET

Correlations		Statistics = Pearson correlation												
	Site	Rating	Date	Affect	Posemo	Negemo	Aesth	Capacity	Ent	Connect	Feat	Love	Big	Small
Rating	Amazon	1	0.009	0.282 ^a	0.387 ^a	-0.200 ^a	0.061	0.064	0.216 ^a	0.002	0.128	0.273 ^a	0.015	-0.024
	CNET	1	-0.012	0.095	0.319 ^a	-0.433 ^a	0.024	-0.058	0.044	0.145 ^b	-0.118	0.373 ^a	0.091	-0.053
Date	Amazon	0.009	1	-0.087	-0.046	-0.118	-0.082	0.013	0.073	0.008	-0.040	0.022	-0.156 ^b	-0.095
	CNET	-0.012	1	0.144 ^b	0.145 ^b	0.011	-0.009	-0.203 ^a	0.114	0.127 ^b	-0.102	-0.006	-0.106	-0.001
Affect	Amazon	0.282 ^a	-0.087	1	0.910 ^a	0.350 ^a	-0.049	-0.098	-0.043	-0.187 ^a	0.049	0.450 ^a	-0.001	-0.036
	CNET	0.095	0.144 ^b	1	0.865 ^a	0.263 ^a	0.367 ^a	-0.036	0.111	0.036	0.108	0.411 ^a	-0.096	0.034
Posemo	Amazon	0.387 ^a	-0.046	0.910 ^a	1	-0.056	0.005	-0.052	0.032	-0.164 ^b	0.064	0.473 ^a	0.006	-0.015
	CNET	0.319 ^a	0.145 ^b	0.865 ^a	1	-0.253 ^a	0.409 ^a	-0.019	0.156 ^b	0.106	0.104	0.514 ^a	-0.038	-0.056
Negemo	Amazon	-0.200 ^a	-0.118	0.350 ^a	-0.056	1	-0.117	-0.140 ^b	-0.194 ^a	-0.104	-0.030	-0.013	0.026	-0.050
	CNET	-0.433 ^a	0.011	0.263 ^a	-0.253 ^a	1	-0.086	-0.026	-0.087	-0.139 ^b	0.000	-0.205 ^a	-0.119	0.167 ^a
Aesth	Amazon	0.061	-0.082	-0.049	0.005	-0.117	1	0.131	-0.019	0.016	0.005	-0.055	0.126	0.003
	CNET	0.024	-0.009	0.367 ^a	0.409 ^a	-0.086	1	-0.025	0.040	-0.052	0.291 ^a	0.015	-0.072	-0.053
Capacity	Amazon	0.064	0.013	-0.098	-0.052	-0.140 ^b	0.131	1	0.055	0.052	-0.044	-0.010	-0.046	0.144 ^b
	CNET	-0.058	-0.203 ^a	-0.036	-0.019	-0.026	-0.025	1	0.079	-0.177 ^a	-0.079	-0.048	-0.025	0.020
Ent	Amazon	0.216 ^a	0.073	-0.043	0.032	-0.194 ^a	-0.019	0.055	1	0.139 ^b	-0.022	-0.061	0.069	0.063

(continued)

Table 6 (continued)

Correlations		CNET	0.044	0.114	0.111	0.156 ^b	-0.087	0.040	0.079	1	0.023	-0.141 ^b	0.072	0.055	-0.012
Connect	Amazon	0.002	0.008	-0.187 ^a	-0.164 ^b	-0.104	0.016	0.052	0.139 ^b	0.007	1	0.007	-0.055	-0.077	-0.009
	CNET	0.145 ^b	0.127 ^b	0.036	0.106	-0.139 ^b	-0.052	-0.177 ^a	0.023	0.008	1	0.008	0.139 ^b	0.038	-0.056
Feat	Amazon	0.128	-0.040	0.049	0.064	-0.030	0.005	-0.044	-0.079	-0.022	0.007	1	0.000	-0.019	-0.024
	CNET	-0.118	-0.102	0.108	0.104	0.000	0.291 ^a	-0.079	-0.141 ^b	0.008	1	1	-0.086	-0.045	-0.096
Love	Amazon	0.273 ^a	0.022	0.450 ^a	0.473 ^a	-0.013	-0.055	-0.010	-0.061	-0.055	0.000	1	-0.016	-0.016	-0.048
	CNET	0.373 ^a	-0.006	0.411 ^a	0.514 ^a	-0.205 ^a	0.015	-0.048	0.072	0.139 ^b	0.139 ^b	-0.086	1	0.078	0.044
Big	Amazon	0.015	-0.156 ^b	-0.001	0.006	0.026	0.126	-0.046	0.069	0.069	-0.077	-0.019	-0.016	1	0.055
	CNET	0.091	-0.106	-0.096	-0.038	-0.119	-0.072	-0.025	0.055	0.055	0.038	-0.045	0.078	1	0.059
Small	Amazon	-0.024	-0.095	-0.036	-0.015	-0.050	0.003	0.144 ^b	0.063	0.063	-0.009	-0.024	-0.048	0.055	1
	CNET	-0.053	-0.001	0.034	-0.056	0.167 ^a	-0.053	0.020	-0.012	-0.012	-0.056	-0.096	0.044	0.059	1

^aCorrelation is significant at the 0.01 level (2-tailed)

^bCorrelation is significant at the 0.05 level (2-tailed)

Table 7 OLS regression coefficient estimates. Affect and capacity by time and Amazon vs. CNET

Dependent variable		B	Std. error
ln(capacity)	(Intercept)***	0.275	0.058
	Is 2nd Gen	0.024	0.081
	Is CNET***	0.407	0.069
	Is 2nd Gen × CNET***	-0.546	0.158
ln(affect)	(Intercept)***	1.916	0.048
	Is 2nd Gen	-0.043	0.068
	Is CNET	0.063	0.057
	Is 2nd Gen × CNET*	0.275	0.132

p < 0.10
 *p < 0.05
 **p < 0.01
 ***p < 0.001

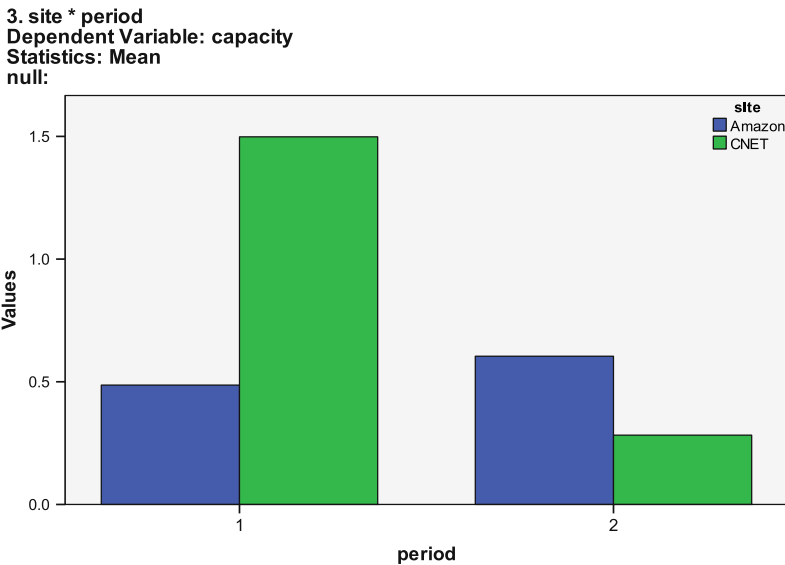


Fig. 1 Mean number of capacity words by site and time period

with the correlation analysis. We observe that in general expert reviewers discussed capacity more than nonexperts ($\hat{\beta} = 0.407$, $p < 0.001$). However, as predicted by Hypothesis 4, such discussions decreased after the release of the second-generation iPod ($\hat{\beta} = -0.546$, $p < 0.001$).

Affect also changes differentially in each group (Fig. 2). The OLS analysis (Table 7) shows that in the first time-period, affective language is roughly equivalent, but experts on CNET use more affective language in the second time-period than they do in the first time-period ($\hat{\beta} = 0.275$, $p < 0.05$). In short, site and period have a positive interactive effect on affective expressions. These are just two examples of

Table 8 Regression coefficients: predictors of product rating for experts vs. nonexperts

Coefficients						
Site	Category	Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. error	Beta		
Amazon	(constant)	3.839	0.137		27.932	0.000
	Aesthetics	0.145	0.175	0.058	0.833	0.406
	Capacity	0.064	0.087	0.051	0.732	0.465
	Problems	-0.015	0.086	-0.012	-0.174	0.862
	Entertainment	0.150	0.047	0.221	3.178	0.002
	Connect	-0.035	0.073	-0.033	-0.476	0.635
	Features	0.174	0.088	0.136	1.972	0.050
CNET	(constant)	3.799	0.144		26.373	0.000
	Aesthetics	0.031	0.031	0.062	0.978	0.329
	Capacity	-0.029	0.042	-0.043	-0.697	0.486
	Problems	-0.290	0.195	-0.091	-1.484	0.139
	Entertainment	0.011	0.040	0.017	0.277	0.782
	Connect	0.100	0.049	0.128	2.062	0.040
	Features	-0.126	0.059	-0.137	-2.138	0.033

3. site * period
Statistics: Mean
null:
Dependent Variable: affect

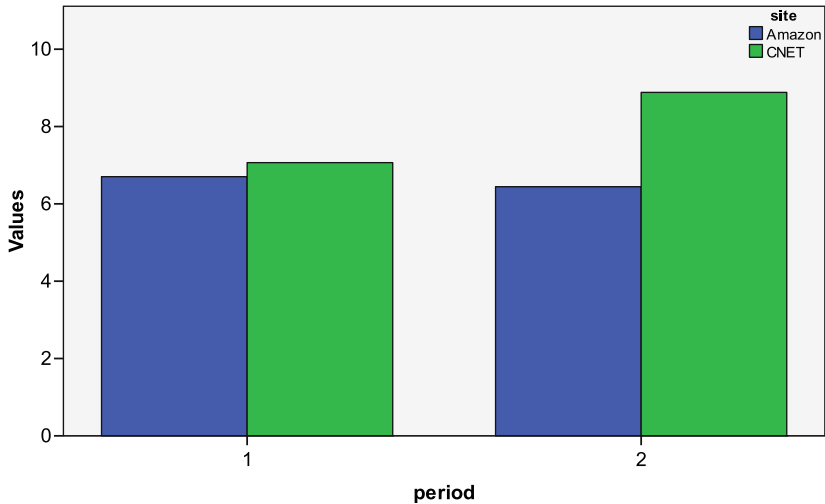


Fig. 2 Mean number of affect words by site and time period

how automated content analysis can be used to assess changes in word-of-mouth communication.

Regression with ratings. Now that relationships between semantic elements in the text have been discerned, their relationship to other, nonsemantic variables is of interest. For example, what factors impact ratings for experts vs. nonexperts? To test the impact of discourse on rating, an OLS regression was run with rating as the dependent variable and the discursive categories as the independent variables. Several discursive variables were significant predictors of ratings overall ($F_{\text{Amazon}} = 2.55, p < 0.05$; $F_{\text{CNET}} = 2.30, p < 0.05$). Results are shown in Table 8. These reveal that the ratings of nonexperts were influenced by entertainment and features, while the ratings of experts were affected by connectability and by the (negative) evaluation of the features. This provides support for H5 and H6. However, they also indicate a more complicated relationship. Features are correlated with both expert and nonexpert ratings. However, for nonexperts, features are positively correlated with ratings while for experts, they are negatively correlated. Problems and cost, although much discussed in the posts, appeared to have little effect on ratings. The unimportance of cost may be explained by the fact that the ratings data are nonbehavioral, that is, most posters had already purchased the device.

Stage 6: Validation

The previous analyses revealed there were systematic differences in the number of words used between experts and nonexperts. To assess construct validity, we used a triangulation approach to explore the relationships between the concepts through a correlation analysis of word association within comment (Table 7). This means that we are looking for how the dictionary categories occur together within one post. To assess construct validity of affect, we included another operationalization of affect, star rating, in the correlational analysis. We calculated Pearson correlations for all categories in the set and compared them with cosine similarities. Both tables produced directionally similar results, and here we report Pearson correlations, as it accounts for both presence and absence of collocation. First, a few expected correlations between categories were checked. For both sites, positive emotion is correlated with rating ($r_{(\text{posem}, \text{rating})} = 0.335, p < 0.01$), as one would expect. Negative emotion is negatively correlated with positive emotion ($r_{(\text{negemo}, \text{posemo})} = -0.348, p < 0.01$). More can be learned, however, by comparing word association in expert versus nonexpert groups.

In general, nonexperts use positive language alongside distal uses for the iPod such as work and family ($r_{(\text{work}, \text{posem}|\text{Amazon})} = 0.243, p < 0.01$ and $r_{(\text{family}, \text{posem}|\text{Amazon})} = 0.190, p < 0.01$). For the non-experts, negative emotion is correlated with problems, as one would expect ($r_{(\text{problems}, \text{negem}|\text{Amazon})} = .470$). For experts, positive emotion occurs alongside aesthetics ($r_{(\text{aesth}, \text{posem}|\text{CNET})} = 0.409, p < 0.01$). For experts, there is also a positive correlation between Apple and love ($r_{(\text{Apple}, \text{love}|\text{CNET})} = 0.312, p < 0.01$).

$r_{\text{CNET}} = 0.203$, $p < 0.01$) that does not exist for nonexperts. These correlations indicate that aesthetics are viewed positively by experts and that they are involved with not only the device but the brand as well. Cosine similarities produce directionally similar results.

Secondly, features are interpreted differentially between the two groups. Novices interpret some features using standards of other categories (like an mp3 player), while experts are more willing to judge them relative to the standards for a new category. For example, from the correlation between small and capacity among the nonexpert group ($r_{(\text{capacity,small}|\text{Amazon})} = 0.144$, $p < 0.01$), one can conclude that posters feel the capacity is too small. No such correlation exists for experts. This could be because the *iTouch* is a product without a known category. Experts can interpret size for this ambiguous product, but novices are uncertain about what capacity is appropriate for the device. These are just a few of the findings that can be gleaned using a correlation Table. A full spatial analysis might compare the network of meanings in the Amazon group to the network of meanings in the CNET group.

For the binary logistic classification, k -fold cross-validation was performed, and per convention, we set $k = 10$. The resulting comparisons between predicted values based on our model and the real values show that overall the model is 80.13% accurate (95% accuracy confidence interval = [0.7624, 0.8363]). Table 9 shows the confusion matrix.

In sum, the automated text analysis presented here shows that that experts evaluate new products in a systematically different way from nonexperts. Using comparison between groups, we show that experts evaluate products by focusing on features while nonexperts focus on the uses and benefits of the devices. Using correlation analysis, we find that experts associate aesthetics with positive emotion while nonexperts associate positive emotion with uses of the device and negative emotion with problems. Further, the correlation analysis provides some validation for the method of automated content analysis by demonstrating the correlation between positive emotion and ratings, a variable used in previous studies of online word-of-mouth communication (Godes and Mayzlin 2004, 2009). We find that, over time, experts focus less on problematic features like capacity and speak more affectively about the product. A regression analysis of the elements of discourse on ratings demonstrates that ratings for experts are driven by features, while ratings by nonexperts are better predicted by both features and the amount of talk about entertainment, a benefit. Note that, like field research, these findings make sense in convergence with previous findings from experimental data and provide ecological validity to previous findings obtained in laboratory settings. These are not meant to be a rigorous test of expertise,

Table 9 Confusion matrix from tenfold cross-validation. Accuracy = 0.8013. p -Value [accuracy > no information rate] = $< 2e-16$

Prediction			
		Expert	Not expert
Expert	Expert	237	62
	Not expert	32	142

but rather an illustration of the way in which text analysis can provide convergent evidence that is meaningful to consumer researchers.

Conclusion and Future Directions

Developments in text analysis have opened a large and fascinating arena for marketing research. Theoretically, marketing research can now incorporate linguistic theory to understand consumer attitudes, interaction, and culture (Humphreys and Wang 2018). While most approaches have focused on analyzing word frequencies, a vast world of looking at text structure at higher, conversational levels remain open. For example, understanding where a word like “great” falls within the text itself (early, middle, or late in a sentence or paragraph) may shed light on the importance of the word in predicting, for example, consumer sentiment. Drawing inferences on the sentence or paragraph level may yield more meaningful results in some contexts (Büschken and Allenby 2016). Lastly, pragmatics, the area of linguistic research aimed at understanding the effect of context on word meaning may help marketing researchers capture more about the nature of consumer communication online.

Practically, incorporating this kind of data allows researchers and managers to integrate the abundance of textual data with existing and growing datasets of behavioral data collected online or through devices. And yet one must be aware of the many limitations of using machines to interpret a human language that has developed socially in face-to-face contexts over 100,000 years. Text analysis can often be used to gather information about top-line patterns of attention or relatively wrote patterns of interaction, but capturing the subtly of human communication remains allusive to machines. Further, due to the ambiguity of language, careful and transparent analysis and interpretation are required at each step of text analysis, from cleaning textual markers that may be misleading to correctly interpreting correlations and differences. Despite these challenges, marketing researchers have clearly shown the theoretical, practical, and managerial insight that can be distilled through the seemingly simple process of counting words.

Cross-References

- ▶ [Return on Media Models](#)
- ▶ [Social Network Analysis](#)

References

- Alba, J. W., & Hutchinson, J. W. (1987). Dimensions of consumer expertise. *Journal of Consumer Research*, 13(4), 411–454.
- Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. *Journal of Marketing Research*, 4, 291–295.

- Arsel, Z., & Bean, J. (2013). Taste regimes and market-mediated practice. *Journal of Consumer Research*, 39(5), 899–917.
- Arvidsson, A., & Caliandro, A. (2016). Brand public. *Journal of Consumer Research*, 42(5), 727–748.
- Barasch, A., & Berger, J. (2014). Broadcasting and narrowcasting: How audience size affects what people share. *Journal of Marketing Research*, 51(3), 286–299.
- Belk, R. W., & Pollay, R. W. (1985). Images of ourselves: The good life in twentieth century advertising. *Journal of Consumer Research*, 11(4), 887.
- Berelson, B. (1971). *Content analysis in communication research*. New York: Hafner.
- Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2), 192–205.
- Blei, David M., Andrew Y. Ng, & Michael I. Jordan. (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3, 993–1022.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computer Science*, 2(1), 1–8.
- Boyd, R. L., & Pennebaker, J. W. (2015a). Away with words. In *Consumer psychology in a social media world* (p. 222). Abingdon: Routledge.
- Boyd, R. L., & Pennebaker, J. W. (2015b). Did Shakespeare write double falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science*, 26(5), 570–582.
- Brown, J. J., & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3), 350–362.
- Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6), 953–975.
- Carley, K. (1997). Network text analysis: The network position of concepts. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah: Lawrence Erlbaum.
- Chung, C. K., & Pennebaker, J. W. (2013). Counting little words in Big Data. In *Social cognition and communication* (p. 25). New York: Psychology Press.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407.
- Constant, N., Davis, C., Potts, C., & Schwarz, F. (2009). The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung*, 33(1–2), 5–21.
- De Choudhury M., Sundaram H., John A., & Seligmann D. D. (2008). Can blog communication dynamics be correlated with stock market activity? In *Proceedings of the nineteenth ACM conference on hypertext and hypermedia*, ACM, pp. 55–60
- Duhachek, Adam, and Katie Kelting. (2009). Coping repertoire: Integrating a new conceptualization of coping with transactional theory. *Journal of Consumer Psychology* 19(3), 473–485.
- Dunphy, D. M., Bullard, C.G., & Crossing, E.E.M. (1974). Validation of the general inquirer Harvard Iv Dictionary. Paper presented at the 1974 Pisa conference on content analysis, Pisa, Italy.
- Ertimur, B., & Coskuner-Balli, G. (2015). Navigating the institutional logics of markets: Implications for strategic brand management. *Journal of Marketing*, 79(2), 40–61.
- Fishbein, M., & Ajzen, I. (1972). Attitudes and opinions. *Annual Review of Psychology*, 23(1), 487–544.
- Fiss, P. C., & Hirsch, P. M. (2005). The discourse of globalization: Framing and sensemaking of an emerging concept. *American Sociological Review*, 70(1), 24p.
- Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *The American Journal of Sociology*, 95(1), 1–37.
- Gandolfo, A., Tuan, A., Corciolani, M., & Dalli, D. (2016). What do emerging economy firms actually disclose in their CSR reports? A longitudinal analysis. In *CSR-HR Project (Corporate Social Responsibility and Human Rights Project)*. Research Grant of University of Pisa (PRA_2015_0082).

- Garrett, D. E. (1987). The effectiveness of marketing policy boycotts: Environmental opposition to marketing. *Journal of Marketing*, 51(2), 46–57.
- Giesler, M. (2008). Conflict and compromise: drama in marketplace evolution. *Journal of Consumer Research*, 34(6), 739–753.
- Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, 23(4), 545–560.
- Godes, D., & Mayzlin, D. (2009). Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Science*, 28(4), 721–739.
- Grayson, K., & Rust, R. (2001). Interrater reliability assessment in content analysis. *Journal of Consumer Psychology*, 10(1/2), 71–73.
- Grice, H. P. (1975). *Logic and Conversation*. Syntax and Semantics, vol.3 edited by P. Cole and J. Morgan, Academic Press. Reprinted as ch.2 of Grice 1989, 22–40.
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297.
- Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, 52(5), 629–641.
- Hong, J., & Sternthal, B. (2010). The effects of consumer prior knowledge and processing strategies on judgments. *Journal of Marketing Research*, 47(2), 301–311.
- Humphreys, A. (2010). Megamarketing: The creation of markets as a social process. *Journal of Marketing*, 74(2), 1–19.
- Humphreys, A., & Latour, K. A. (2013). Framing the game: Assessing the impact of cultural representations on consumer perceptions of legitimacy. *Journal of Consumer Research*, 40(4), 773–795.
- Humphreys, A., & Thompson, C. J. (2014). Branding disaster: Reestablishing trust through the ideological containment of systemic risk anxieties. *Journal of Consumer Research*, 41(4), 877–910.
- Humphreys, A. (2015). *Social media: Enduring principles*. New York/Oxford: Oxford University Press.
- Humphreys, A., & Wang, R. J.-H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research*, 44(6), 1274–1306. <https://doi.org/10.1093/jcr/ucx104>
- Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Jia, L., Clement, Y., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM conference on information and knowledge management*: ACM, pp. 1827–1830.
- Jurafsky, D., Chahuneau, V., Routledge, B. R., & Smith, N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4). <https://doi.org/10.5210/fm.v19i4.4944>.
- Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research*, 4(1), 8–19.
- Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research*, 18(2), 243–250.
- Kovács, B., Carroll, G. R., & Lehman, D. W. (2013). Authenticity and consumer value ratings: Empirical tests from the restaurant domain. *Organization Science*, 25(2), 458–478.
- Kozinets, R. V. (2010). Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of Marketing*, 74(2), 71–89.
- Kranz, P. (1970). Content analysis by word group. *Journal of Marketing Research*, 7(3), 377–380.
- Krauss, J., Nann, S., Simon, D., Gloor, P. A., & Fischbach, K. (2008). Predicting movie success and academy awards through sentiment and social network analysis. In *ECIS*, pp. 2026–2037.
- Lasswell, H. D., & Leites, N. (1949). *Language of politics; studies in quantitative semantics*. New York: G. W. Stewart.

- Lee, F., Peterson, C., & Tiedens, L. Z. (2004). Mea culpa: Predicting stock prices from organizational attributions. *Personality and Social Psychology Bulletin*, *30*(12), 1636–1649.
- Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881–894.
- Ludwig, S., Ko, d. R., Friedman, M., Brüggem, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing*, *77*(1), 87–103.
- Ludwig, S., Van Laer, T., De Ruyter, K., & Friedman, M. (2016). Untangling a web of lies: Exploring automated detection of deception in computer-mediated communication. *Journal of Management Information Systems*, *33*(2), 511–541.
- Maheswaran, D., & Sternthal, B. (1990). The effects of knowledge, motivation, and type of message on ad processing and product judgments. *Journal of Consumer Research*, *17*(1), 66–73.
- Maheswaran, D. (1994). Country of origin as a stereotype: Effects of consumer expertise and attribute strength on product evaluations. *Journal of Consumer Research*, *21*(2), 354–365.
- Maheswaran, D., Sternthal, B., & Gurhan, Z. (1996). Acquisition and impact of consumer expertise. *Journal of Consumer Psychology*, *5*(2), 115.
- Mahoney, J. (2003). Strategies of causal assessment in comparative historical analysis. In J. Mahoney & D. Rueschemeyer (Eds.), *Comparative historical analysis in the social sciences*. Cambridge, UK/New York: Cambridge University Press. pp. xix, 444.
- Mankad, S., Han, H. S., Goh, J., & Gavirneni, S. (2016). Understanding online hotel reviews through automated text analysis. *Service Science*, *8*(2), 124–138.
- Mehl, M. R., & Gill, A. J. (2008). Automatic text analysis. In S. D. G. J. A. Johnson (Ed.), *Advanced methods for behavioral research on the internet*. Washington, DC: American Psychological Association.
- Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PLoS One*, *8*(8), e71226.
- Meyers-Levy, J., & Tybout, A. M. (1989). Schema congruity as a basis for product evaluation. *Journal of Consumer Research*, *16*(1), 39–54.
- Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and methods of scientific investigation*. London: J.W. Parker.
- Moe, Wendy W., and Michael Trusov. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research* *48*(3), 444–456.
- Moe, W. W., & Schweidel, D. A. (2014). *Social media intelligence*. Cambridge, UK/New York: Cambridge University Press.
- Mogilner, C., Kamvar, S. D., & Aaker, J. (2010). The shifting meaning of happiness. *Social Psychological and Personality Science*, *2*(4), 395–402.
- Money, R. B., Gilly, M. C., & Graham, J. L. (1998). Explorations of national culture and word-of-mouth referral behavior in the purchase of industrial services in the United States and Japan. *Journal of Marketing*, *62*, 76–87.
- Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2009). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403.
- Moore, S. G. (2015). Attitude predictability and helpfulness in online reviews: The role of explained actions and reactions. *Journal of Consumer Research*, *42*(1), 30–44.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.
- Opoku, R., Abratt, R., & Pitt, L. (2006). Communicating brand personality: Are the websites doing the talking for the top South African business schools? *Journal of Brand Management*, *14*(1–2), 20–39.
- Packard, G., Moore, S. G., & McFerran, B. (2014). How can “I” help “you”? The impact of personal pronoun use in customer-firm agent interactions. MSI report, pp. 14–110.
- Packard, G. M., & Wooten, D. B. (2013). Compensatory knowledge signaling in consumer word-of-mouth. *Journal of Consumer Psychology* *23*(4), 434–450.

- Palmquist, M. E., Carley, K., & Dale, T. (2009). Analyzing maps of literary and non-literary texts. In K. Krippendorff & M. A. Bock (Eds.), *The content analysis reader* (pp. 4120–4415). Thousand Oaks: Sage.
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: Liwc 2001* (Vol. 71). Mahway: Lawrence Erlbaum Associates.
- Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 17p.
- Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, 37(10), 1915.
- Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research*, 44(4), 333–348.
- Plaisant, C., Rose, J., Bei, Y., Auvil, L., Kirschenbaum, M. G., Smith, M. N., Clement T., & Lord G. (2006). Exploring erotics in emily Dickinson's correspondence with text mining and visual interfaces. In *Proceedings of the 6th ACM/IEEE-CS joint conference on digital libraries*, ACM, pp. 141–150.
- Potts, C., & Schwarz, F. (2010). Affective 'this'. *Linguistic Issues in Language Technology*, 3(5), 1–30.
- Quantcast. (2010a) Cnet monthly traffic (estimated). (www.quantcast.com/cnet.com).
- Quantcast. (2010b). Amazon monthly traffic (estimated). (www.quantcast.com/amazon.com).
- Rayson, P. (2009). Wmatrix: A web-based corpus processing environment. Edited by C. Department, Lancaster University, UK.
- Salton, Gerard, and Michael J. McGill. (1983). Introduction to modern information retrieval McGraw-Hill. New York.
- Schweidel, D. A., & Moe, W. W. (2014). Listening in on social media: A joint model of sentiment and venue format choice. *Journal of Marketing Research*, 51(4), 387–402.
- Sennett, R. (2006). *The culture of the new capitalism*. New Haven: Yale University Press.
- Sneffjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological Science*, 26(9), 1449–1460.
- Spiller, S. A., & Belogolova, L. (2016). On consumer beliefs about quality and taste. *Journal of Consumer Research*, 43(6), 970–991.
- Stephen, A. T., & Toubia, O. (2010). Deriving value from social commerce networks. *Journal of Marketing Research*, 47(2), 215–228.
- Stevenson, T. H., & Swayne, L. E. (1999). The portrayal of African-Americans in business-to-business direct mail: A benchmark study. *Journal of Advertising*, 28(3), 25–35.
- Stone, P. J. (1966). *The general inquirer; a computer approach to content analysis*. Cambridge: MIT Press.
- Sujan, M. (1985). Consumer knowledge: Effects on evaluation strategies mediating consumer judgments. *Journal of Consumer Research*, 12(1), 31–46.
- Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31(2), 198–215.
- Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, 51(4), 463–479.
- Van de Rijt, A., Shor, E., Ward, C., & Skiena, S. (2013). Only 15 minutes? The social stratification of fame in printed media. *American Sociological Review*, 78(2), 266–289.
- Van Laer, T., Escalas J. E., Ludwig S., & Van den Hende E. A. (2017). What happens in Vegas stays on TripAdvisor? Computerized text analysis of narrativity in online consumer reviews.
- Ordenes, V., Francisco, S. L., Ko, D. R., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. *Journal of Consumer Research*, 43(6), 875–894.
- Weber, K. (2005). A toolkit for analyzing corporate cultural toolkits. *Poetics*, 33(3/4), 26p.

-
- Weber, M. (1924). Towards a sociology of the press. Paper presented at the first congress of sociologists, Frankfurt.
- Winer, R. S. (2009). New communications approaches in marketing: Issues and research directions. *Journal of Interactive Marketing*, 23(2), 108–117. <https://doi.org/10.1016/j.intmar.2009.02.004>.
- Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.



Image Analytics in Marketing

Daria Dzyabura, Siham El Kihal, and Renana Peres

Contents

Introduction	666
Top Research Questions by Area Using Image Analytics	669
Product Design	670
Advertising	671
Branding	672
Online Shopping Experience	674
Consumer Perspective	675
Data: Consumer Vs. Firm Images	675
Consumer-Generated Images	676
Firm-Generated Images	679
Methods	681
Feature Extraction	681
Model Training	683
Model Evaluation and Validation	684
Model Application	685
Integrating It All Together	685
Conclusion	688
Cross-References	689
References	689

D. Dzyabura

New Economic School and Moscow School of Management SKOLKOVO, Moscow, Russia
e-mail: ddzyabura@nes.ru

S. El Kihal (✉)

Frankfurt School of Finance and Management, Frankfurt, Germany
e-mail: s.elkihal@fs.de

R. Peres

School of Business Administration, Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: renana.peres@mail.huji.ac.il

Abstract

Recent technical advances and the rise of digital platforms enhanced consumers' abilities to take and share images and led to a tremendous increase in the importance of visual communication. The abundance of visual data, together with the development of image processing tools and advanced modeling techniques, provides unique opportunities for marketing researchers, in both academia and practice, to study the relationship between consumers and firms in depth and to generate insights which can be generalized across a variety of people and contexts.

However, with the opportunity come challenges. Specifically, researchers interested in using image analytics for marketing are faced with a triple challenge: (1) To which type of research questions can image analytics add insights that cannot be obtained otherwise? (2) Which visual data should be used to answer the research questions, and (3) which method is the right one?

In this chapter, the authors provide a guidance on how to formulate a worthy research question, select the appropriate data source, and apply the right method of analysis. They first identify five relevant areas in marketing that would benefit greatly from image analytics. They then discuss different types of visual data and explain their merits and drawbacks. Finally, they describe methodological approaches to analyzing visual data and discuss issues such as feature extraction, model training, evaluation, and validation as well as application to a marketing problem.

Keywords

Image analytics · Visual information · Image processing · Image tagging · Firm images · Consumer images · Feature extraction · Deep neural networks · High-level features · Low-level features · Human-coded features · Color histograms · Gabor filters

Introduction

“The drawing shows me at one glance what might be spread over ten pages in a book.”

Ivan S. Turgenev, Fathers and Sons, 1862.

In the past two decades, images have been playing an increasing role in the marketing arena. Social media outlets have become more image rich, new versions of mobile phones have enhanced ability to take, store, and share photos, and storage and communication infrastructures have become more accessible. These processes have immensely increased the significance of images in consumer life in general, and in marketing in particular.

Images have always been an important part of firms' marketing efforts. Visuals convey a sense of proximity and closeness, and thus, are able to represent objects better than words (Amit et al. 2009). Relative to text, visual information was found to

be better processed and better remembered by humans (MacInnis and Price 1987). Therefore, many of the components of product design, packaging, brand elements, advertising, and design of shopping outlets, use visuals.

These visuals impact consumer response and purchase. For example, Raghurir and Greenleaf (2006) found that certain geometrical ratio of rectangular packaging and print ads influence consumers' relative preferences and purchase intentions. Meyers-Levy and Zhu (2008) showed that various visual elements of store design, such as architecture, freestanding in-store structures, display surfaces, type and arrangement of display cases, mirror orientation, and artwork, relate to consumers' choice and shopping behaviors.

In marketing communications, images are dominant in brands' paid media, such as print, outdoor, and television advertising. Over time, the proportion of pictorial content in a typical magazine ad has grown, while the proportion of text has been gradually shrinking (McQuarrie 2008). A higher proportion of pictorial content in an ad is more efficient in attracting attention (Pieters and Wedel 2004), and is associated with more positive attitude toward the ad (Radach et al. 2003) and with a better memory of the advertised brand (Wedel and Pieters 2000). The layout of the ad, and the size and color combination of its elements have a strong impact on consumers' perceptions and attitudes toward the ad as well as the brand (Janiszewski 1998; Wedel and Pieters 2014; Cho et al. 2008). For example, studies have found that ads containing colors with lighter shade (high "value" in color theory terms) lead to greater liking of the ad. This effect is mediated by stronger feelings of relaxation induced by higher value colors. Higher levels of chroma (a color dimension that relates to the intensity of the color) induce feelings of excitement, which in turn also increase the likeability of the ad (Gorn et al. 1997). Finally, upward looking angles are aligned with perception of potency while photos taken from a downward looking angle were found to lead to a more detailed recall of the brand (Peracchio and Meyers-Levy 2005).

In addition to product design and marketing communications, images are prominently used in marketing research. Consumers express their thoughts, perceptions, and emotions through images. Images have been demonstrated to successfully disrupt people's well-rehearsed narratives and reflect authentic thoughts and deep metaphors. For this reason, visual research methods often better reflect emotions, cultural practices, and attitudes compared to verbal methods (Reavey 2012). Qualitative visual methods are used to arrange brand associations on a map (John et al. 2006), to create brand collages (Zaltman and Coulter 1995; Zaltman and Zaltman 2008), and to elicit brand associations. Other studies used lab experiments (Peracchio and Meyers-Levy 2005) or user-generated digital content (Liu et al. 2017; Klostermann et al. 2018; Pavlov and Mizik 2019; Dzyabura and Peres 2021) to create visual representation of brand associations and connect them with brand characteristics.

The development of digital platforms has further increased the role of images in consumers' lives. Pictures became an important part of brands' owned media – websites, apps, and social media outlets, as well as of brands' earned media – that is, brand content posted by users on social media. Industry reports indicate that 74% of

the content generated by firms contains some form of visual elements, including photos, illustrations, videos, and data visualization (Venngage 2020). Much of the visual activity happens through social media outlets: every minute 136,000 photos are uploaded to Facebook. Every day five Million photos are uploaded to Instagram, added to its corpus of 50 billion photos, which are viewed and receive 3.5 billion likes per day by Instagram's one billion users (ibid.). Consumers use images to communicate with each other and share their experiences, feelings, and impressions. Brands use visual data to learn about consumers' needs and perceptions, to create and communicate value, to shape consumers' attitudes and drive them into action. This ongoing activity has created a rich, dynamic, and vibrant visual ecosystem, which provides a fertile ground for marketing research and marketing activity.

The abundance of visual data, together with the development of image processing tools and advanced modeling techniques, provides unique opportunities for marketing researchers, in both academia and practice, to study the relationship between consumers and firms in depth and to generate insights which can be generalized across a variety of people and contexts.

However, with the opportunity come challenges. Specifically, researchers interested in using image analytics for marketing are faced with a triple challenge. First is the formulation of the research question. Since working with visuals requires elaborate data collection and elaborate analysis, one should identify a research question to which image analytics can add insights that are difficult to obtain in other, more conventional ways.

The second challenge is the choice of data. Visual data sources include user-generated content on brands' web pages (e.g., comments on the brand's Facebook page), data from consumer interactions with other consumers (e.g., one's own Instagram), firm-generated content, general photo repositories (e.g., Flickr), visual product presentation in shopping outlets (eBay, Airbnb), or directly elicited visuals (e.g., online collages). Each of these data sources has its own merits as well as limitations and sometimes, once the research question has been identified, the right data source needs to be carefully chosen. Sometimes, none of the existing data sources contains all the information of interest and the researcher must find ways to combine several sources or supplement the dataset with additional data collection.

The third challenge is the choice of method – most methods used in image analytics were developed in engineering and computer science and were not necessarily optimized for marketing questions. Therefore, using image analytics in marketing requires tailoring existing methods to better fit the data and the research question, or developing new methods altogether.

These three challenges are not independent of each other – the data and method need to be congruent with the research question. For example, a research which seeks to elicit brand associations (Dzyabura and Peres 2021) will need to use interpretable features (such as objects in the pictures), rather than low-level image patterns, and consequently can use image tagging methods and tag-based classifiers to extract high-level features. On the other hand, forecasting the success of a brand based on consumer reviews (Zhang and Luo 2019) has more freedom in choosing the features, but requires showing that pictorial content contains information which cannot be retrieved by a straightforward sentiment or content analysis of the review

text. Finding the right combination of research question, data source, and method is the key to producing meaningful image analytics research in marketing.

Our goal in this chapter is to provide guidance as to how to approach this triple challenge: formulate a worthy research question, select the appropriate data source, and apply the right method of analysis. We start by identifying research questions in five areas that would greatly benefit from using image analytics. We then discuss the different types of visual data including firm-generated and user-generated data and explain their merits and limitations. While data sources constantly change, we suggest guidelines for their characterization and evaluation. We further describe the methodological approaches to analyzing visual data, discussing issues such as feature extraction, model training, classification, and deep learning. We conclude with a decision matrix which can be used as a tool to assist in matching the data and method to the problem at hand. To provide the novice researcher with a gateway to start implementing the ideas, we provide a hands-on tutorial (available on https://github.com/dariasil/image_tutorial), which contains code implementation of several fundamental image analytics tasks and explanations on the required software tools and libraries. We hope that the set of research questions, data sources, richness of methods, code and examples, and guidelines as to how to bring them all together, will help marketing researchers to maximize the tremendous potential of image analytics methods in order to expand the understanding of important research problems and gain meaningful, valuable insights for the benefit of the field.

Top Research Questions by Area Using Image Analytics

The rapid evolution of the visual ecosystem has created unprecedented opportunities to obtain new perspectives on enduring marketing questions. At the same time, it also evoked a large number of new managerial decisions and consumer behaviors which need to be studied. We are just beginning to scratch the surface of this fascinating realm. We outline below the five major areas in marketing that have been most affected by this ecosystem and offer, within each of them, a set of research questions that could lead the further research using image analytics. These questions are summarized in Fig. 1.



Fig. 1 Summary of future research questions for image analytics in marketing

Product Design

In many product categories, design is a dominant factor in consumer choices (Bloch 1995; Rubera 2015). Firms use product design and aesthetics to differentiate themselves (Crilly et al. 2004) and to strategically position their brand among competitors (Keller 2003).

Research has demonstrated how specific elements of product and package design impact consumer perception (Greenleaf and Raghurir 2008). Studies of product aesthetics are mostly focused on one or several specific visual aspects of a design, such as characteristic lines, silhouettes, ornamentation, color, or texture (Orsborn et al. 2009; Eisenman et al. 2016; Chan et al. 2018). Image analytics, on the other hand, allows taking a more holistic view and study the joint, synergetic effect of the overall product design to customer decision or product performance. It also makes it possible to automatically compare a large number of designs and derive quantitative insights and predictions.

Characterizing Designs: How Can Designs Be Characterized Above and Beyond Their Specific Visual Elements?

Images can be used to classify and characterize product designs without the need to break them down into specific predefined visual elements. Such classification can help to:

1. Measure similarity and differences between designs. Specifically, quantify the distance of a focal design from the “average” design, to evaluate how unique the focal design is. This distance could be used to construct a metric measuring design differentiation and design innovativeness.
2. Map designs to brand perceptual dimensions. For example, whether a car design looks family friendly, or a shoe looks rugged, or a sofa looks modern.
3. Match the product design to the customer’s personal style. Such matching can be used to identify and assemble the products to recommend to customers (see [stitchfix.com](https://www.stitchfix.com)).
4. Creating new designs. Models of image analytics can be used to augment the creative process of product design by suggesting novel and unexpected combinations of existing design elements. Algorithms of generative adversarial networks (GAN) that use computer vision to assist in the process of product design. Burnap et al. (2019) demonstrate how image analytic algorithms of GAN to generate models of cars for the design team to consider.

Quantifying the Value of Designs: How Can We Assess and Predict Consumer Attitudes Toward Various Product Designs?

Traditionally, demand models are based on quantifiable product attributes (e.g., miles per gallon, battery life, screen size, brand name, safety rating, price). The design of a product, that is, its overall appearance, vibes, emotion, and symbolism, is hard to be decomposed into quantifiable product attributes, yet they are critical

factors in consumer choice. When a model is estimated using only these traditional functional attributes, these design characteristics end up in the error term.

Image analytics can improve the accuracy of such models by incorporating product images alongside the traditional characteristics in the demand model. For example, they can incorporate information about the success of previous designs in order to forecast various aspects of demand such as product liking, purchase in different channels, product returns or word-of-mouth. Combining image analytics with traditional models requires the development of new models and estimation methods. Specifically, the challenge is to retain the interpretability and un-biasness of some of the traditional coefficients such as price.

Advertising

Image analytics opens new possibilities for taking a systematic, quantitative approach to selecting, adjusting, and optimizing the visual composition of print and video advertisements.

Assessing Ad Creativity: How Can Print and Video Advertisements Be Rated According to Their Level of Creativity? What Are the Combinations of Visual Elements That Make an Ad Perceived as Creative?

Creativity is an important property of advertising messages that is associated with ad recall and effectiveness (Ang et al. 2007). It is sometimes defined as being composed of two factors: divergence and relevance. Divergence is the originality of the ad, and relevance is the extent to which at least some ad or brand elements are meaningful or valuable to the consumer (Smith et al. 2007).

Identifying the visual qualities which construct creativity is challenging. Therefore, creativity of visual ads is typically evaluated by human judges, using research tools such as surveys (Yang and Smith 2009; Sheinin et al. 2011), creativity awards (Lehnert et al. 2014), or crowdsourcing platforms (Kireyev et al. 2020).

Image analytics can allow for automated and scalable assessment of ad creativity. This can be done, for example, by comparing the focal ads to award winning ads, or to a corpus of candidate ads. Toubia and Netzer (2017) developed a prototypicality-based measure of text creativity. Based on their approach, a similar measure can also be constructed for images, either over predefined attributes or as a self-emergent arrangement of the visual space.

Linking Visuals to Emotional and Cognitive Effects: What Visuals Should Be Included in an Ad in Order to Achieve a Desired Outcome? What Objects, Colors, Shades, or Visual Structures Can Be Used to Spark Laughter, Fear, Urgency, Attention, Long-Term Recall, or Other Effects?

Image analytics could address these questions by taking large repositories of photos and their corresponding consumer reactions, and identifying images with certain emotional and cognitive effects.

Initial steps in this direction were taken by Rietveld et al. (2020), who extracted emotional information (i.e., arousal and valence levels) from Instagram photos of different brands and combined them with text analysis to predict customer engagement with the brand. However, there is need for further research in order to achieve a more complete visual-emotional-cognitive mapping.

Monetizing the Value of Images: What Is the Value of Images in Various Stages of the Customer Journey?

For the first time, differential effectiveness of visuals throughout the purchase funnel can be quantified by image analytics. Specifically:

1. Which visual features are most appropriate for various stages of the purchase funnel – what visuals get consumers' attention? Enhance awareness? Increase consideration, liking, and purchase intentions?
2. Which visuals should a firm present to customers at various stages in the customer life cycle? For example, are different images effective for customer acquisition vs. repeat purchase, upgrade, development, and retention?
3. How should visual features in ads be priced? Image analytics could revolutionize the way ads are priced. While media outlets are priced according to their reach, creative advertising is priced based on the effort invested and the reputation of the creative team. Quantifying the value of various visual components of an ad can lead to a differential pricing scheme. For example, measuring the relative value of a face in an ad versus a white space, or scenery, enables value-based pricing of ad creatives.

Branding

Images play a key role in consumer brand perception, recall, and associations (Peracchio and Meyers-Levy 1994, 2005). Image analytics opens new opportunities for brands to execute their desired positioning through visuals, manage their brand portfolio, and foster brand collaborations.

Visual Brand Representation: What Is the Visual Representation of Brand Associations? How Does It Align with Brand Characteristics? What Is the Role of the Visual Brand Elements and Brand Communications in Shaping Brand Perception and Associations?

A recently proposed tool to explore the visual representation of brand perception and elicit brand associations was described in the work of Dzyabura and Peres (2021). They developed a platform for eliciting brand associations through creating and analyzing online collages of images and showed how these collages can be used to retrieve a visual representation of brand associations and to connect it to brand personality and brand equity metrics. Such approaches have the potential to address many additional questions relating to the nature of these associations, their dynamics over time, their representation in brand communications, and their connection to various brand metrics.

Every brand has a unique set of visual brand elements (logo, colors, fonts, etc.) created by designers in collaboration with brand managers. These brand elements

reflect the brand positioning, foster the desired associations and differentiate the brand from its competitors. Through image analytics, marketing scholars and brand managers can evaluate to what extent a proposed design achieves these goals (Dew et al. 2019).

Brand Hierarchy: What Are the Optimal Relationships Between the Visual Elements of Brands in a Brand Portfolio?

Sub-brands within a brand hierarchy require identities which are distinct from one another and yet convey the identity of the master brand. Brands vary in the extent the master brand dominates these sub-brand identities. For example, Fig. 2 shows the brand hierarchy of FedEx and Gillette. For FedEx, the master brand visual elements are clearly dominant, while for Gillette, the sub-brands have distinct visual elements of their own with the master brand being represented to a much lesser degree.

Image analytics can assist in achieving the desired balance between these two extremes. First, image analytics methods can be used to measure the level of visual coherence within the brand hierarchy. Second, it identifies the visual elements that create the perception of similarity. Third, it can connect the overall visuals of the hierarchy to brand performance metrics.

Brand Strategic Collaborations: When Brands Collaborate with Each Other, What Is the Right Mix of Their Visual Elements Which Will Ensure that Both Brands Are Fairly Represented?

Creating a visual identity for a collaboration of brands is often challenging and complicated for the collaborating parties to agree which visual elements should be taken from each brand and how to combine them together. Consider, for example, the two designs of the joint Philadelphia-Milka brand illustrated in Fig. 3. Design A contains more Milka colors, but a larger Philadelphia logo than Design B. Do they manage to achieve parity? Image analytics can help address such dilemma by evaluation to what extent a proposed design represents the desired collaborative identity.



Source: FedEx (2020)

Fig. 2 Examples of the brand hierarchies of FedEx and Gillette



Fig. 3 Comparison of collaborative design packages with different mix of visual elements

Online Shopping Experience

Image analytics can help firms make better decisions with respect to physical store design and the visual elements of online shopping outlets.

The Role of Visuals in Online Product Display: How Does the Composition of Visual Elements, Objects, Size, Background, and Relative Location Impact the Search, Click, and Purchase Propensity?

When photographing a shirt for the online shop, the retailer has numerous options as to how to present the item: folded neatly on a flat surface, laying more carelessly, hanging against the wall, worn by a model, photographed against a solid color background, in an outdoor or indoor location, etc. All these factors influence consumer reactions and expectations from the product.

For example, Zhang et al. (2019) demonstrate that the photographic properties of homes displayed on Airbnb, such as diagonal dominance and rule of thirds, influence demand. Li et al. (2019a, b) show, on the same platform, that the order and layout in which the photos are presented also influence demand. Peng et al. (2020) show that the facial attractiveness of hosts on such rental platforms also influences the occupancy of these homes. More research is needed to explore a larger variety of situations, context, product categories, and consumer behaviors and understand their underlying mechanisms.

The Role of Visuals in Ecommerce Website Design: How Do the Visual Components of an Ecommerce Website Contribute to Profitability?

Designing an ecommerce website is a visual challenge. Designers must make decisions on the sizes and colors of the items that are displayed on the website (e.g., product images or buttons) and create a design that helps users to search for products, explore assortments, get inspired, and discover new products. At the same

time, the items should be presented in a way that will match their brand identity. This raises several practical questions which can be answered by employing image analytics methods:

1. How do images change/affect consumers' propensity to keep searching on the website? How does this propensity change at different stages of the search process?
2. How to create more personalized website layouts? For example, Hauser et al. (2009, 2014) use a multi-armed bandit approach that balances exploration and exploitation to automatically match the look-and-feel of the website to customers' cognitive styles.

Consumer Perspective

Studies have shown that consumers use photos to express emotions and attitudes as well as to document their experiences (Van House et al. 2005). This usage has greatly increased with the abundance of mobile phone cameras, storage space, and sharing apps. While traditionally photos were taken on special occasions, people have moved to continuously documenting and sharing their daily routines.

Uncovering Consumer Attitudes: What Are the Hidden Consumer Traits and Attitudes That Can Be Revealed Through Images and Go Beyond the Standard Metrics?

The rich body of consumer-generated photos can be used by researchers to gain a deeper understanding of the consumer experience, to profile and characterize a wide range of experiences, and, in addition, to segment consumers based on dimensions that could not be revealed otherwise.

For example, photos taken by consumers (either posted on social media or collected directly through mobile diaries) can show what is the actual choice set that consumers face when walking around the supermarket; what their environment looks like when sitting in a restaurant; what food brands are served at the same meal; what is their personal style and how it relates to the brands they buy, etc.

Data: Consumer Vs. Firm Images

Once the research question has been formulated, constructing the appropriate data is the next key step. A good dataset for image analysis should satisfy the following criteria: First, it should capture the specific constructs being studied. In many cases this involves the combination of images and additional data. For example, photos that users post on restaurant reviews and the corresponding restaurant financial performance (Zhang and Luo 2019), or photos of Airbnb properties accompanied

by property price, location, history, and host characteristics (Zhang et al. 2019, 2021; Li et al. 2019a, b).

Second, the dataset should be large enough to allow drawing insights. The state-of-the-art deep neural network models are trained on the ImageNet dataset, a freely available dataset which contains over 1.2 million images, organized into one thousand categories (<http://image-net.org/>). In marketing such big sizes are rare, but the datasets still have to contain thousands of images for researchers to be able to draw meaningful insights. Third, the dataset should contain minimal biases that could interfere with the main constructs. For example, using user-generated content to understand brand perception should be done carefully, since the sample is not controlled, and since users often post strategically to signal something about themselves (rather than about the brand) to their peers.

Below we describe the main data sources for image analytics in marketing, as summarized in Fig. 4. As illustrated in the figure, the data sources can be classified into consumer-generated images and firm-generated images.

Consumer-Generated Images

Consumer images include all the images created by consumers for different purposes: as a part of their own documentation of experiences and memories, for the purpose of sharing with other consumers, and for sharing with firms. They can be retrieved by researchers either through mining Internet and social media outlets, or directly elicited through surveys, diaries, panels, and collage making tasks.

Images from Internet and Social Media

Consumers increasingly share images on social media platforms such as Instagram or Facebook, and also on review platforms such as Yelp or [Booking.com](https://www.booking.com). For example, both Rietveld et al. (2020) and Liu et al. (2020) use Instagram images to monitor how brands are portrayed by consumers, and compare their perception to firm-generated visuals. Zhang and Luo (2019) use consumer-posted images on Yelp as a leading indicator of restaurant survival. They show that photos are more predictive of restaurant survival than reviews. Jalali and Papatla (2016) use brand images posted by users on Instagram to see how the color composition of the photo

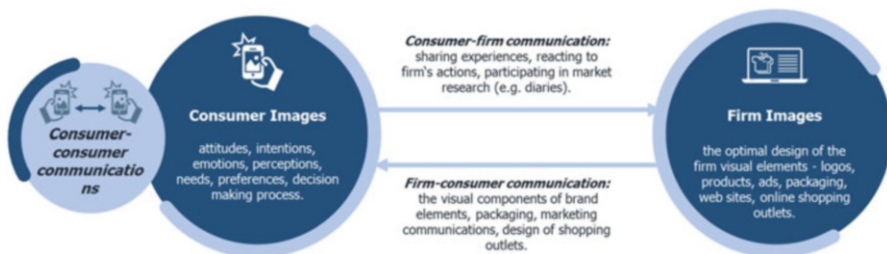


Fig. 4 Sources for image data in Marketing

influences its click-through rate, when a photo is curated by the website of the brand. They found that click-rates are higher for photos that include higher proportions of green and lower proportions of red and cyan. They also found that photos with higher click-rates are characterized by higher chroma of red and blue. Klostermann et al. (2018) use brand-related Instagram posts to derive insights on how consumers think and feel about a brand (McDonald's) in different brand-related situations.

The appeal of Internet and social media as sources of consumer-generated images is that they are abundant, free, unaided, and cover a broad range of topics. However, for many relevant research questions, these data will not capture the constructs of interest. First, social media data are available for only certain categories and brands: while the brand Nike generates a lot of social media commentary, finding social media posts on other brands such as Colgate, is difficult. Second, it is difficult to control the characteristics of the content contributors. For example, users who have a stronger relationship with the brand (Labrecque 2014), or hold a particularly strong positive or negative opinion, may contribute more than those who have only mild opinions (Lovett et al. 2013). Finally, it is important to carefully interpret the content, since consumers' posts may serve a self-signaling or other purpose (Han et al. 2010).

Social media resources are also valuable in constructing the visual representation of concepts. In many cases the images are tagged and labeled by the users, and these tags can be used as a means of describing the content of the picture. This labeling goes beyond object detection. It can be used to interpret the visual representation of emotions (e.g., happiness), abstract characteristics (e.g., glamorous), and general concepts (e.g., big-city life). For example, several researchers have used Flickr to gather an annotated dataset of images (e.g., Dhar et al. 2011; McAuley and Leskovec 2012; Zhang et al. 2012; Dzyabura and Peres 2021). Flickr lends itself well to gathering an annotated image dataset, because it provides a search engine that returns the most relevant images for a keyword. The search is based on text labels provided by users, image content, and clickstream data (Stadlen 2015). An image ranked at the top for a particular query has often been validated by tens of thousands of users who clicked on the image, reflecting a large population consensus regarding a strong association between the image and the query term.

Directly Elicited Images

Another approach for retrieving visual data from consumers is direct elicitation, namely, asking respondents to provide, create, rate, or choose images according to certain criteria.

Elicitation can go in one of two directions – one is presenting the respondents with an image and asking them to indicate the properties of interest in this image. For example, does the image look fun (Liu et al. 2020), does a clothing item in an image have asymmetry (Dzyabura et al. 2020), does a logo look modern (Dew et al. 2019), etc. Typically, several human judges are required to rank each image. Management of such tasks can be done using commercially available software tools such as Amazon MTurk or Appen.

The other direction is to provide respondents with a concept (a brand, an emotion, a mood, etc.) and ask them to select or create the images that best represent this

concept to them. Such a technique is used by Dzyabura and Peres (2021) who developed a brand visual elicitation platform that allows firms to ask consumers to create collages of images that they associate with a brand. Collage making is a projective technique that has long been used for qualitative research by psychologists (Koll et al. 2010) and brand researchers (Zaltman and Coulter 1995; Zaltman and Zaltman 2008). Typically, participants select images representing the concept in focus, and then explain to the moderator why they chose these images. Dzyabura and Peres (2021) have used image analytics to transform this task into a quantitative market research method. Using online data collection, image processing, and machine learning techniques, collage making now allows researchers to retrieve a large number of images for any concept of interest, over a large number of respondents. Figure 5 describes examples of collages (with some verbal descriptions) created using this method for the brand Starbucks.

Another useful method to directly elicit visuals from consumers is mobile diaries. Mobile diaries are a trending tool to collect repeated self-reports about experiences. They have been used as a research tool in variety of domains including psychology, geography, health, medicine (e.g., Hektner et al. 2007; Heinonen et al. 2012; Hensel et al. 2012; Hofmann and Patel 2015), marketing practice, and recently also by academic researchers (Lovett and Peres 2018). In mobile diary visual studies, respondents are usually asked to take photos of certain experiences – for example, photograph what they see on the shelf, windows of



Fig. 5 Examples for four collages for the brand Starbucks, created by four different respondents, with verbal descriptions. (Source: Dzyabura and Peres (2021))

stores they stop by, or the content of their refrigerator. These data can be later analyzed to create a data-driven representation of consumer stimuli, choice set, environment, and experience.

Note that data sources can be combined. For example, directly elicited data can be used to validate conclusions derived from social media, to provide insights on the underlying mechanisms, or complement the social media-generated data with interpretable features. For example, Hartmann et al. (2021) complement data from Twitter and Instagram with results from a lab experiment to show that the mechanism behind higher click-through rates of brand selfies (images of consumers holding a branded product, but face not showed in the frame) is that brand selfies induce more self-related thoughts. Peng et al. (2020) use surveys to elaborate on the mechanism behind the U-shaped relationship between facial attractiveness of the seller and product sales.

Firm-Generated Images

Firms continuously create visuals as part of their marketing efforts. Data originating from firms come in different formats such as visual brand elements (logo, colors, fonts etc.), product images on online stores, images used in advertising, and the firm's social media outlets. All these provide rich data for visual research. Unlike consumer-generated images, firm-generated images are typically curated and created by professional teams to meet the firm positioning goals. Thus, they constitute a visual representation of the firm strategy and can be used to study market structure and competitive landscape. The reactions to these images by consumers can, in turn, be used to study consumer response to various marketing actions. We list below several main sources for firm-generated data.

Product Images on Retail Websites

Retail websites are a great source for product images, since they contain many products from various vendors. The images are typically of good quality, focused on the product itself, and often capture the product from various angles. Many retail sites also require standardization of the images. For example, the shoe retail website [Zappos.com](https://www.zappos.com) photographs all shoes in the same way: from seven angles, against a solid white background. The uniformity makes it easier for image processing algorithms to focus on the product image. Retail sites typically provide other relevant product information such as price, materials, size, manufacturer, and brand, which can complement the analysis.

The challenge with using image data from retail websites is that they often lack data on many dependent variables of interest, such as clicks, likes, purchases, profitability, product returns, and repeat purchase rates. Answering research questions regarding these variables requires collaboration with the firm. For example, Dzyabura et al. (2020) used product images from an apparel retailer's online shop and collaborate with the firm to obtain information on the corresponding products' online and offline purchases and product return rates. They found that incorporating

the image in the prediction model in addition to the nonvisual attributes (e.g., price, category, season, size) greatly improved its accuracy.

Some multi-vendor retail platforms such as Alibaba, eBay, Airbnb, and Amazon provide consumer reviews and information about the product popularity, which can serve as a proxy for some of these variables. Zhang et al. (2021), for example, use property images posted by Airbnb hosts and combine it with the occupancy rate of the property.

Images on the Firm Social Media Pages

The social media pages of firms are also a rich source of image data. They contain, in addition to the products themselves, other components of the firm's visual representation – endorsers, users and usage scenarios, print and video ads, brand elements, sceneries, locations, activities, and all the visuals curated by the firms to create and enhance its brand associations. These data can be used to study the competitive landscape by identifying differences and similarities between the visual representation of competing brands, as well as between the brand self-representation and consumer perceptions of the firm. Unlike the retail sites, social media outlets also contain more dependent variables, such as consumer likes, shares, reactions, and comments. For example, Li and Xie (2020) used photos of major airlines and sport utility vehicle brands collected from Twitter and Instagram, and measured the engagement they created through retweets and likes. One of their findings is that the presence of a human face and the fit between the image and the textual content of the post can induce higher user engagement on Twitter but not on Instagram.

The Firm Brand Communications

Researchers can use visual elements of brands of interest in order to explore questions related to brand associations and brand image. This often requires the research team to assemble their own dataset. For example, Dew et al. (2019) assembled a dataset consisting of logos, textual description of firms, industry labels, and brand personality rating of 706 major brands. Then they used image analytics to explore the visual elements of logos they assembled and show how they can be used to create new brand identities and spark ideation.

Advertising Databases

Advertising visuals are continuously being generated by firms, displayed in social media outlets, websites, magazines, TV channels, and billboards. Interestingly, central repositories of advertising images are hard to find. Studies on advertising design effectiveness are mostly behavioral (Wedel and Pieters 2008) and use specific manipulations to test theories. A notable exception is the paper of Pieters et al. (2007), who used data from advertisements by several chains of grocery retailers in the Netherlands to measure the relative importance of the pictorial content in the ad in getting consumer attention.

Large-scale advertising image data across multiple firms, if assembled, could be used to study aspects of parity and differentiation between similar offerings and explore how the competitive landscape is reflected in the visual space. Combined

with brand perception, consumer responses, and advertising expenditure, these data can be used to study advertising effectiveness and provide guidance for the optimal design for an ad. Such data repositories are already available in many domains such as fashion (Xiao et al. 2017), autonomous driving (Caesar et al. 2020), and medical imaging. A common general repository is ImageNet, a public dataset of 1.2 million images labeled by humans, which was used to train many state-of-the-art models (e.g., Krizhevsky et al. 2017). A joint data collection effort in advertising could lead to many new and impactful insights on the visual aspects of advertisements.

Methods

Marketing researchers who study visual data have an unprecedented opportunity of access to state-of-the-art methods and analysis techniques. These methods are rapidly improving due to the increased computational power and ongoing efforts of the machine-learning community to broaden the scope of the analysis tools and make them publicly available and user-friendly. An image analytic process is typically composed of the four stages: feature extraction, model training, model evaluation and validation, and model application to the marketing problem.

Feature Extraction

The first step of the image analytics process is determining what feature space to work in. A key challenge of working with images is that the raw input elements – the pixels that make up the images – are not suitable features. A single pixel in isolation does not lend itself to meaningful interpretation. Compared to text, for example, this challenge is particularly pronounced. In text, the basic unit of analysis is words, which carry a meaning, a positive or negative valence, and can be grouped by topic. Pixels, on the other hand, have none of these properties. Therefore, a critical step in any modeling of image data is generating features which will provide a meaningful representation of the images.

There are multiple approaches to feature generation. One is predefined feature extraction. Researchers have developed a variety of predefined features. Perhaps the simplest are **color** histograms, which capture the distribution of the color composition of the image. Such histograms are created by discretizing the colors in the image into bins, based on a color space, and counting the number of image pixels in each bin. The most common color spaces are RGB (red, green, blue) and HSV (hue, saturation, value). Another dimension of interest is **shape**, where the features are line directions, corners, and curves. A third common property is **texture** – defining the repeating patterns in the image, such as line and color intensity. Texture is most commonly measured by a Gabor filter, which detects repeating frequencies of color in certain parts of the image. For videos, Li et al. (2019b) add a dynamic component by defining a measure of visual variation, calculated by decomposing a video into a

number of static frames and then computing the visual distance between consecutive frames.

The role of the feature extraction step has been revolutionized with the development of deep neural networks (NN). In an NN, the feature selection and the model training are done simultaneously, so the network automatically extracts the features that are optimal for the specific research problem. For example, it would extract different features for determining whether there is a pedestrian or a traffic light in an image, versus determining whether an image appears fun or serious. A deep NN does such simultaneous training by applying several “layers” of non-linear transformations on the raw pixel data. The outputs of each transformation serve as the features, or predictor variables, for the next layer. Through multiple layers of such transformations, the network extracts higher- and higher-level representations of the data, allowing the final layer to easily classify the data. For example, lower layers of a deep-learning model may extract edges and textures, whereas higher layers detect motifs, object parts, and complete objects (Goodfellow et al. 2016). The final layer maps the resulting features onto the target variables with a classification function.

There are many neural network architectures that are used in different applications that differ in the types of functions captured by their nodes, their depth, data flow, etc. – together, these make up the network architecture. The networks that work best for image analytics problems are Convolutional Neural Networks, or ConvNets. ConvNets are characterized by the first several layers of the network being *convolutional layers*: each neuron applies a particular transformation to a small part of the image. Rather than applying a transformation to the entire image, it processes small “batches” of the image separately, to detect shapes and edges in different parts of the image. Two commonly used ConvNet architectures are ResNet and VGG19.

Feature extraction using deep NN almost always results in higher predictive accuracy than human-coded features. However, it has two major caveats. One is that the features are not interpretable – they are complex nonlinear transformations of pixels, which have no meaning on their own. While this is not a disadvantage if the main task is prediction, it is, if interpretable insights are desired. For example, interpretability is important if the goal is to understand what people associate with a brand (Klostermann et al. 2018; Dzyabura and Peres 2021), what kind of image content gets most engagement on social media (Li and Xie 2020), to give recommendations for photographing a home for rent on Airbnb (Zhang et al. 2019), or to create a promotional video for a project in a crowd funding platform (Li et al. 2019b). In such cases, the modeling must be done on interpretable features. One way to obtain them is by using tagging software or a tagged dataset from sources such as Flickr, in order to identify the objects, activities, sceneries, and themes presented in the image. Thus, the image is described by a set of words or tags, which serve as the features, and the image analysis task is transformed into a text analysis task. This opens a wide range of options for the analysis: using word embeddings, various dictionaries such as LIWC, sentiment analysis techniques, and topic modeling (chapter ► “Automated Text Analysis”).

The other caveat is that training a deep neural network from scratch is extremely challenging: it requires a very large annotated dataset, massive memory and computational power, and complex engineering. Additionally, a lot of modeling choices, such as the number, type, and order of layers, how much regularization to use, and learning rate, are made through trial and error. To simplify the training stage, most deep learning applications in marketing use *transfer learning*: rather than designing and training a new neural network for every task, they use a network that has already been trained by someone else for a different purpose. The reasoning behind it is that knowledge gained from performing one task can be used to perform another. For example, knowledge gained from recognizing image aesthetics could be applied to forecast product demand or recognize brand perceptual attributes (Bengio et al. 2011; Bengio 2012).

The transfer can be done either by taking the final layers of the trained network as is, or by fine tuning the trained NN. Dzyabura et al. (2020) took the first approach – they used the second-to-last pre-output of ResNet, which is trained on the ImageNet data, as features in a random forest model to predict the return rates of clothing items.

The fine tuning approach does train the NN, but instead of initializing the model parameters with random numbers, the model is initialized with parameters learned from another NN. The idea is similar to using an informed prior in Bayesian estimation. Relative to training the model from scratch, fine tuning significantly increases model performance and avoids overfitting (Donahue et al. 2014; Girshick et al. 2014; Yosinski et al. 2014). Li et al. (2019a, b) employ fine tuning by using ResNet50 to train their model and learn picture quality and room type (e.g., bedroom, bathroom) for images from Airbnb postings. The resulting features are used to predict occupancy rates of the properties. Interestingly, Zhang et al. (2019) also predict image quality on Airbnb, but they use a different pretrained model, VGG16 (Simonyan and Zisserman 2015), also pretrained on ImageNet. The paper uses the results to explain the decision-making process of the hosts who use pictures of lower quality even when a high-quality option is free and available.

Model Training

Regardless of what approach was chosen for feature extraction, the resulting feature space for an image problem will be very large, often larger than the number of observations. Standard statistical methods which assume linear models and estimate their coefficients cannot be applied. The large number of coefficients makes it impossible to identify every single coefficient without bias. Therefore, researchers apply machine learning methods which are tailored for working in very large feature spaces.

Many image analytics problems can be formulated as a *supervised classification task* – determining whether an image belongs to one or multiple predefined categories or classes. For example, in Liu et al. (2020), image classification is used to determine whether an image exhibits a brand perceptual attribute: does the image look fun? rugged healthy? glamorous? A classifier is trained on an annotated dataset

of images labeled with the desired classes, that is, images that are known to belong or not to belong to the classes (e.g., images which are glamorous and not glamorous). As explained in “[Product Design](#),” the annotated datasets can be taken from publicly available sources, company data, or collected by the researchers. After training, the classifier can assign a class to a new, unlabeled image.

Some research problems do not involve classifying images into predefined categories. Instead, the researcher is looking to identify patterns in the data – such as recurring objects, colors, shapes, and themes. This is most commonly done by representing the images in a feature space (either using predefined features or with deep neural networks), and then using *unsupervised methods* (e.g., clustering using K-means or nearest neighbors) to group them together (Dew et al. 2019; Peng et al. 2020; chapter ► “[Cluster Analysis in Marketing Research](#)”). If the feature extraction is based on tagging, then one could use unsupervised methods for text analysis – such as topic modeling (Dew et al. 2019; Nanne et al. 2020; Peng et al. 2020). In some cases, the image analytic task aims at creating novel combinations of existing patterns, for example – to create new designs of the product, using generative models (Burnap et al. 2019), or predicting “design gaps” in a certain market (Burnap and Hauser 2018).

Model Evaluation and Validation

Once the model has been trained, it is important to evaluate and validate it. Evaluation establishes its performance and validation ensures that the output obtained from the images measures the construct of interest.

A proper evaluation is done by testing the model performance on a different sample than the one it was trained on. In most machine learning algorithms, a portion of the sample is held out and used to test the model. This out-of-sample test is important since the large size of the feature space can easily lead to overfitting. This is only true for supervised learning. For unsupervised learning, model accuracy is hard to evaluate, because there is no specific independent target variable.

Validation depends greatly on the nature of the task. Basically, validation ensures that the model was successful in capturing the construct it intended to measure. For example, if the analysis was done to assess how happy a face is, the results should be validated by testing that faces that were identified as happy are indeed perceived as happy by people. Validation is particularly challenging and particularly important in the unsupervised case. Since predictive accuracy cannot be shown, how can one prove that one clustering is better than another clustering, or that the identified patterns are true? For example, Dzyabura and Peres (2021), used images to extract brand associations from collages of images created by users. They used two layers of validation to demonstrate that the extracted associations are the correct ones: users were first asked to match extracted associations to a collage, and second, to guess the brand based on the associations. Dew et al. (2019) built a model to predict the visual features of a logo based on the verbal descriptions of the brand from the company website, and validated it by taking the brand ShakeShack, using the model to predict

the visual elements of its logo and comparing it against the existing logo. Peng et al. (2020) studied face attractiveness and whether it can predict product sales in ecommerce platforms. After extracting facial features, they validated the model using a group of coders that were asked to rank the attractiveness of the faces. In Zhang and Luo (2019), user-posted images were used to predict survival of restaurants. In such tasks, when looking at the examples the algorithm misclassified, one can notice which types of restaurant the classifier fails, and add information accordingly. In addition, by calculating the proportion of mistakes of each type, we can have a better understanding of the precision rate that is possible.

Model Application

The final step of the analysis is applying the model to the research problem, whether it be computing a brand metric (Liu et al. 2020), forecasting demand (Peng et al. 2020), or optimizing the visual communication (Li and Xie 2020; Li et al. 2019b). This stage is important, since in marketing, clustering or classifying images is rarely the end goal. The images are a manifestation of a more fundamental underlying construct, and their analysis is typically an intermediate step in deriving meaningful insights with respect to this construct and its relationship with perceptual, behavioral, and economic variables.

Integrating It All Together

Image analytics could very easily go wrong. The researcher is faced with numerous data sources, code packages, constantly improving methods, and pre-trained models. All of these open a broad range of research opportunities, yet they often create confusion as to the right choice of the model components. Specifically, the researcher has to carefully match the research problem, data, and method. This is a challenging task: the data, although very rich, might not contain the variables of interest; the model might be good in classifying images but incapable of yielding interpretable insights; the data can suffer from various biases and confounds, such as user strategic posting and self-signaling. Many failures in image analytics tasks are caused by incorrect matching between the various components, leading to none, or even worse – misleading insights.

To ensure an optimal match between the research question, data, and method in order to produce the highest quality analytics with meaningful insights, the researcher should ask herself two questions: first, whether or not there is a single dependent variable that is the crux of the research question. Such a variable could be demand (Zhang et al. 2019, 2021), crowdfunding success (Li et al. 2019b; Peng et al. 2020), business survival (Zhang and Luo 2019), ad recall (Rosbergen et al. 1997), or product return rates (Dzyabura et al. 2020). Second, whether interpretability of the features is important for the task. That is, do the desired insights involve interpretation of specific elements of the image? The answers to these questions determine

the appropriate methods and data type. They can be described in the following 2×2 matrix presented in Fig. 6.

Most computer vision tasks fall into the *top right quadrant* of the matrix: there is a specific target variable of interest and interpretation of the features is not necessary. This is the quadrant where most engineering computer vision problems belong. A typical computer vision task is to identify, for instance for a self-driving car, whether an image contains a pedestrian or a traffic light. A good algorithm for such problems is engineered to detect the objects of interest with a low probability of error. It does not need to be able to say what about the picture forms a pedestrian. The set of methods in this quadrant are by definition supervised, and are typically based on deep neural networks.

Thanks to the rapid growth and development of the computer vision field, research questions in marketing that fall into this quadrant have a rich and constantly improving set of methods to choose from. The choice of method depends on the nature of the problem. Nanne et al. (2020) compared different computer vision algorithms to monitor user-generated content, and found that they have different strengths: Google Cloud Vision is more accurate in object detection, whereas Clarifai provides more useful labels to interpret the portrayal of a brand and YOLOV2 did not prove to be useful to analyze visual brand-related UGC. This should be taken into account when conducting the analysis, and one might need to use several methods and assess their performance for the specific research problem.

In order to apply these methods, the researcher needs to obtain good data from a reliable source. The image data must be annotated with the target variable and it must be very large: contain thousands of annotated images in the training set. In marketing, this is often challenging to accomplish. Most brands, for example, do not

		Does the task require interpretability of the features	
		Yes	No
Is there a single target variable?	Yes	Interpretable features – theory based, or tagging. Econometric models.	Dataset that is labelled with the target variable. Supervised – deep learning models.
	No	Interpretable features – theory based, or tagging. Unsupervised classification	Structure the problem – either by finding target variables that capture the construct, or identifying a relevant feature space

Fig. 6 A classification matrix for combining data, features, and image analytic model

produce thousands of images. To get a large enough dataset, one may need to pool data across multiple brands or over time, risking introducing noise to the data.

Unlike engineering, many marketing questions do involve some sort of interpretation of the image characteristics and their relationship with the target variable. For example, in the case of forecasting demand based on product images, one likely wants to know what it is about the product or the way it was photographed that leads to high or low demand. Such cases belong to the *top left quadrant*. In this quadrant, one faces a tradeoff, because the most accurate predictive models are not based on interpretable features. Once one introduces a constraint on the types of features, predictive accuracy will be compromised.

One way in which interpretable features can be generated is using image tagging software which extracts the image content and produces a list of the objects, sceneries, activities, moods, and other themes in the image (Klostermann et al. 2018; Rietveld et al. 2020; Nanne et al. 2020). Another way is to apply domain expertise to identify the relevant features. In photography, these may be diagonal dominance and rule of thirds (Zhang et al. 2019); in apparel, these may be types of prints, graphics, collar type, sleeve length, symmetry or metallic details (Vilnai-Yavetz and Tifferet 2015). These can be extracted using either machine learning classifiers or using human judges. Once the features were extracted, one can use econometric methods, such as regression, to obtain insights as to how they relate to the target variable.

If both interpretable insights and accurate predictions are important, the analysis should include both: a deep learning model trained to optimize the features and the model for maximum predictive accuracy, and regression analyses over interpretable features. For example, Dzyabura et al. (2020) use deep learned features to predict the return rate of a product based on its image. Then, they have four independent judges manually label the images with respect to industry standard design elements such as symmetry, pattern (solid, floral, striped, geometric/abstract), and additional details (text, metallic/sequin, graphic, lace). The authors then analyze which of these are associated with higher return rates in a regression.

The *bottom left quadrant* comprises situations in which there are no dependent variables of interest. Instead, we ask an open-ended question such as – how is a brand perceived by consumers? How do consumers use the product? What visual features of logos are associated with what brand perceptions?

These questions often require the combined use of interpretable features with unsupervised learning algorithms. The features create a meaningful and managerially relevant space, into which all the relevant observations can be mapped. The unsupervised algorithms, in turn, detect patterns and identify data-driven classifications in this space. For example, Dew et al. (2019) use features taken from theories of logo semantics to form a “visual dictionary” that describes logos in a way that is meaningful to designers (e.g., the amount of white space, corners, and edges). They then use a probabilistic modeling framework to flexibly capture the linkages between the brand descriptions, logo features, industry labels, and brand personality metrics. Dzyabura and Peres (2021) used tagging to identify image content, and then use unsupervised topic modeling to reveal latent topics. Klostermann et al. (2018) tag

objects and situations using object detection software, and then employ unsupervised clustering algorithms to form associative networks connecting image content to consumer sentiment. The resulting map of associations is informative for brand management, communications, and monitoring the response of consumers to new products and features.

Finally, questions that belong to the *bottom right quadrant* do not have a specific target variable and do not require interpretability of the features. This quadrant is challenging since it applies structure neither to the features nor to the dependent variable. In order to impose structure on these problems one might consult with domain experts to identify either some relevant dependent variables that are of interest to managers or a set of features that represent the space in a meaningful way.

Conclusion

In the past two decades, major technological advances and the popularity of digital platforms made taking and sharing images a crucial part of consumers' daily lives. In addition to the abundance of visual data, image processing tools and advances in modeling techniques created unprecedented opportunities to obtain new perspectives on important marketing questions. We are now able to study new phenomena, investigate the relationship between consumers and firms and obtain insights that would have been difficult or impossible to obtain otherwise.

Using image analytics to generate insights is not trivial though. Researchers are faced with different sources of data, various analysis techniques, and continuously improving methods. In order to benefit from implementing image analytics in solving relevant marketing problems, matching a good research question with the right visual data and appropriate method comes with many challenges. However, once the researcher is able to surmount these challenges, many marketing areas can benefit from image analytics to gain new insights. In the area of **Product Design**, researchers can for example explore how to characterize designs above and beyond their specific elements. Moreover, they can use image analytics to quantify the value of designs by incorporating product images in traditional consumer demand models. In the area of **Advertising**, image analytics can allow for a holistic quantitative approach to selecting, adjusting, and optimizing the visual composition of print and video advertisements. In **Branding**, image analytics opens new perspectives for firms to strategically position their brands, manage their brand portfolio, and identify new collaborations. Image analytics can also help firms make well-grounded decisions to enhance consumers' **Online Shopping Experience** by identifying the role of visuals in ecommerce websites for example. Finally, from a **Consumer Perspective**, image analytics has the potential to reveal through images more about consumers than we knew so far. For example, firms can understand how consumers see brands, how they think about consumption, and how they perceive and evaluate their environments.

Cross-References

- ▶ [Automated Text Analysis](#)
- ▶ [Cluster Analysis in Marketing Research](#)

References

- Amit, E., Algom, D., & Trope, Y. (2009). Distance-dependent processing of pictures and words. *Journal of Experimental Psychology: General*, 138(3), 400.
- Ang, S. H., Lee, Y. H., & Leong, S. M. (2007). The ad creativity cube: Conceptualization and initial validation. *Journal of the Academy of Marketing Science*, 35(2), 220–232.
- Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.), *Proceedings of the ICML workshop unsupervised transfer learn* (pp. 17–36). Bellevue.
- Bengio, Y., Bergeron, A., Boulanger-Lewandowski, N., Breuel, T., Chherawala, Y., Cisse, M., & Erhan, D. (2011). Deep learners benefit more from out-of-distribution examples. In G. Gordon, D. Dunson, & D. Miroslav (Eds.), *Proceedings of the 14th international conference artificial intelligence statist* (pp. 164–172). Fort Lauderdale, FL.
- Bloch, P. H. (1995). Seeking the ideal form: Product design and consumer response. *Journal of Marketing*, 59(3), 16–29.
- Burnap, A., & Hauser, J. (2018). Predicting “design gaps” in the market: Deep consumer choice models under probabilistic design constraints. *arXiv preprint arXiv*, 1812.11067.
- Burnap, A., Hauser, J., & Timoshenko, A. (2019). *Design and evaluation of product aesthetics: A human-machine hybrid approach*. Available at SSRN 3421771.
- Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). NuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11621–11631).
- Chan, T. H., Mihm, J., & Sosa, M. E. (2018). On styles in product design: An analysis of US design patents. *Management Science*, 64(3), 1230–1249.
- Cho, H., Schwarz, N., & Song, H. (2008). Images and preferences: A feelings-as-information analysis. In M. Wedel & R. Pieters (Eds.), *Visual marketing: From attention to action* (pp. 259–276). New York: Lawrence Erlbaum Associates.
- Crilly, N., Moultrie, J., & Clarkson, P. J. (2004). Seeing things: Consumer response to the visual domain in product design. *Design Studies*, 25(6), 547–577.
- Dew, R., Ansari, A., & Toubia, O. (2019). *Letting logos speak: Leveraging multiview representation learning for data-driven logo design*. Available at SSRN 3406857.
- Dhar, S., Ordóñez, V., & Berg, T. L. (2011, June). High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011* (pp. 1657–1664). IEEE.
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014, January). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (pp. 647–655).
- Dzyabura, D., & Peres, R. (2021). Visual elicitation of brand perception. *Journal of Marketing*, forthcoming.
- Dzyabura, D., El Kihal, S., Ibragimov, M., & Hauser, J. (2020). *Leveraging the power of images in managing product return rates*. Available at SSRN 3209307.
- Eisenman, M., Frenkel, M., & Wasserman, V. (2016). Toward a theory of effective aesthetic communication. In *Academy of management proceedings* (Vol. 2016, p. 12822). Briarcliff Manor: Academy of Management.

- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT Press.
- Gorn, G. J., Chattopadhyay, A., Yi, T., & Dahl, D. W. (1997). Effects of color as an executional cue in advertising: They're in the shade. *Management Science*, *43*(10), 1387–1400.
- Greenleaf, E., & Raghubir, P. (2008). Geometry in the marketplace. In M. Wedel & R. Pieters (Eds.), *Visual marketing: From attention to action* (pp. 113–143). New York: Lawrence Erlbaum Associates.
- Han, Y. J., Nunes, J. C., & Drèze, X. (2010). Signaling status with luxury goods: The role of brand prominence. *Journal of Marketing*, *74*(4), 15–30.
- Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The Power of brand selfies. *Journal of Marketing Research*, *58*(6), 1159–1177. <https://doi.org/10.1177/00222437211037258>
- Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science*, *28*(2), 202–223.
- Hauser, J. R., Liberali, G., & Urban, G. L. (2014). Website morphing 2.0: Switching costs, partial exposure, random exit, and when to morph. *Management Science*, *60*(6), 1594–1616.
- Heinonen, R., Luoto, R., Lindfors, P., & Nygård, C. H. (2012). Usability and feasibility of mobile phone diaries in an experimental physical exercise study. *Telemedicine and e-Health*, *18*(2), 115–119.
- Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Thousand Oaks: Sage Publications, Inc.
- Hensel, D. J., Fortenberry, J. D., Harezlak, J., & Craig, D. (2012). The feasibility of cell phone based electronic diaries for STI/HIV research. *BMC Medical Research Methodology*, *12*(75), 1–12.
- Hofmann, W., & Patel, P. V. (2015). Survey signal a convenient solution for experience sampling research using participants' own smartphones. *Social Science Computer Review*, *33*(2), 235–253.
- Jalali, N. Y., & Papatla, P. (2016). The palette that stands out: Color compositions of online curated visual UGC that attracts higher consumer interaction. *Quantitative Marketing and Economics*, *14*(4), 353–384.
- Janiszewski, C. (1998). The influence of display characteristics on visual exploratory search behavior. *Journal of Consumer Research*, *25*(3), 290–301.
- John, D. R., Loken, B., Kim, K., & Monga, A. B. (2006). Brand concept maps: A methodology for identifying brand association networks. *Journal of Marketing Research*, *43*(4), 549–563.
- Keller, K. L. (2003). Brand synthesis: The multidimensionality of brand knowledge. *Journal of Consumer Research*, *29*(4), 595–600.
- Kireyev, P., Timoshenko, A., & Yang, C. L. (2020). *Scaling human effort in idea screening and content evaluation*. INSEAD Working Paper No. 2020/42/MKT, HEC Paris Research Paper No. MKG-2020-1384, Available at SSRN: <https://ssrn.com/abstract=3685882> or <https://doi.org/10.2139/ssrn.3685882>
- Klostermann, J., Plumeyer, A., Böger, D., & Decker, R. (2018). Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing*, *35*(4), 538–556.
- Koll, O., Von Wallpach, S., & Kreuzer, M. (2010). Multi-method research on consumer–brand associations: Comparing free associations, storytelling, and collages. *Psychology & Marketing*, *27*(6), 584–602.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.
- Labrecque, L. I. (2014). Fostering consumer–brand relationships in social media environments: The role of Parasocial interaction. *Journal of Interactive Marketing*, *28*(2), 134–148.

- Lehnert, K., Till, B. D., & Ospina, J. M. (2014). Advertising creativity: The role of divergence versus meaningfulness. *Journal of Advertising*, 43(3), 274–285.
- Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1), 1–19.
- Li, H., Simchi-Levi, D., Wu, M. X., & Zhu, W. (2019a). *Estimating and exploiting the impact of photo layout in sharing economy*. Available at SSRN.
- Li, X., Shi, M., & Wang, X. S. (2019b). Video mining: Measuring visual information using automatic methods. *International Journal of Research in Marketing*, 36(2), 216–231.
- Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. *Journal of Advertising*, 46(2), 236–247.
- Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4), 669–686.
- Lovett, M. J., & Peres, R. (2018). Mobile diaries – Benchmark against metered measurements: An empirical investigation. *International Journal of Research in Marketing*, 35(2), 224–241.
- Lovett, M. J., Peres, R., & Shachar, R. (2013). On brands and word of mouth. *Journal of Marketing Research*, 50(4), 427–444.
- MacInnis, D. J., & Price, L. L. (1987). The role of imagery in information processing: Review and extensions. *Journal of Consumer Research*, 13(4), 473–491.
- McAuley, J., & Leskovec, J. (2012, October). Image labeling on a network: Using social-network metadata for image classification. In *European conference on computer vision* (pp. 828–841). Berlin/Heidelberg: Springer.
- McQuarrie, E. F. (2008). Differentiating the pictorial element in advertising – A rhetorical perspective. In M. Wedel & R. Pieters (Eds.), *Visual marketing: From attention to action* (pp. 91–112). New York: Psychology Press.
- Meyers-Levy, J., & Zhu, R. (2008). Perhaps the store made you purchase it: Toward an understanding of structural aspects of indoor shopping environment. In M. Wedel & R. Pieters (Eds.), *Visual marketing: From attention to action* (pp. 193–224). New York: Psychology Press.
- Nanne, A. J., Antheunis, M. L., van der Lee, C. G., Postma, E. O., Wubben, S., & van Noort, G. (2020). The use of computer vision to analyze brand-related user generated image content. *Journal of Interactive Marketing*, 50, 156–167.
- Orsborn, S., Cagan, J., & Boatwright, P. (2009). Quantifying aesthetic form preference in a utility function. *Journal of Mechanical Design*, 131(6), 061001.
- Pavlov, E., & Mizik, N. (2019). *Increasing consumer engagement with firm-generated social media content: The role of images and words*. Working Paper, University of Washington.
- Peng, L., Cui, G., Chung, Y., & Zheng, W. (2020). The faces of success: Beauty and ugliness premiums in e-commerce platforms. *Journal of Marketing*, 84(4), 67–85.
- Peracchio, L. A., & Meyers-Levy, J. (1994). How ambiguous cropped objects in ad photos can affect product evaluations. *Journal of Consumer Research*, 21(1), 190–204.
- Peracchio, L. A., & Meyers-Levy, J. (2005). Using stylistic properties of ad pictures to communicate with consumers. *Journal of Consumer Research*, 32(1), 29–40.
- Pieters, R., & Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing*, 68(2), 36–50.
- Pieters, R., Wedel, M., & Zhang, J. (2007). Optimal feature advertising design under competitive clutter. *Management Science*, 53(11), 1815–1828.
- Radach, R., Lemmer, S., Vorstius, C., Heller, D., & Radach, K. (2003). Eye movements in the processing of print advertisements. In R. Radach & H. Deubel (Eds.), *The mind's eye* (pp. 609–632). Amsterdam: Elsevier Science Publishers.
- Raghubir, P., & Greenleaf, E. A. (2006). Ratios in proportion: What should the shape of the package be? *Journal of Marketing*, 70(2), 95–107.
- Reavey, P. (Ed.). (2012). *Visual methods in psychology: Using and interpreting images in qualitative research*. Routledge. London.

- Rietveld, R., van Dolen, W., Mazloom, M., & Worring, M. (2020). What you feel, is what you like influence of message appeals on customer engagement on Instagram. *Journal of Interactive Marketing, 49*, 20–53.
- Rosbergen, E., Pieters, R., & Wedel, M. (1997). Visual attention to advertising: A segment-level analysis. *Journal of Consumer Research, 24*(3), 305–314.
- Rubera, G. (2015). Design innovativeness and product sales' evolution. *Marketing Science, 34*(1), 98–115.
- Sheinin, D. A., Varki, S., & Ashley, C. (2011). The differential effect of ad novelty and message usefulness on brand judgments. *Journal of Advertising, 40*(3), 5–18.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of *International Conference on Learning Representations (ICLR)*. Available at <https://arxiv.org/abs/1409.1556>
- Smith, R. E., MacKenzie, S. B., Yang, X., Buchholz, L. M., & Darley, W. K. (2007). Modeling the determinants and effects of creativity in advertising. *Marketing Science, 26*(6), 819–833.
- Stadler, A. (2015). *Find every photo with Flickr's new unified search experience*. Available at <https://blog.flickr.net/en/2015/05/07/flickr-unified-search/>
- Toubia, O., & Netzer, O. (2017). Idea generation, creativity, and prototypicality. *Marketing Science, 36*(1), 1–20.
- Van House, N., Davis, M., Ames, M., Finn, M., & Viswanathan, V. (2005). The uses of personal networked digital imaging: An empirical study of cameraphone photos and sharing. In *CHI'05 extended abstracts on human factors in computing systems* (pp. 1853–1856). ACM.
- Venngage. (2020). *14 Visual content marketing statistics to know for 2020*. Available at <https://venngage.com/blog/visual-content-marketing-statistics/>
- Vilnai-Yavetz, I., & Tifferet, S. (2015). A picture is worth a thousand words: Segmenting consumers by Facebook profile images. *Journal of Interactive Marketing, 32*, 53–69.
- Wedel, M., & Pieters, R. (2000). Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Science, 19*(4), 297–312.
- Wedel, M., & Pieters, R. (2008). A review of eye-tracking research in marketing. *Review of Marketing Research, 4*(2008), 123–147.
- Wedel, M., & Pieters, R. (2014). *Looking at vision* (p. 2014). Abingdon: Routledge.
- Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv, 3*.
- Yang, X., & Smith, R. E. (2009). Beyond attention effects: Modeling the persuasive and emotional effects of advertising creativity. *Marketing Science, 28*(5), 935–949.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328). Available at <https://arxiv.org/abs/1411.1792>
- Zaltman, G., & Coulter, R. H. (1995). Seeing the voice of the customer: Metaphor-based advertising research. *Journal of Advertising Research, 35*(4), 35–51.
- Zaltman, G., & Zaltman, L. H. (2008). *Marketing metaphoria: What deep metaphors reveal about the minds of consumers*. Boston: Harvard Business Press.
- Zhang, M., & Luo, L. (2019). *Can User-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp*. Available at SSRN 3108288.
- Zhang, H., Korayem, M., You, E., & Crandall, D. J. (2012). Beyond co-occurrence: Discovering and visualizing tag relationships from geo-spatial and temporal similarities. In *Proceedings of the fifth ACM international conference on web search and data mining* (pp. 33–42). Available at <https://doi.org/10.1145/2124295.2124302>
- Zhang, S., Mehta, N., Singh, P. V., & Srinivasan, K. (2019). *Can lower-quality images lead to greater demand on AirBnB?* Technical report, working paper, Carnegie Mellon University.
- Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2021) "What makes a good image? Airbnb demand analytics leveraging interpretable image features". *Management Science*. Forthcoming.



Social Network Analysis

Hans Risselada and Jeroen van den Ochtend

Contents

Introduction	694
The Relevance of Network Analyses for Marketing Purposes	695
Network Metrics	697
Network	698
Basic Notation	698
Actor Level	699
Tie Level	701
Network Level	702
Network Data and Sampling Methods	703
Data Collection	703
Network Sampling	706
Social Network Analysis in R	707
Data	707
Calculating Actor Level Metrics	709
Calculating Tie Level Metrics	710
Modeling Social Contagion	712
A Word of Caution	713
Conclusion	715
Cross-References	715
References	715

Abstract

The increased awareness about the presence of social effects in consumer networks has inspired marketers to better understand and address the needs of their consumers through network analyses. In this chapter we consider network

H. Risselada (✉)
University of Groningen, Groningen, The Netherlands
e-mail: h.risselada@rug.nl

J. van den Ochtend (✉)
University of Zürich, Zürich, Switzerland
e-mail: jeroen.vandenochtend@business.uzh.ch

analyses as a set of techniques which allows researchers to analyze how the social structure of relationships around consumers affects their attitudes and behavior, and vice versa, how attitudes and behavior may affect the social structure. We focus on the types of network analyses that are currently most prominent within the field of marketing. We provide basic network theory and notation with references to key publications in the field. We also provide suggestions for software (packages) and useful functions including code snippets to support researchers and practitioners in setting up their first social network analyses. At the end of the chapter we discuss several more advanced network analysis methods and list several resources that might be useful to the interested reader.

Keywords

Social Networks · Social Influence · Social Contagion · Network Analysis · Consumer Networks

Introduction

While marketers have shifted from mass marketing techniques to personalized marketing strategies in which the focus lies on the individual consumer, a broad stream of literature in the social sciences has shown that consumers do not operate as separate entities. That is, consumers typically affect other consumers with their behavior and at the same time are themselves affected by the behavior of others. For example, Bikhchandani et al. (1992) show that consumers are likely to conform to the behavior of other consumers in order to reduce the risk associated with a commercial decision. This behavior is based on the presumption that the majority must be right. Another example is provided by Burnkrant and Cousineau (1975) who provide experimental evidence that consumers perceive products as more favorable after they observed other consumers evaluating the product positively. Thus, it has been known for decades that ignoring social effects prevents marketers from fully understanding consumers' behavior and leads to missed marketing opportunities. Fortunately, the increasing availability of consumer network data creates a widespread opportunity for marketers to fully realize the potential of marketing campaigns that incorporate social effects to better address the consumers' needs. This chapter provides both scholars and practitioners with fundamental information on social network analyses and allows them to get their first hands-on experience in estimating social effects from observational data.

In this chapter, we consider social network analysis as a set of techniques, which allows researchers to analyze how the social structure of relationships around consumers affects their behavior and attitudes, and vice versa, how behavior and attitudes may affect the social structure. That is, marketers can leverage network data in several ways. On the one hand, social networks among consumers facilitate the working of social influence. For example, networks facilitate the diffusion of information and therefore, observing the social network of consumers allows

marketers to strategically spread information about new products or services through word-of-mouth mechanisms (e.g., Peres et al. 2010). Not only do consumers regularly become aware of new products or competing offers through their peers, they also consider their peers to be the most trustful source of information (Nielsen 2015). As a result, consumer behavior is influenced by information provided by their social network on a daily basis. On the other hand, observing the behavior and attitudes of a consumer's social network can reveal latent information about the consumer her/himself (e.g., Goel and Goldstein 2014). For example, little is known about a consumer before acquisition. However, consumers tend to be part of a network that consists of individuals with similar preferences and characteristics, a phenomenon called homophily (Aral et al. 2009). Hence, when data about the consumer of interest is not directly available, data about their already acquired peers can provide a solution. Similarly, consumers' decisions to adopt or defect can inform marketers about the propensity of connected consumers to adopt or defect as well (e.g., Nitzan and Libai 2011; Landsman and Nitzan 2020). In summary, with the right network data and analyses, marketers are able to (1) identify consumers that can either be acquired or should be actively retained based on the behavior of their peers, and (2) try to acquire or retain these consumers by leveraging social effects within their personal network. Of course, the possibilities that network analysis provides to marketers are much broader, and the rapidly expanding research stream on social influence continuously provides additional insights about the relevance and usability of social effects among consumers, e.g., by integrating consumers' influence in their customer lifetime value (Kumar et al. 2010) or by identifying the most influential consumers (Goldenberg et al. 2009). The goal of this chapter is to provide a comprehensive overview of marketing insights that can be generated from network data, and to summarize and illustrate the most common methods to generate these insights. In the next section, we briefly discuss how network analyses can benefit both marketing researchers and practitioners. Subsequently, we discuss the most important components and metrics of network data. Next, we illustrate the suitability of various types of data and sampling methods for network analyses. After this, we will illustrate how to derive the network metrics and apply some of the discussed network analyses on a publicly available dataset. We will conclude with a few words on more advanced network analysis methods and list several resources that might be useful to the reader.

The Relevance of Network Analyses for Marketing Purposes

Network analysis enables marketers to derive a broad range of customer insights. We start with a brief overview of two broad research fields within the social sciences and marketing in particular: (1) social influence, which includes the study of the underlying social mechanisms and (2) the role of influencers, which includes the study of consumer network characteristics at an aggregated and individual level. These types of network analyses are broadly researched and applied in practice and will be the main focus throughout the chapter. However, network analyses can serve a much

broader set of marketing purposes. We refer to Valente (2012) for a brief overview of possible network interventions that can be used to accelerate behavioral change.

The first question marketers might ask themselves is how relevant social influence is for their customer acquisition, development, and retention processes. While the presence of social influence among consumers has been acknowledged for decades, the role of social influence can fiercely vary depending on the type of consumer decision and the underlying mechanism of social influence. First, *social normative pressure*, i.e., the discomfort a consumer might experience if they don't own a product their peers purchased, is typically high for products and services that are displayed to a broad audience (Burnkrant and Cousineau 1975). For example, peer groups can easily observe whether their members conform to the established fashion norms and praise those who do or reprimand those who don't. On the contrary, this social mechanism might be less relevant for brands that offer services and products that are consumed in a private setting or that are not strongly associated with consumer identity (Iyengar et al. 2015). For example, it is ambiguous to establish social norms on the use of household products because it is difficult to observe compliance and praise/reprimand accordingly. Second, the presence of *social learning*, i.e., gathering information from the decisions and experiences of peers, is typically strong for consumer decisions that involve high risk, while it will be much weaker for decisions that involve little investment and commitment (Iyengar et al. 2015). For example, consumers interested in the services of a telecom provider generally commit to a long-term contract and therefore reduce the decision risk by gathering information from their social environment (Nitzan and Libai 2011; Haenlein 2013). on the other hand it is unlikely to find strong social effects for purchase decisions of cheap consumer goods, as consumers can learn from their own experiences against low informational costs. Third, there are some other less common social influence mechanisms that are context-specific. For example, *network effects* cause an increase in product/service value with an increase in adoption rate, e.g., the value of online social platforms increases with the number of users (imagine being the only Twitter user, it would get boring quite fast). Further, *competitive concerns* cause influence among firms, as failing to adopt new innovations that are adopted by the competition can lead to competitive disadvantages (Van den Bulte and Lilien 2001).

Thus, the relevance of social influence varies between industries, brands, and the type of consumer decision, e.g., adoption versus repurchases (Hahn et al. 1994). Self-evidently, marketing tools like influencer marketing, brand ambassadors, and referral campaigns will be much more efficient when there is a sufficient level of social influence among the target consumers. Therefore, studying the presence and nature of social influence in the relevant context is a critical step in the development of a successful social network strategy. We will discuss different measures of social influence and provide an example of the identification of social influence on a publicly available dataset later on in the chapter.

When social influence is present, a first step toward leveraging these social mechanisms is to understand the role of hubs, also commonly referred to as influencers or opinion leaders. Typically, within a network, there are certain

consumers that exert a greater influence (e.g., Goldenberg et al. 2009). First, these consumers might have a higher level of persuasiveness because they are considered experts, because they represent a desirable image, or because of a combination of both. As it is sometimes ambiguous to identify these features from pure network data, additional survey data can provide a solution (Iyengar et al. 2011). Second, consumers might exert a greater influence because of their reach. Obviously, a consumer with many connections can spread information quicker to a broad audience than consumers with a small number of connections (Hinz et al. 2011). However, besides the number of connections, a consumer's position within a network is at least as important (Burt 2004; Granovetter 1983). Consumers that are connected to other well-connected consumers can help to spread information even faster. Understanding the role that such consumers play in the working of social effects gives marketers valuable insights about whom to target first with their marketing campaigns.

Even in the absence of social influence, network analyses can be used to generate valuable consumer insights. For example, it is well known that consumers that are more alike tend to be clustered within a network (Aral et al. 2009). As such, different clusters within the network might differ in preferences, in behavior, or in their attitude toward the brand (chapter ► “Market Segmentation” in this Handbook). Identifying the value and preferences of consumers within a cluster reveals information about the potential to acquire the consumer and helps to predict whether the consumer will have a high customer lifetime value (chapter ► “Modeling Customer Lifetime Value, Retention, and Churn” in this Handbook; Haenlein and Libai 2013). Further, identifying the overall attitude of a network segment is especially important for managers that seek to manage positive and negative word-of-mouth on social media channels (Homburg et al. 2015). However, while on the one hand homophily can be an important driver of social influence, the clustering of preferences and attitudes within a network also causes severe challenges in the identification of social influence. We will highlight these benefits and challenges further throughout the chapter. Next, we will first introduce some basic network metrics that help marketers to better understand the network that they aim to analyze.

Network Metrics

A social network is made up of two components, i.e., actors and ties. A social network is defined as the set of actors and the ties between them. As terminology differs across research disciplines, actors are also referred to as individuals, nodes, vertices, agents, players, and in the marketing field typically as consumers. Commonly used alternatives to the term tie are dyad, link, edge, and relationship. We will use these terms interchangeably throughout this chapter.

Social networks are a subset of all possible networks and have the unique characteristic that the actors are human beings. Examples of other networks are infrastructure networks (e.g., actors: train stations, ties: railroads) or the Internet (e.g., actors: websites, ties: links). This observation illustrates the interdisciplinary

nature of the social network field. Theories and models to guide the use of social network analysis in marketing could come from a broad set of research fields, such as logistics (i.e., understanding the flow of products through infrastructure networks), computer science (i.e., understanding information flow and the ranking of search results on the Internet), sociology (i.e., fundamental theories on social interactions), and medicine (i.e., understanding the spread of viruses through a population) among others.

We distinguish two types of social network analysis, namely, the analysis of the evolution of the structure of a network and the analysis of behavior and information flow within an existing network. In this chapter, we focus on the latter, because it facilitates the discussion of basic network characteristics and analyses. There is an interesting body of literature on the evolution of networks (e.g., Snijders et al. 2010) which is beyond the scope of this chapter as most of the marketing studies have used static networks only.

Network

A first step in social network analysis is describing a social network by means of simple metrics. We can measure networks at three different levels, the actor level, the tie level, and the network level. These metrics provide information on characteristics of the actors, ties, and the network as a whole, respectively. They enable marketers to get a first impression of which consumers might be influential and which relationships are likely to be crucial for the diffusion of innovations. In the following paragraphs, we introduce basic network notation and a selection of network metrics, their formulas and corresponding examples on how to calculate and interpret these metrics. We provide only a selection, because the complete set of metrics described in the literature is large and once the intuition on this type of metrics is clear, it should be easy to find and calculate the metric that is particularly relevant to answer a specific research question.

Basic Notation

We use notation that is widely used by others, e.g., Jackson (2010). We define a set of actors: $N = \{1, \dots, i, \dots, j, \dots, n\}$ and a $n \times n$ matrix A where each element of this matrix, A_{ij} , represents a tie between actors i and j . In the simplest case, A is a binary matrix where $A_{ij} = 1$ if there is a tie between actors i and j , and $A_{ij} = 0$ otherwise. We refer to A_{ij} as an adjacency matrix. Instead of just the presence or absence of a tie, it is also possible to indicate the strength of the tie between two actors by any real number for A_{ij} . We then refer to this matrix as a weight matrix. In both cases, we denote the resulting network or graph as (N, A) . To illustrate several metrics, we use an example network, Fig. 1a, based on the adjacency matrix in Fig. 1b. We assume that the distance between all adjacent pairs is 1.

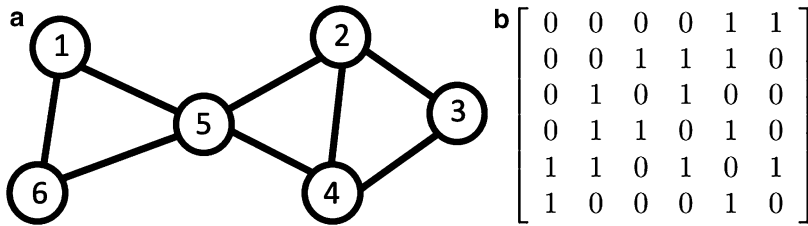


Fig. 1 (a) Example network, (b) Adjacency matrix of the example network

Actor Level

We start our discussion on metrics at the actor level. These metrics capture individual characteristics and therefore complement the set of consumer-level variables typically used in marketing, such as socio-demographic or service usage data. These so-called centrality measures each relate to a particular role of an individual in the network, such as influencer or information broker. Marketers can use this additional individual level information to enrich their customer segmentations and prediction models for behaviors such as product adoption and churn.

Degree Centrality

The most basic and most commonly used centrality measure is degree centrality. It is simply the number of ties an actor has, for example, the number of friends on Facebook or the number of connections on LinkedIn. Those with an extremely high degree centrality are called (social) hubs and are generally seen as influential individuals.

To calculate degree centrality one only accounts for the presence or absence of a tie and ignores the potential direction of the relationship. However, in many real networks, relationships are generally asymmetric or only go in one direction. For example, on Twitter you may be a follower of person j (the tie goes in the direction from i to person j , $i \rightarrow j$), but person j may not be following you (there is no tie going back from person j to i). Undirected ties are graphically represented by either a line $i-j$ or double arrows $i \rightleftarrows j$. An undirected network is formalized by a symmetric adjacency matrix while a directed network is formalized by an asymmetric adjacency matrix. To account for this directionality, we use the metrics in-degree centrality and out-degree centrality. In-degree centrality measures the number of incoming ties and out-degree measures the number of outgoing ties.

Assuming an undirected network in which $A_{ij} = A_{ji}$ for all i and j , you calculate degree centrality of actor i , $d_i(A)$, as:

$$d_i(A) = \#\{ j : A_{ij} = 1 \}$$

In the case of a directed network, we define in-degree centrality, $d_{i,in}$, and out-degree, $d_{i,out}$, as follows:

$$d_{i,in}(A) = \#\{ j : A_{ji} = 1 \}$$

$$d_{i,out}(A) = \#\{ j : A_{ij} = 1 \}$$

Some authors divide this measure by the maximum number of ties ($n-1$) to obtain values between 0 and 1. This standardization accounts for the size of the network and facilitates comparison of degree centralities across networks. The standardized version is most insightful in small networks where having ties to all other actors would be feasible, e.g., a class in school or a sports club, and you would observe values in the entire (0, 1) range. Table 1 shows the degree centralities of our example network.

Betweenness Centrality

The second, slightly more complex centrality measure that we discuss is betweenness centrality. This measure is based on the number of times an actor occurs on the shortest path between any pair of actors in a network. To calculate betweenness centrality, you first need to determine the shortest paths between all possible pairs of actors and then count how many times an actor appears on each of these paths. The difficulty here is that there may be multiple shortest paths and an actor might be on none, some, or all of these shortest paths.

A high betweenness centrality implies that a large amount of information flows “through” an actor on its way from sender to receiver. This gives the actor a high level of influence over which and how much information to pass on through the network and thus provides the actor with brokerage power (Burt 2004).

We define the number of shortest paths between actors j and k as $P(kj)$ and the number of shortest paths between j and k that go through actor i as $P_i(kj)$. Then, the ratio $P_i(kj)/P(kj)$ indicates how crucial actor i is in the relationship between actors j and k . The betweenness centrality of actor i is as follows:

$$BC_i = \sum_{k \neq j, i \text{ not in } \{k,j\}} P_i(kj)/P(kj)$$

Some authors divide this measure by the total number of pairs of actors ($n-1$) ($(n-2)/2$) to obtain average values. The latter would facilitate comparison across networks, but other than that the measures are equally useful. Table 1 shows the betweenness centralities of our example network.

Table 1 Actor level metrics of our example network

Actor	Degree centrality	Betweenness centrality	Closeness centrality	Eigenvector centrality
1	2	0	1.8	0.54
2	3	1.5	1.4	0.88
3	2	0	2	0.62
4	3	1.5	1.4	0.88
5	4	6	1.2	1.00
6	2	0	1.8	0.54

Closeness Centrality

Closeness centrality is the average distance between an actor and all other actors in the network. Distance could reflect the number of steps between two actors (1: friend, 2: friend of a friend, etc.), or it could be a weighted version of the number of steps by, for example, tie strength. It indicates quite literally how close everyone else in the network is to an actor and thus how much effort it would take to reach all others. Using the length of the shortest path $l(i,j)$ between actor i and any other actor j , the closeness centrality of actor i is:

$$CC_i = \sum_{j \neq i} l(i,j) / (n - 1)$$

Some authors take the inverse of this metric such that higher values of closeness centrality correspond to shorter paths and thus increased closeness. Which one to choose is a matter of conceptual preference and/or convention in a research domain. Table 1 shows the degree centralities of our example network.

Eigenvector Centrality

The previous three metrics reflect how central an actor is directly based on the actor's characteristics. Eigenvector centrality on the other hand is more complex and indirect in that it reflects how central an actor is based on how central or well-connected the actor's neighbors are. This metric builds on the notion that an actor can be considered central even if she/he only has a limited number of connections (i.e., low degree centrality) when these actors in turn are well-connected. Calculating this metric by hand is rather complex, because it is self-referential: the eigenvector centrality of actor i partly depends on the eigenvector centrality of his/her neighboring actor j , but the eigenvector centrality of this actor j again depends partly on the eigenvector centrality of actor i . Table 1 shows the eigenvector centralities of our example network.

Tie Level

We now move to the metrics on tie level. We distinguish two types of tie level characteristics, (1) those that measure a characteristic of the tie itself (e.g., strength or direction) and (2) those that measure similarities and differences between the two actors forming the tie (e.g., homophily). These tie characteristics provide a lot of information in addition to the actor level variables. They capture the social context in which the actor operates and allow you to put the actor characteristics in perspective. For example, for contagion purposes it might matter whether a female consumer is mainly connected to other males or females and whether her ties with others are generally strong or weak.

Tie Strength

Measuring tie strength is less straightforward than measuring the actor characteristics from the previous section. Following Granovetter, we define tie strength as “a

combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie” (Granovetter 1973, p. 1361). The challenge is how to operationalize this definition, because many social network datasets lack information on one or several of the dimensions from the definition. Most papers in the marketing literature use a proxy for tie strength based on interaction intensity, see Table 2.

Homophily

Homophily refers to the phenomenon that we tend to form relationships with others that are in some way similar to us. A well-known expression for this phenomenon is “birds of a feather flock together” (McPherson et al. 2001). When analyzing social network data, homophily is important to take into account, because otherwise observed similarities in behavior of connected individuals might erroneously be attributed to social influence or contagion (Aral et al. 2009; Manski 2000). The reason for this is that connected individuals tend to be similar and similar individuals with similar preferences tend to behave similarly regardless of any social influence occurring between them. Only after adjusting your analysis for homophily one can infer the impact of social influence properly.

Network Level

Metrics on the network level provide information on the network as a whole. They provide the context in which the lower level (actor, tie) action takes place. The role of these network level metrics is similar to the role of descriptive statistics for cross-sectional data: it gives you a good first impression of the information contained in the data at a glimpse.

Size

The most basic network level metric is the size of the network, i.e., the number of actors. Like the sample size in cross-sectional data, it is an important metric to report as it helps to calculate all other lower-level network metrics.

Table 2 Operationalizations of tie strength

<i>Data source</i>	<i>Papers</i>	<i>Tie strength measure</i>
Call detail records (telecom)	Haenlein (2013) Meyners et al. (2017) Nitzan and Libai (2011) Onnela et al. (2007) Risselada et al. (2014)	Aggregated duration of calls between users A and B over a reciprocated tie (relative to an actor’s total calling duration).
Direct messaging network	Aral et al. (2009)	Number/fraction of exchanged messages.

Density

Network density reflects the proportion all existing ties in a network to all possible ties in a network. It is closely related to the actor level variable degree centrality discussed above. You calculate the density of a network by dividing the average degree centrality of all actors by $(n-1)$, with n the number of actors in the network. It tells you how much of the network potential is actually used.

Degree Distribution

The degree distribution of a network is the relative frequency of actors with a certain degree. The distribution is therefore not a metric but rather a feature of a social network. Many social networks have a degree distribution which is highly skewed, where many actors have only a few ties, and only a few actors have many ties. This phenomenon is caused by preferential attachment, i.e., actors typically want to connect to popular (high-degree) actors. A common term for networks with such properties is scale-free networks. On the contrary, when the degree distribution follows a normal distribution, the network is said to be random, i.e., the likelihood that actors connect is independent of their degree. See Barabasi and Bonabeau (2003) and Broido and Clauset (2019) for more information about the properties of common real-world networks.

Network Data and Sampling Methods

Scaling down to networks that are relevant for marketers, there are still numerous networks that differ substantially in the nature of ties and actors, and the structure of the network. Networks that are often analyzed in marketing include product networks (for recommendation systems), networks of companies (for B2B marketing), organizational networks (for social influence among employees), and market networks (for buyer–seller relationships). As before, we focus on networks of consumers, as this is the most common network type in marketing research and practice. A large collection of different types of networks is provided by the Colorado Index of Complex Networks (<https://icon.colorado.edu/>).

Data Collection

To gather insights in consumer networks, there are multiple data solutions. Table 3 illustrates the types of online and offline network data and their benefits and disadvantages. A solution that is as straightforward as it is effective is the use of geographic proximity (Wuyts et al. 2011; Meyners et al. 2017). Naturally, consumers that live close to each other tend to interact with each other. In addition, local social influence effects can be quite strong due to the high level of perceived similarity between consumers that live in the same area (i.e., perceived homophily). For example, Nam et al. (2010) estimate the effect of social influence on

Table 3 Types of network data with selection of papers

Type of network data	Benefits	Disadvantages	Examples within the marketing literature
Geographic proximity	Easy to obtain data Highly scalable	Low accuracy No information about the individual ties No detailed information about homophily	Bell and Song (2007) Nam et al. (2010) Choi et al. (2010)
Survey data	High accuracy Complete network Tie-strength info Information about nature of relationship Detailed information about homophily	Hard to obtain data Not highly scalable Self-report bias	Iyengar et al. (2011) Iyengar et al. (2015) Nair et al. (2010)
CDR data	High accuracy Highly scalable Tie-strength info Some information about homophily	Privacy issues No information about nature of relationship	Nitzan and Libai (2011) Onnela et al. (2007) Risselada et al. (2014)
Social network platform	Medium hard to obtain data Highly scalable Tie-strength info Detailed information about homophily Information about WOM content	Limited data access through APIs Strong sampling bias dependent on type of platform	Ma et al. (2015) Trusov et al. (2009) Trusov et al. (2010) Aral and Walker (2014) Valsesia et al. (2020)
Instant messenger	Highly scalable Tie-strength info	Hard to obtain data Little information about nature of relationship	Aral et al. (2009)
Community networks	High accuracy Detailed information about homophily Information about WOM content	Only information about current customers	Zhang and Godes (2018) Park et al. (2018)

the adoption of a video-on-demand service by measuring how local quality differences drive the adoption within the same geographical area. The benefit of geographical data is that it is relatively easy and cheap to obtain for a large group of consumers. However, the data is often an imperfect representation of the true network and does not allow for detailed information about the ties, such as tie strength and homophily. A second method that requires more effort is the use of surveys (chapter ► [“Crafting Survey Research: A Systematic Process for Conducting Survey Research”](#) in this Handbook). The benefit of survey data is that one can measure the more detailed tie attributes such as tie-strength and homophily. For example, Iyengar et al. (2015) are able to distinguish individuals that are seen as experts from individuals that are seen as discussion partners. Besides

the high costs of collecting data, researchers should be aware of possible self-report biases. For example, friendships are not always reciprocal, and the number of friends or the own social status can be overestimated (Wuyts et al. 2011; Haenlein 2013). To generate a connected network from survey data, one can engage in snowball sampling or the so-called referral chain sampling (Ebbes et al. 2016; Reingen and Kernan 1986). This type of sampling refers to a specific type of survey distribution, in which the surveyed consumers forward the questionnaire to his/her connections. As such, one can obtain a network in which each node is connected with the entire network. A third technique that is widely applied in marketing is to leverage third party data of a telecom provider. The so-called “call detail records” (CDR) data contains all phone calls of the customers of the telecom company. Based on these calls, one can derive the relevant network of the customer. For example, Nitzan and Libai (2011) identify the effect of social influence on churn behavior from such a dataset. Studies show that such networks accurately represent offline networks (Eagle et al. 2009). Features such as the weight of a tie can be estimated by taking the frequency and the duration of calls into account (Risselada et al. 2014). However, little information about the nature of the relationship is known (e.g., friends or colleagues?) and the rising privacy regulations make it more difficult to obtain CDR data. A fourth method is to analyze the network from a social network platform such as Twitter or LinkedIn. Often such data can be gathered through an API provided by the platform and gives a good indication of the individuals network, see <https://developer.twitter.com/en/docs> for APIs that can be used to download data from Twitter and see <https://www.linkedin.com/developers/> for APIs that can be used to download data from LinkedIn. When the platform also tracks features such as individual messaging, or interactions with user-generated content, one can use these as tie strength measures. In addition, these platforms often gather demographic and behavioral data of their users, which allows for precise homophily measures between the nodes. An example of the use of such data is the study by Valsesia et al. (2020), which shows that the number of out-degree ties, conditional on the number of in-degree ties, has a negative effect on the perceived and actual influence of social media users. Similar data can be derived from online communication tools, such as networks based on e-mail traffic or online message services. While the true nature of the ties is often unobserved in such networks, they are able to provide a sufficient proxy for the offline network of consumers. However, not every online network is a representation of the offline network. Typically, the boundary between friends and strangers deteriorates when moving from an offline to an online network. Nevertheless, online networks are highly relevant for marketers as consumers can be influenced by both close friends as well as by acquaintances or complete strangers (Zhang et al. 2015). Finally, other typical online networks are community networks. Firms create community networks to foster shared consumption experiences, collaboration, or competition between their customers. In addition, it provides their customers with a single point of concentration information about the products or services of a firm. For example, analyzing the network of an online gaming community, Park et al. (2018) identify a positive effect of social contagion on users’ spending behavior.

Network Sampling

Consumer networks can include millions of nodes and the square number of ties resulting in large and complex data. As such, to map out and analyze the entire network is a computationally expensive task. Even though some of the techniques to model networks discussed later on are scalable, many methods are limited in their capacity and require a lot of computational power. To reduce the size and complexity of the network, researchers can apply several sampling techniques. These techniques differ in their ability to recover the different network characteristics. As such, their suitability depends on the goal of the researchers.

When the goal of the analyses is to identify the impact of social influence on consumer behavior at an individual level, researchers can rely on random sampling. To measure social influence, the data needs to include (1) the behavior of an individual node, (2) the nodes that are connected to the focal node, and (3) the behavior of the connected nodes. As such, the overall network structure is less relevant, and the interest focuses mainly on the ego-network, i., the focal actor (“ego”) and the nodes that are directly connected to the actor (e.g., Risselada et al. 2014; Nitzan and Libai 2011; Haenlein 2013). Such direct connections are typically referred to as first-degree neighbors. Scholars that are interested in social influence across multiple nodes or on a global level can expand the ego-network to include second- or higher-degree neighbors. Figure 2 illustrates the relevant ego network versus the complete network.

As soon as marketers are interested in social influence at the global level or want to predict the results of possible marketing initiatives that leverage social influence, network sampling techniques are required. To derive a representative sample of the network, it is important to recover important network characteristics such as betweenness and closeness centrality, and the degree distribution. This can be achieved through the so-called subgraph sampling methods. These methods differ

Fig. 2 Ego network (gray = ego) versus complete network



from the random sampling described above. In the procedure of collecting data from ego-networks, we first select a group of random nodes and subsequently select their links. In subgraph sampling, we sample the nodes and links jointly. There are four widely applied network sampling methods to derive a subgraph from a population network, i.e., the random sampling method, the snowball method, the random walk method, and the forest fire method. To illustrate the different methods, we define a population network as $G = (V, E)$, with the set of actors $V = \{v_1, \dots, v_N\}$ and the set ties $E = \{(v_i, v_j)\}$. We define a subgraph of G as $G^* = (V^*, E^*)$, where the sampled actor set and tie set of graph G^* are $V^* \subseteq V$ and $E^* \subseteq E$. In the case of random sampling, we take a random selection of actors, V^* , and then include all ties between the actors, E^* , to build the subgraph $G^* = (V^*, E^*)$. For snowball sampling, we start with selecting one actor, v_1^* , then select all neighbors, v_2^* , and select all their unselected neighbors, v_3^* , up to v_k^* , until we've reached a large enough set of actors $V^* \{v_1^*, \dots, v_k^*\}$. In the case of the random walk method, we only select one neighbor at random from the entire set of unselected neighbors. In the case of the forest fire method, we select a certain percentage (i.e., the burn rate) of the remaining unselected neighbors at random for each round. Ebbes et al. (2016) compare the performance of all methods and conclude that (1) forest-fire sampling with a burn-rate around 50% should be used in research on local influence, as this method is best in recovering the degree distribution of the graph, and (2) the random walk method or forest-fire sampling with a low burn-rate (e.g., 20%) should be used for research on influence at a network level as it is best in retrieving the centrality measures.

Social Network Analysis in R

In this section, we illustrate how to calculate the metrics discussed above and how to estimate a basic model to quantify social influence. The main package we use for the network-related analyses is the *igraph* package (Csardi and Nepusz 2006) in R (R Core Team 2018). This package is also available for Python.

Data

We use the classic Coleman's Drug Adoption dataset "Innovation among Physicians," which is publicly available in the *spatialprobit* R-package (Wilhelm and de Matos 2015). You can find a detailed description of the dataset in the package by typing? CKM in the R console after loading the *spatialprobit* package.

```
> library(spatialprobit) #load the spatialprobit package
> ?CKM #this calls for the help on the CKM dataset
```

The dataset contains information on 246 physicians in four cities and was collected in 1966. Table 4 shows the variables we use in our examples.

To be able to work with the dataset, you need to load it in R.

Table 4 Variable names and descriptions as given in the *spatialprobit* package

Variable name	Description
city	a numeric vector; City: 1 Peoria, 2 Bloomington, 3 Quincy, 4 Galesburg
adoption.date	an ordered factor with levels November, 1953; December, 1953; January, 1954; February, 1954; March, 1954; April, 1954; May, 1954; June, 1954; July, 1954; August, 1954; September, 1954; October, 1954; November, 1954; December, 1954; December/January, 1954/1955; January/February, 1955; February, 1955; no prescriptions found; no prescription data obtained
med_sch_yr	Years in practice
friends	friends
community	Time in the community
specialty	Medical specialty

```
> data(CKM) #load the dataset
```

We use the data tables format in R to make the data manipulation and coding as easy as possible. See <https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html> for more information regarding the data table format and the *data.table* package (Dowle and Srinivasan 2019).

```
> library(data.table) #load the data.table package
> CKM <- as.data.table(CKM) #make CKM a data.table object
> CKM[,date := as.numeric(adoption.date)] #change the variable
type of adoption.date to numeric for ease of analysis
```

We add a new variable called *id* to have a clear identifier for all individuals in the dataset.

```
> CKM[,id := 1:.N] #assign IDs to all individuals in the dataset
```

Now we are all set to create the network. Here we use an adjacency matrix approach where each element of the matrix simply indicates absence (0) or presence (1) of a tie. In this example, we base the adjacency matrix on three matrices for which the original authors collected data by means of a sociometric approach. The three relevant questions were “When you need information or advice about questions of therapy where do you usually turn?”, “Who are the three or four physicians with whom you most often find yourself discussing cases or therapy in the course of an ordinary week – last week for instance?”, and “Would you tell me the first names of your three friends whom you see most often socially?”. These three questions resulted in an advice matrix (A1), a discussion matrix (A2), and a friend matrix (A3), respectively. For illustrative purposes, we combine the three matrices in a unique adjacency matrix A4 by setting its elements to 1 if it was 1 in A1, A2, or A3.

```
> A4 <- A1 + A2 + A3 #add up the advice, discussion, and friend
matrices
```



```
> A4[A4 > 0] <- 1 #create a binary adjacency matrix
> library(igraph) #load the igraph package
> medinnovNetw <- graph_from_adjacency_matrix(A4) #create the
network
```

To get a first impression of the network you can use the *plot* function to generate a simple network plot. This will only provide useful output for smaller networks. Large networks will be messy or even unreadable. Visualizing large and complex networks requires specialized software, e.g., Gephi (<https://gephi.org>).

Calculating Actor Level Metrics

Figure 3 shows that the network is not fully connected. It consists of several disconnected clusters. For illustrative purposes we will use the largest cluster and calculate the actor level metrics discussed above. The reason for this is that a metric

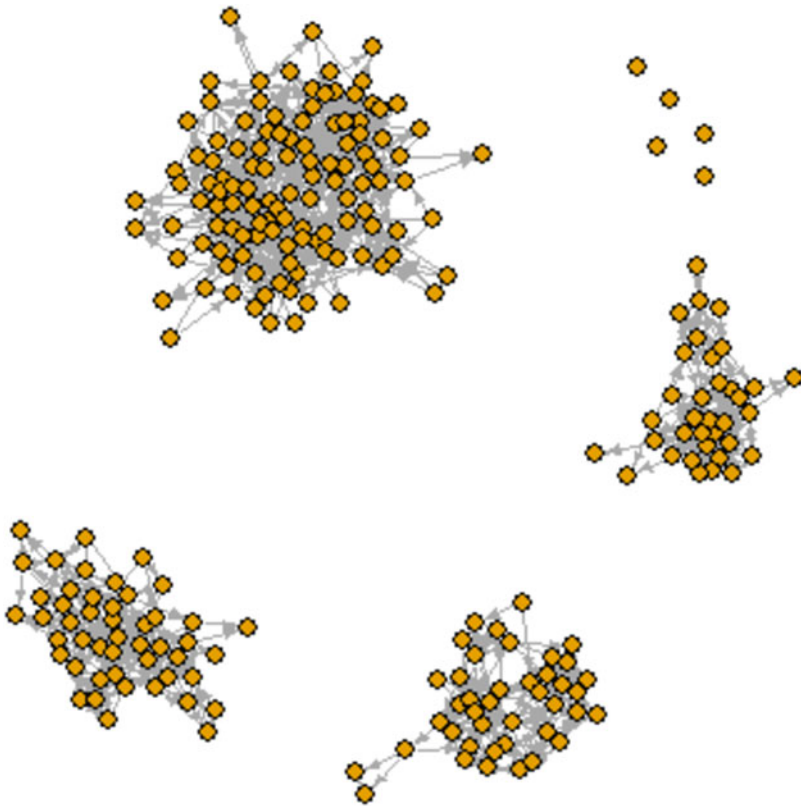


Fig. 3 Plot of the network. (Notes: generated by `plot(medinnovNetw, vertex.size = 5, edge.arrow.size = 0.3, vertex.label = "")`)

like closeness centrality cannot be calculated for disconnected parts of the network, because the distance between two actors in different clusters would be infinite.

```
> medinnovNetw.components <- clusters(medinnovNetw) #identify the
clusters in the network; the largest cluster consists of 117 actors
> medinnovNetw.components.graphs <- decompose.graph(medinnovNetw)
#decompose the network into separate clusters; each cluster is an
element in a list
> medinnovNetw.subgraph1 <- as.undirected(medinnovNetw.
components.graphs[[1]]) #we store the largest cluster as an
undirected network for illustrative purposes
> gorder(medinnovNetw.subgraph1) #check whether the subgraph is
indeed the one with 117 vertices
```

Now that we have our new network, we can easily calculate all network metrics using standard functions in the *igraph* package. We show the code here and provide summary statistics in Table 5.

```
> subgraph1.degree.centralty <- degree(medinnovNetw.subgraph1)
> subgraph1.betweenness.centralty <- betweenness(medinnovNetw.
subgraph1)
> summary(subgraph1.closeness.centralty)
> subgraph1.eigenvector.centralty <- eigen_centralty
(medinnovNetw.subgraph1)#the first element of this list contains the
centralities
```

Calculating Tie Level Metrics

We now switch back to using the complete dataset. Most papers using communication-based network data, for example, call detail records or online social network data, use interaction intensity as a proxy for tie strength. However, the dataset that we use here is not based on communication, but on survey responses and the network is based on sociometric questions related to advice, discussion, and friendship. We used the binary adjacency matrices to construct the network and we can use the weight matrices derived from the adjacency matrices (and included as W1, W2, and W3 in the CKM data) as our tie strength measure. The weight matrices are just the normalized versions of the adjacency matrices. For example, if physician 1 indicated

Table 5 Descriptive statistics of the actor level metrics

Actor level metric	Mean	S.D.	Min.	Max.
Degree centrality	7.966	4.026	2.000	26.000
Betweenness centrality	91.880	99.114	0.000	717.120
Closeness centrality	0.003	0.0003	0.003	0.004
Eigenvector centrality	0.262	0.153	0.031	1.000

7 friends, the weight of each of those friends would be $1/7$. The implicit assumption here would be that the more friends you have, the weaker the tie per friend.

The variables in the dataset that we use to calculate homophily are the city (*city*), years in practice (*med_sch_yr*), time in the community (*community*), and medical specialty (*specialty*). Before we can perform these calculations, we need to prepare the data. Note that we continue with the adjacency matrix A4 that we created earlier.

```
> dimnames(A4) <- list(as.character(CKM$id), as.character(CKM$id))
> g <- graph_from_adjacency_matrix(A4)
> g.list <- data.table(get.edgelist(g)) #this returns a data table
with all edges (relationships)
> friends <- g.list[,list(id=as.numeric(V1), friend=as.numeric
(V2))] #set names
> friends <- rbind(friends, friends[,list(id=friend, friend=id)])
#the original data is directed, but here we make an edgelist for an
undirected graph
> friends <- merge(friends,
+                 CKM[,list(friend=id,
+                           d.adopt=date,
+                           f.city=city,
+                           f.med_sch_yr=med_sch_yr,
+                           f.community=community,
+                           f.specialty=specialty)],
+                 by="friend") #we add the adoption dates and the
variables we need to calculate homophily later for the friends
> friends[,f.adopt := 1]
> friends[d.adopt > 17, f.adopt := 0] #in these two steps we create
an adoption dummy and set it equal to zero when the adoption month
is larger than 17 (i.e. the end of the observation period)
> friends2 <- merge(friends,
+                 CKM[,list(id =id,
+                           id.city=city,
+                           id.med_sch_yr=med_sch_yr,
+                           id.community=community,
+                           id.specialty=specialty)],
+                 by="id") #we add the variables we need to calculate
homophily later for the ids
> setkey(friends, id) #we create a key to sort the data and speed up
the data manipulation
```

To check whether we got the desired results we use the `head()` function to display the first few rows of the dataframe, see Fig. 4. In the table `friends2` we see the

```
> head(friends2)
  id friend d.adopt f.city f.med_sch_yr f.community f.specialty f.adopt id.city id.med_sch_yr id.community id.specialty
1: 1     8       3     1       4           5           1     1     1           2           6           3
2: 1    58       4     1       2           6           4     1     1           2           6           3
3: 1    78       1     1       3           5           3     1     1           2           6           3
4: 1    87      19     1       6           2           4     0     1           2           6           3
5: 1    90      19     1       2           3           4     0     1           2           6           3
6: 1   110      20     1       9           2           4     0     1           2           6           3
```

Fig. 4 The result of `head(friends2)`

two physician identifiers (*id*, *friend*) forming a tie, the adoption date (*d.adopt*) and dummy (*f.adopt*) of the friend, and the variables we use for the homophily calculation of both *id* and *friend*.

We can now calculate the homophily variable in two steps. First, we create a new variable per dimension attaching a weight of $\frac{1}{4}$ when *id* and *friend* have the same value for that dimension. We add the conditions that end with “ $!= 9$ ” to exclude the cases where the respondent indicated “no answer,” because two individuals not answering does not make them more similar. In the second step we add all the weights to get a homophily score per *id*, the variable *HOMOPH*.

```
> friends3 <- friends2 %>%
+   mutate(
+     city.hom = ifelse(id.city == f.city, 1/4, 0),
+     med_sch_yr.hom = ifelse(id.med_sch_yr == f.med_sch_yr & id.
med_sch_yr != 9, 1/4, 0),
+     community.hom = ifelse(id.community == f.community & id.
community != 9, 1/4, 0),
+     specialty.hom = ifelse(id.specialty == f.specialty & id.
specialty != 9, 1/4, 0),
+     HOMOPH = city.hom + med_sch_yr.hom + community.hom +
specialty.hom
+   )
```

Modeling Social Contagion

Assessing the relationship between the adoption by the *id* and the adoption(s) by his/her friend(s) is our main objective here. We start by creating a clean version of the original CKM dataset.

```
> CKM <- CKM[date < 19] #remove all missing values on adoption
(date == 19 or 20)
> CKM[,adoption := 1] #we create an adoption dummy
> CKM[date == 18,adoption := 0] #set it to 0 if id did not adopt
(date == 18)
> CKM <- CKM[discuss != 9 & friends != 9] #remove if no answer on
friends or discussion
```

Then we create a panel dataset with a unique row per *id*-month combination. The last month per *id* is the month in which the *id* adopted (i.e., *adoption* = 1). If *id* did not adopt at all during the observation period the maximum number of rows is 17 where *adoption* = 0 for all rows. The time-independent covariates are the same in every row per *id*. Figure 5 below shows the top of the created panel dataset CKM.panel.

```
> CKM.panel <- CKM[,list(month=seq(from=1,to=17,by=1)),by=id]
> CKM.panel <- merge(CKM.panel,CKM[,list(id,date,jours,patients,
med_sch_yr,specialty)],all.x=T,by="id")
```

id	month	date	jours	patients	med_sch_yr	specialty	adoption
1	1	1	8	5	2	3	1
2	1	12	4	4	6	1	0
2	2	12	4	4	6	1	0
2	3	12	4	4	6	1	0
2	4	12	4	4	6	1	0
2	5	12	4	4	6	1	0

Fig. 5 First six rows of the CKM.panel dataset

```
> CKM.panel[,adoption := 0]
> CKM.panel[date == month, adoption := 1]
> CKM.panel <- CKM.panel[month <= date]
```

We need to create another dataset based on the friends dataset we created earlier in order to be able to link the adoptions by friends to the adoption by id in a certain month. The code below sums up the number of adoptions (`f.adopt = sum(f.adopt)`) per id-date (`by = c("id", "d.adopt")`) combination and stores this in a new dataset called friends.adopt.

```
> friends.adopt <- friends[,list(f.adopt=sum(f.adopt)),by=c
("id","d.adopt")]
> setnames(friends.adopt,"d.adopt","month")
```

We can now merge the friends.adopt dataset with the CKM.panel to obtain the analysis set.

```
> CKM.panel <- merge(CKM.panel,friends.adopt,all.x=T,by=c
("id","month"))
> CKM.panel[is.na(f.adopt),f.adopt := 0]
> CKM.panel[,c.adopt := cumsum(f.adopt),by=id]
```

Our social contagion model simply regresses adoption by the id in a certain month to the number of adoptions of friends until and including that month (chapter ▶ [“Regression Analysis”](#) in this Handbook). Figure 6 shows the results. The parameter of the social influence variable is positive and significant ($\beta = 0.367, p < 0.001$), which implies that the likelihood of adoption by an individual is greater when the number of friends who already adopted is larger.

A Word of Caution

To identify social influence is not an easy task. For example, the model above is fairly simple and misses many important variables. To increase the causal evidence, we could include the homophily variable (HOMOPH) that we created earlier in our social contagion model. One way to do this is by using the homophily variable as a

```

> summary(glm(adoption ~ f.adopt ,data=CKM.panel,family=binomial("cloglog")))

Call:
glm(formula = adoption ~ f.adopt, family = binomial("cloglog"),
    data = CKM.panel)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1429  -0.4567  -0.4567  -0.4567   2.1505

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2607     0.1195 -18.920 < 2e-16 ***
f.adopt       0.3670     0.1089   3.368 0.000757 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 595.09  on 831  degrees of freedom
Residual deviance: 585.13  on 830  degrees of freedom
AIC: 589.13

```

Fig. 6 R code and estimation results of our social contagion model

weight for the adoptions (e.g., Risselada et al. 2014). Another refinement would be to take the recency of adoptions into account where adoptions of more than a certain number of periods ago would no longer contribute to the social influence variable. Alternatively, one could use a stock variable approach which is commonly for advertising expenditures in sales response models. Further, scholars have shown that controlling for marketing initiatives removes the evidence of social influence that we found in the dataset above (van den Bulte and Lilien 2001).

However, when analyzing social influence based on network data, there are three additional challenges that can lead to a biased estimate of social influence, i.e., simultaneity, external shocks, and homophily (Manski 2000). Simultaneity or reflection arises when two consumers influence each other simultaneously. When this is the case, it remains unclear whether the actor is influenced by her/his connections, or whether the connections are influenced by the actor. To avoid a bias due to simultaneity, scholars can use a lagged variable of social influence. That is, the behavior of connections in the past can influence the current behavior of the actor, but not vice versa. External shocks refer to a change in the environmental conditions that can influence the behavior of consumers that are connected at the same time. For example, changes at a geographical local level might impact the behavior of multiple connected consumers. Finally, there is the issue of homophily. While we can use similarity measures based on observed characteristics to capture homophily and even use it as moderator of social influence, it is likely that we do not observe all relevant characteristics. That is, there are latent variables, such as unobserved consumer preferences, that both drive the network formation as well as the behavior of interest.

As such, observed clustered behavior within a network might not be the cause of social influence, but rather a result of the clustered unobserved preferences. Several solutions to this problem have been proposed, such as propensity score matching on observed variables (Aral et al. 2009), the use of fixed effects with longitudinal data (Park et al. 2018), the use of instrumental variables (Aral and Nicolaides 2017), or the use of specific network models such as the latent space model (Davin et al. 2013) or the spatial error model (Ansari et al. 2011). Currently, the problem of homophily and the mentioned solutions are still subject to an ongoing discussion (Shalizi and Thomas 2011). Both, the issue of external shocks and homophily are similar to the endogeneity problem caused by omitted variables (chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#) in this Handbook).

Conclusion

While the existence of social influence among consumers has been studied for decades, the increasing availability of network data and the tools to analyze these allows marketers to better understand and address consumers’ needs. However, collecting and analyzing network data brings new challenges to marketing researchers and practitioners. In this chapter we provide basic network theory and notation with references to key publications in the field. We also provide suggestions for software (packages) and useful functions including code snippets to support you in preparing and running your first social network analyses. Our aim was neither to provide a complete literature overview nor was it the aim to go into the most advanced social contagion models. We hope that this chapter is a good starting point for those willing to discover the exciting social network domain.

Cross-References

- [Crafting Survey Research: A Systematic Process for Conducting Survey Research](#)
- [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- [Market Segmentation](#)
- [Modeling Customer Lifetime Value, Retention, and Churn](#)
- [Regression Analysis](#)

References

- Ansari, A., Koenigsberg, O., & Stahl, F. (2011). Modeling multiple relationships in social networks. *Journal of Marketing Research*, 48(4), 713–728.
- Aral, S., & Nicolaides, C. (2017). Exercise contagion in a global social network. *Nature Communications*, 8(1), 1–8.
- Aral, S., & Walker, D. (2014). Tie strength, embeddedness, and social influence: A large-scale networked experiment. *Management Science*, 60(6), 1352–1370.

- Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51), 21544–21549.
- Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5), 60–69.
- Bell, D. R., & Song, S. (2007). Neighborhood effects and trial on the internet: Evidence from online grocery retailing. *Quantitative Marketing and Economics*, 5(4), 361–400.
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992–1026.
- Broido, A. D., & Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1), 1–10.
- Burnkrant, R. E., & Cousineau, A. (1975). Informational and normative social influence in buyer behavior. *Journal of Consumer Research*, 2(3), 206–215.
- Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2), 349–399.
- Choi, J., Hui, S. K., & Bell, D. R. (2010). Spatiotemporal analysis of imitation behavior across new buyers at an online grocery retailer. *Journal of Marketing Research*, 47(1), 75–89.
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 1–9.
- Davin, J. P., Gupta, S., & Piskorski, M. J. (2013). Separating homophily and peer influence with latent space. Available at SSRN 2373273.
- Dowle, M., & Srinivasan, A. (2019). data.table: Extension of `data.frame`. R package version 1.12.6. <https://CRAN.R-project.org/package=data.table>
- Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36), 15274–15278.
- Ebbes, P., Huang, Z., & Rangaswamy, A. (2016). Sampling designs for recovering local and global characteristics of social networks. *International Journal of Research in Marketing*, 33(3), 578–599.
- Goel, S., & Goldstein, D. G. (2014). Predicting individual behavior with social networks. *Marketing Science*, 33(1), 82–93.
- Goldenberg, J., Han, S., Lehmann, D. R., & Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of Marketing*, 73(2), 1–13.
- Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380.
- Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, 1, 201–233.
- Haenlein, M. (2013). Social interactions in customer churn decisions: The impact of relationship directionality. *International Journal of Research in Marketing*, 30(3), 236–248.
- Haenlein, M., & Libai, B. (2013). Targeting revenue leaders for a new product. *Journal of Marketing*, 77(3), 65–80.
- Hahn, M., Park, S., Krishnamurthi, L., & Zoltners, A. A. (1994). Analysis of new product diffusion using a four-segment trial-repeat model. *Marketing Science*, 13(3), 224–247.
- Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75(6), 55–71.
- Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, 52(5), 629–641.
- Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Rejoinder – Further reflections on studying social influence in new product diffusion. *Marketing Science*, 30(2), 230–232.
- Iyengar, R., Van Den Bulte, C., & Lee, J. Y. (2015). Social contagion in new product trial and repeat. *Marketing Science*, 34(3), 408–429.
- Jackson, M. O. (2010). *Social and economic networks*. Princeton: Princeton university press.
- Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., & Tillmanns, S. (2010). Undervalued or overvalued customers: Capturing total customer engagement value. *Journal of Service Research*, 13(3), 297–310.
- Landsman, V., & Nitzan, I. (2020). Cross-decision social effects in product adoption and defection decisions. *International Journal of Research in Marketing*, 37(2), 213–235.

- Ma, L., Sun, B., & Kekre, S. (2015). The squeaky wheel gets the grease – An empirical analysis of customer voice and firm intervention on twitter. *Marketing Science*, 34(5), 627–645.
- Manski, C. F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, 14(3), 115–136.
- McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444.
- Meyners, J., Barrot, C., Becker, J. U., & Goldenberg, J. (2017). The role of mere closeness: How geographic proximity affects social influence. *Journal of Marketing*, 81(5), 49–66.
- Nair, H., Manchanda, P., & Bhatia, T. (2010). Asymmetric social interactions in physician prescription behavior: The role of opinion leaders. *Journal of Marketing Research*, 47(5), 883–895.
- Nam, S., Manchanda, P., & Chintagunta, P. K. (2010). The effect of signal quality and contiguous word of mouth on customer acquisition for a video-on-demand service. *Marketing Science*, 29(4), 690–700.
- Nielsen. (2015). Global trust in advertising. Available online at <http://www.nielsen.com/us/en/insights/reports/2015/global-trust-in-advertising-2015.html>. Updated on 09-28-2015, checked on 5/3/2017.
- Nitzan, I., & Libai, B. (2011). Social effects on customer retention. *Journal of Marketing*, 75(6), 24–38.
- Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., et al. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18), 7332–7336.
- Park, E., Rishika, R., Janakiraman, R., Houston, M. B., & Yoo, B. (2018). Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing*, 82(1), 93–114.
- Peres, R., Muller, E., & Mahajan, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, 27(2), 91–106.
- R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>
- Reingen, P. H., & Kernan, J. B. (1986). Analysis of referral networks in marketing: Methods and illustration. *Journal of Marketing Research*, 23(4), 370–378.
- Risselada, H., Verhoef, P. C., & Bijmolt, T. H. A. (2014). Dynamic effects of social influence and direct marketing on the adoption of high-technology products. *Journal of Marketing*, 78(2), 52–68.
- Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2), 211–239.
- Snijders, T. A. B., van de Bunt, G. G., & Steglich, C. E. G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1), 44–60.
- Trusov, M., Bucklin, R. E., & Pauwels, K. H. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of Marketing*, 73(5), 90–102.
- Trusov, M., Bodapati, A. V., & Bucklin, R. E. (2010). Determining influential users in internet social networks. *Journal of Marketing Research*, 47(4), 643–658.
- Valente, T. W. (2012). Network interventions. *Science*, 337(6090), 49–53.
- Valesia, F., Proserpio, D., & Nunes, J. C. (2020). The positive effect of not following others on social media. *Journal of Marketing Research*. forthcoming.
- Van den Bulte, C., & Lilien, G. L. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*, 106(5), 1409–1435.
- Wilhelm, S., & de Matos, M. G. (2015). spatialprobit: Spatial Probit Models. R package version 0.9–11. <https://CRAN.R-project.org/package=satialprobit>.
- Wuyts, S. H. K., Dekimpe, M. G., Gijbrenchts, E., & Pieters, F. G. M. R. (2011). *The connected customer: The changing nature of consumer and business markets*. Routledge.
- Zhang, Y., & Godes, D. (2018). Learning from online social ties. *Marketing Science*, 37(3), 425–444.
- Zhang, J., Liu, Y., & Chen, Y. (2015). Social learning in networks of friends versus strangers. *Marketing Science*, 34(4), 573–589.



Bayesian Models

Thomas Otter

Contents

Introduction: Why Use Bayesian Models?	720
Bayesian Essentials	722
Bayesian Estimation	731
Examples of Posterior Distributions in Closed Form	732
Posterior Distributions Not in Closed Form	736
Model comparison	761
Numerical Illustrations	765
A Brief Note on Software Implementation	765
A Hierarchical Bayesian Multinomial Logit Model	766
Mediation Analysis: A Case for Bayesian Model Comparisons	769
Conclusion	771
Cross-References	771
Appendix	772
MCMC for Binomial Probit Without Data Augmentation	772
HB-Logit Example	776
References	778

Abstract

Bayesian models have become a mainstay in the tool set for marketing research in academia and industry practice. In this chapter, I discuss the advantages the Bayesian approach offers to researchers in marketing, the essential building blocks of a Bayesian model, Bayesian model comparison, and useful algorithmic approaches to fully Bayesian estimation. I show how to achieve feasible Bayesian inference to support marketing decisions under uncertainty using the Gibbs sampler, the Metropolis Hastings algorithm, and point to more recent developments – specifically the no-U-turn implementation of Hamiltonian Monte Carlo sampling available in Stan. The emphasis is on the development of an

T. Otter (✉)

Goethe University Frankfurt am Main, Frankfurt am Main, Germany

e-mail: otter@marketing.uni-frankfurt.de

appreciation of Bayesian inference techniques supported by references to implementations in the open source software R, and not on the discussion of individual models. The goal is to encourage researchers to formulate new, more complete, and useful prior structures that can be updated with data for better marketing decision support.

Keywords

Marketing decision-making · Bayesian inference · Gibbs sampling · Metropolis Hastings · Hamiltonian Monte Carlo · R · bayesm · Stan

Introduction: Why Use Bayesian Models?

Bayesian models have gained popularity over the past 30 years both among academics in marketing and marketing research practitioners. There are several reasons for this popularity. First, many marketing problems involve data in the form of relatively short panels but with many observational units (large N , small T). Each observational unit, e.g., a respondent, a customer, or a store supplies only a limited amount of data, but there are many observational units in the data set. In the vast majority of these applications, decision makers know a priori that observational units are heterogeneous in their underlying, at least partially unobserved characteristics that generated the data. And the successful marketing of differentiated goods that involves market segmentation, targeting and positioning requires measures of heterogeneity in the population of observational units. Estimating separate, independent models for each observational unit results in unreliable estimates, and in many applications, individual level time series are too sparse for individual level maximum likelihood estimates to be defined. Hierarchical Bayes models offer a convenient and practical solution to this problem.

Second, the overwhelming majority of marketing data sets involve so-called limited dependent variables, e.g., choices, ratings, rankings, or generally dependent variables that have strongly noncontinuous features. Although a number of non-Bayesian estimators are available for models with such dependent variables (see e.g., Amemiya 1985; Long 1997), the assessment of statistical uncertainty in estimates relies on large sample asymptotic arguments. In marketing, large samples that allow for inference based on asymptotic arguments are the exception, even in an era where big data has become a ubiquitous buzzword. Big data, by definition, involves large data sets. However, the size of the data set usually does not translate into more statistical information about individual target parameters. Big data are always “big” because of their dimensionality spanning across, e.g., tens of thousands of customers, products, and time points, and include a myriad of potentially useful conditioning arguments. The dimensionality of the data at the very source of its size, or “bigness,” regularly translates into similarly high-dimensional models and estimation problems, such that the amount of statistical information about individual target parameters is small yet again. Bayesian models allow for coherent inference

even in small samples, or more generally in situations where there is little data-based information about individual parameters. Moreover, a number of relatively simple yet powerful computational algorithms facilitate the estimation of limited dependent variable models.

Third, in marketing, inference about model parameters or more generally about different models, i.e., the statistical assessment of the likely mechanisms that bring about consumers' and competitors' behaviors in a market is usually not an end in itself but input to the decisions of marketing managers in companies. The likely benefit from various alternatives for, e.g., product design, product line composition, pricing, or advertising schedules can be expressed as a function of a model and its parameters. However, knowledge of model parameters and generally the model that generated the observed market behaviors will never be perfect. Bayesian modeling facilitates the accurate incorporation of any remaining uncertainty about the mechanism behind observed market behaviors in managerial decisions.

Fourth, computational resources become more powerful and affordable every year, facilitating the estimation of ever more realistic and thus complex models in academic and industry applications. In addition, freely available software such as, e.g., the R-package `bayesm` (see Rossi et al. 2005) makes a collection of Bayesian models useful for marketing applications readily accessible (The latest version of `bayesm` is written for speed using the R-package `Rcpp` (Eddelbuettel and François 2011; Eddelbuettel 2013). The last complete version mostly written in plain R is version 2.2–5. The R-files are available from the CRAN-archives and often a useful start when developing your own routines). In fact, one reason for the popularity of Bayesian modeling among market research practitioners has been the adoption of hierarchical Bayes models for inference by companies like Sawtooth software (Orme 2017) that revolutionized how market research consultants approach the analysis of, for example, choice-based conjoint experiments. Finally, `Stan` (Carpenter et al. 2017) appears as a big step towards freeing creative modeling from having to invest substantial amounts of time in the development of efficient Bayesian estimation routines.

Fifth, because Bayesian estimation is simply the exact reverse of the data generating process (DGP), it is naturally attractive to researchers that are interested in the development and the empirical test of their own marketing models. Some researchers view the need to specify a complete DGP as a drawback. The argument is that theory never is precise enough to do so, and that this requirement leads to arbitrary choices that unduly impact the inference for quantities the data are more or less directly informative about. The Bayesian response to this criticism is to specify highly flexible DGPs in instances where theory is lacking. This strategy is facilitated by algorithms that adaptively determine a reasonable dimensionality of a flexibly formulated model. This determination is based on statistical evidence that potentially favors a lower dimensional, simpler model and not just fails to reject that model as in classical hypothesis testing.

All that said, it usually still takes longer to estimate a fully Bayesian model than it takes to compute maximum likelihood estimates, in case they exist. I have also heard people “complain” about the amount of information contained in large samples from

posterior distributions as produced by modern numerical Bayesian inference tools (Compared to a collection of maximum likelihood estimates and their standard errors). However, it seems natural to wait somewhat longer for a more complete answer to a decision problem. And many interesting decision problems cannot be properly addressed based on a collection of maximum likelihood estimates (should they even exist) and especially upon realizing that their standard errors cannot be reliably estimated with the data at hand.

Bayesian Essentials

A Bayesian model consists of a likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ that fully specifies the probability of the data \mathbf{y} given parameters $\boldsymbol{\theta}$, i.e., the process that generates the data for known parameters. In fact, if the researcher only wants to work with one likelihood function, is not interested in comparing across different mechanisms that may have generated the data, any function that is proportional to $p(\mathbf{y}|\boldsymbol{\theta})$ will do, i.e., all functions that differ from $p(\mathbf{y}|\boldsymbol{\theta})$ only by an arbitrary positive constant c are likelihood functions, $\ell(\mathbf{y}|\boldsymbol{\theta}) \equiv c \cdot p(\mathbf{y}|\boldsymbol{\theta})$. We will revisit this point later. A simple example is the linear regression model $y_i = \mathbf{x}'_i\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim iid\mathcal{N}(0, \sigma_\varepsilon^2)$ that implies the following likelihood for the data $p(\mathbf{y}|\boldsymbol{\beta}, \sigma_\varepsilon^2) = \prod_{i=1}^N \mathcal{N}(y_i | \mathbf{x}'_i\boldsymbol{\beta}, \sigma_\varepsilon^2)$.

The second component of a Bayesian model is a prior distribution for the parameters indexing the likelihood $p(\boldsymbol{\theta})$. The notation $p(\boldsymbol{\theta})$ means “the density p evaluated at the value $\boldsymbol{\theta}$.” Further, defining the prior distribution as $p(\boldsymbol{\theta})$ implies that $\boldsymbol{\theta} \sim p$, i.e., that $\boldsymbol{\theta}$ is (a priori) distributed according to density p , or simply is p -distributed. The notation $p(\boldsymbol{\theta})$ is short-hand because it omits the (subjective prior) parameters indexing the prior distribution. For example, in an application the statement that the prior is a multivariate normal distribution is incomplete. We need to add the information about the prior mean and variance, e.g., $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}^0, \boldsymbol{\Sigma}^0)$, where $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}^0, \boldsymbol{\Sigma}^0)$ is the multivariate normal distribution with mean $\boldsymbol{\theta}^0$ and variance-covariance $\boldsymbol{\Sigma}^0$ evaluated at $\boldsymbol{\theta}$. The multivariate normal density can be evaluated in R using the command `dmvn` from the R-package `mvnfast` (Fasiolo 2016) or the command `dmvnorm` from R-package `mvtnorm` (Genz et al. 2018). Both commands support computations on the log-scale which are *essential* for numerical accuracy. For example, a log-likelihood value of -2000 can only be numerically distinguished from a log-likelihood value of, say, -2050 on the log-scale, because both likelihoods, i.e., $\exp(-2000)$ and $\exp(-2050)$ evaluate to an “exact” machine zero at currently available machine accuracies.

The need to specify prior distributions for Bayesian analysis is often viewed as a drawback of the Bayesian approach. There are several aspects to the specification and the role of the prior distribution in a Bayesian model. First, as suggested by the name, the prior distribution is the formal vehicle to bring prior substantive knowledge to bear on the analysis. And it is sometimes overlooked by critics of the Bayesian approach that such knowledge is already required when specifying the likelihood function. Second, from a purely technical point of view, prior

distributions improve the statistical properties of estimators derived from the model (see e.g., Robert 1994, p. 75).

In the regression example, a useful way to probe into prior knowledge is to think about expected changes in y_i as a function of changes in \mathbf{x}_i . Unless the substantive domain the data originates from is unknown, it is extremely likely that the analyst will have some substantive idea about the DGP that should be used in the formulation of prior distributions. In the event where the analysis is a follow-up on previous statistical analyses in the same or a related domain, the choice of prior can build on these results. An example would be market research companies that more or less continuously study demand in a set of markets.

With the specification of a prior distribution, the analyst expresses his beliefs about what parameter values are more likely than other parameter values and by how much, based on his existing substantive understanding of the DGP. If the analyst specifies a prior such that parameters in a relatively small subset of the parameter space are much more likely than other parameters, the prior is usually referred to as an *informative* prior. The most extreme case of an informative prior is a distribution that concentrates all its mass on a singular parameter value. Such a prior is called *degenerate*. Degenerate priors constrain parameters to take particular values known a priori. Conversely, the prior is *weakly informative* or *diffuse* if there is no discernible concentration of prior mass on subsets of the parameter space. However, unless the parameter space is bounded in all directions as, e.g., in the case of a parameter measuring a probability, it is impossible to put exactly equal prior weight on all parameter values without violating the requirement that the prior needs to be in the form of a probability density function (A function $p(\boldsymbol{\theta})$ is a probability density function if $\int p(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$). Priors that fulfill this requirement are also referred to as *proper* priors and priors that do not are *improper* or literally *noninformative*. Finally, if the prior puts zero mass on subsets of the parameter space, e.g., zero mass on positive price coefficients in a demand model, it is called a *constrained* prior.

Bayesian models then apply Bayes' theorem to derive the posterior distribution of model parameters given the data:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} \quad (1)$$

Equation 1 identifies the goal of a Bayesian model as to make probability statements about quantities of interest, $\boldsymbol{\theta}$. More specifically, a Bayesian model extracts information in the data \mathbf{y} via the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ to update prior knowledge about these quantities summarized in the prior distribution $p(\boldsymbol{\theta})$. The updated knowledge is then used to compare among marketing actions a with payoffs that depend on $\boldsymbol{\theta}$. If we define the loss from an action a given $\boldsymbol{\theta}$ as $\mathcal{L}(a, \boldsymbol{\theta})$ the optimal Bayes action minimizes the posterior expected loss:

$$\mathcal{L}(a|\mathbf{y}) = \int \mathcal{L}(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad (2)$$

In marketing applications, the loss usually does not directly depend on θ but on the implied data \hat{y} , usually some manifestation of demand, i.e., $\mathcal{L}(a, \theta) = \int \mathcal{L}(a, \hat{y}) p(\hat{y}|\theta, a) d\hat{y}$. The notation $p(\hat{y}|\theta, a)$ covers the relevant case where the actions under investigation are conditioning arguments to the DGP. A well-known example is finding the coupon strategy that maximizes net revenues, i.e., minimizes the loss defined as negative net revenues in Rossi et al. (1996).

The denominator in Eq. 1, $p(\mathbf{y})$, is known as the marginal likelihood of the data \mathbf{y} or the normalizing constant of the posterior distribution $p(\theta|\mathbf{y})$. As we will see in section “[Bayesian Estimation](#),” knowledge of this quantity is not required for Bayesian inference given a particular model. However, statements about quantities of interest θ in probability form require that $0 < p(\mathbf{y}) < \infty$. Only if this condition is met, the posterior $p(\theta|\mathbf{y})$ will be in the form of a probability density functions, i.e., $\int p(\theta|\mathbf{y}) d\theta = 1$.

In addition, the marginal likelihood of the data is $p(\mathbf{y})$ needed for the comparison across different models for the *same* data where models may be arbitrarily different in terms of the likelihood function, the prior distribution or both. In fact, based on the marginal likelihood of the data given a particular model \mathcal{M} , i.e., $p(\mathbf{y}|\mathcal{M})$, the decision theoretic framework in Eq. 2 can be extended to cover decisions about the DGP itself, and to take uncertainty about the data generating model into account when choosing a marketing action. The optimal action given a set of possible data generating models $\mathcal{M}_1, \dots, \mathcal{M}_K$ and the data minimizes

$$\mathcal{L}(a|\mathbf{y}, \mathcal{M}_1, \dots, \mathcal{M}_K) = \sum_k p(\mathbf{y}|\mathcal{M}_k) \Pr(\mathcal{M}_k) \int \mathcal{L}(a, \theta) p(\theta|\mathbf{y}, \mathcal{M}_k) d\theta \quad (3)$$

where $\Pr(\mathcal{M}_k)$ is the subjective prior probability that model k is the true model that is often chosen to be $1/K$ in the absence of better knowledge. A marketing application following this general idea is presented in Montgomery and Bradlow (1999).

The fundamental appeal of being able to make probability statements about quantities of interest θ is the seamless integration with decision-making based on the expected utility from a set of possible actions. Note that the posterior expected loss in Eq. 2 will only usefully distinguish between different actions a if the posterior $p(\theta|\mathbf{y})$ integrates to 1, i.e., is a valid probability density function. It should be recognized that a *proper* prior distribution $p(\theta)$ essentially guarantees that we can make these probability statements, independent of any data deficiencies that may be present. A Bayesian model therefore quantifies how much the data, through the likelihood, add to our prior understanding of a DGP by comparing the prior distribution $p(\theta)$ to the posterior distribution $p(\theta|\mathbf{y})$. This is different from the classical question what models or model parameters the data can identify.

Consider the following illustrative example. Let us assume that someone measured the preferences for various credit cards on a linear, continuous scale. The cards vary in terms of brand: Mastercard, Visa, Discover; interest rate on outstanding balances: 18%, 15%, 12%; annual fee: no annual fee, \$10, \$20; and finally the credit limit: \$1000, \$2500, \$5000. The researcher has preference

measures for the following eight cards in Table 1, where “1 s” indicate which attribute levels are present.

Dummy coding using the brand Mastercard, 18% interest, no annual fee and a credit limit of \$1000 as base lines, and adding a constant, we obtain the matrix corresponding to the linear regression model $y_i = \beta_0 + x_{1,i}\beta_1 + \dots + x_{8,i}\beta_8 + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ for cards $i = 1, \dots, 8$ in Table 2.

It is easy to verify that the overall nine β -coefficients in this model are not jointly likelihood identified, because there are only eight observations. This can be viewed as toy example of the increasingly common situation, where the number of (potential) explanatory variables exceeds the number of observations, including big data that owe their size to the number of variables in addition to the number of observations (are “broader” than “long”). In such data sets, a purely data-based distinction between connections from explanatory variables to the dependent variable is no longer possible, even if all explanatory variables come from independent processes a priori.

Inspecting the bivariate correlations between covariates in Table 2 that are depicted in Table 3, we can see that these correlations are not too strong, individually. However, we also see that no two design columns are perfectly orthogonal. I further

Table 1 Credit card Designmatrix

#	Brand			Interest			Annual fee			Credit limit		
	Master	Visa	Discover	18%	15%	12%	\$0	\$10	\$20	\$1000	\$2500	\$5000
1	1	0	0	1	0	0	1	0	0	1	0	0
2	1	0	0	0	0	1	0	0	1	0	0	1
3	0	1	0	1	0	0	0	1	0	0	0	1
4	0	0	1	1	0	0	0	0	1	0	1	0
5	0	0	1	0	0	1	0	1	0	1	0	0
6	0	0	1	0	1	0	1	0	0	0	0	1
7	0	1	0	0	0	1	1	0	0	0	1	0
8	1	0	0	0	1	0	0	1	0	0	1	0

Table 2 Credit card Modelmatrix

#	Constant	Brand		Interest		Annual fee		Credit limit	
		Visa	Discover	15%	12%	\$10	\$20	\$2500	\$5000
	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	1	0	1	0	1
3	1	1	0	0	0	1	0	0	1
4	1	0	1	0	0	0	1	1	0
5	1	0	1	0	1	1	0	0	0
6	1	0	1	1	0	0	0	0	1
7	1	1	0	0	1	0	0	1	0
8	1	0	0	1	0	1	0	1	0

investigate the model in Table 2 using regression analysis. Specifically, I regress each column in Table 2 (excluding the constant) on the remaining columns. Every one of these eight regression results in a perfect prediction because we have $9 - 1 = 8$ predictors and 8 observations, each. The rows in Table 4 report the coefficients from regressing the covariate indicated by the row name on the remaining seven covariates in addition to a constant. A dash indicates that the covariate indicated by the column label in Table 4 is the dependent variable. The “NAs” result from perfect predictions of the covariates “Discover,” “12% interest rate,” and “\$10 annual fee” before including the covariate “\$5,000 credit limit” as predictor.

For example, the last line of Table 4 implies the following deterministic equation from regressing the covariate “\$5,000 credit limit” on the remaining covariates in Table 2: $x_8 = 0 + 1x_1 + 0x_2 + 1x_3 + 0x_4 + 0x_5 + 1x_6 - 1x_7$. The contrasts $x_1, x_3,$ and x_6 involving “Visa,” “15% interest,” and “\$20 annual fee” are therefore positively confounded with the contrast involving “\$5,000 credit limit,” and this latter contrast is negatively confounded with the contrast x_7 involving “\$2,500 credit limit.”

Now, what are the implications for modeling the variation in the preference measures y as a function of covariates? In order to arrive at a likelihood-identified regression model, we need to reduce the number of covariates (the number of columns in Table 2) such that the resulting \mathbf{X} -matrix is of full column rank, and the inverse of $\mathbf{X}'\mathbf{X}$ is well defined. As a general rule, we can always throw out covariates that are independent of *all* covariates we would like to keep in the model,

Table 3 Correlations between design columns

	Visa	Discover	15%	12%	\$10	\$20	\$2500	\$5000
Visa	1							
Discover	-0.45	1						
15%	-0.33	0.15	1					
12%	0.15	-0.07	-0.45	1				
\$10	0.15	-0.07	0.15	-0.07	1			
\$20	-0.33	0.15	-0.33	0.15	-0.45	1		
\$2500	0.15	-0.07	0.15	-0.07	-0.07	0.15	1	
\$5000	0.15	-0.07	0.15	-0.07	-0.07	0.15	-0.6	1

Table 4 Design column dependence – regression analysis

	Constant	Visa	Discover	15%	12%	\$10	\$20	\$2500	\$5000
Visa	0	-	0	-1	0	0	-1	1	1
Discover	0.5	-0.5	-	0	0	0	0	0	NA
15%	0	-1	0	-	0	0	-1	1	1
12%	0.5	0	0	-0.5	-	0	0	0	NA
\$10	0.5	0	0	0	0	-	-0.5	0	NA
\$20	0	-1	0	-1	0	0	-	1	1
\$2500	0	1	0	1	0	0	1	-	-1
\$5000	0	1	0	1	0	0	1	-1	-

without biasing our inference for the influence of the latter. Throwing out such covariates, at worst, increases the unexplained variance. In this example, no covariate fulfills this criterion by the mere fact that we have too many covariates to choose from, relative to the number of observations.

As a second general rule, we can eliminate covariates from the model which we strongly believe (know a priori) to have no (direct) effect on the dependent variable. We can do so regardless of how such covariates are related to covariates we would like to keep in the model, for unbiased inference about the influence of the latter.

However, if we eliminate a covariate that actually has a direct effect on the dependent variable that is *not* independent of all covariates we would like to keep in the model, the resulting inference will be biased. For example, whatever the true preference contribution of “\$5,000 credit limit” relative to the baseline of only “\$1,000 credit limit,” the coefficients associated with “Visa,” “15% interest,” and “\$20 annual fee” will be biased upward by this amount, and the coefficient associated with “\$2,500 credit limit” will be biased downward by the same amount upon deleting column x_8 (“\$5,000 credit limit”) for identification in this example. Also, note that the confounds identified here are not automatically resolved upon collecting more data. In fact, even an infinite number of observations from the model in Table 2 will exhibit the same problem. What is required for improved data based identification is not only more but also “different” data, i.e., data generated by \mathbf{X} -configurations different from those in Table 2. However, more data will necessarily be “suitably different” if the processes that generate the covariates are independent, at least conditionally.

In this particular example, there is no obvious choice of covariates that could be omitted based on strong prior beliefs that their direct effect is equal to zero. In fact, a prior understanding of preferences for credit cards would suggest that all covariates likely causally relate to the observed preferences for the different cards. Thus, any likelihood identified model obtained by omitting covariates from Table 2 is likely to yield substantially biased inferences regarding the influence of covariates retained in the model.

At this point, it is useful to relate likelihood-identification by omitting covariates to the formulation of a prior. In a sense, omitting covariates to achieve likelihood-identification corresponds to a degenerate prior concentrated on zero for the effects of omitted covariates, coupled with an improper prior for the effects of covariates retained in the model. In contrast, a Bayesian model for this data defined through a proper prior over *all* observed covariates expresses the belief that these covariates contributed *causally* independently to the observed preferences, with some prior uncertainty about the size of the individual contributions.

From the perspective of different (implied) priors, I believe that essentially nobody would prefer one of the many possible likelihood identified models in this example to the Bayesian model that keeps with the prior causal structure. Mutilating the prior causal structure to overcome data deficiencies and to achieve likelihood-identification (and more generally statistical efficiency) does not seem to be a generally useful strategy. Obviously, one often can (and should) try to obtain more informative data. However, completely discounting the information in only partially informative data seems to be a wasteful strategy.

Importantly, a prior that expresses the belief in invariant structural aspects of the data generating process will eventually translate into accurate posterior measures of the strength of structural relationships, once more likelihood information becomes available. A model (or prior structure) that is formulated in response to observed data deficiencies will not. Thus, the findings from such a model are generally not useful as prior input to future analysis of data from the same process, be it informative, or again deficient per se, potentially in a different way. We will revisit this topic when we discuss and numerically illustrate hierarchical Bayesian models that manage to extract information about the distribution of parameters from a collection of likelihoods that individually fail likelihood-identification (a collection of “deficient” data sets).

A big intellectual step is thus to acknowledge the limits of a perspective that literally asks “for the data to speak.” The decisions that go into “making the data speak,” be it in the form of simple summaries or complicated (likelihood identified) models, always involve prior knowledge. In this context, trading beliefs about an underlying structure for the ability to relate parameters to well-determined functions of the data only regularly voids the thus identified parameters from the meaning sought by the analyst in the first place. In contrast, updating a structurally intact prior with deficient data preserves the structural interpretation of parameters, at the expense of “purely” data-based identification (I put “purely” in quotes, because the decision about how to arrive at a model that can be identified only based on the data at hand always involves subjective, i.e., non-data based prior knowledge).

Now back to our example. When passed to R’s `lm`-function, for example, `lm` automatically deletes the last column from the model for a model that just identifies the remaining β -coefficients. This model computes eight parameters from eight observations and thus trivially fits the data perfectly. Because of the perfect fit of every member of the class of just identified models, the data cannot distinguish among models in this class. However, as mentioned earlier, prior knowledge strongly suggests that no likelihood-identified model obtained by deleting covariates makes much structural sense in this example.

For illustration, I simulate 1000 data sets using the model matrix in Table 2, a coefficient vector $\beta = (4, 2, 0, 1, 1.5, -1, -1.5, 2, 3)$, and $\sigma_\varepsilon^2 = 1$. For each data set, I estimate the regression model in Table 2 dropping column x_8 for identification which corresponds to the default in R’s `lm`-function. I also estimate a fully conjugate (Conjugacy refers to mathematical properties of a prior in combination with a particular likelihood function. So-called conjugate priors result in posteriors of the same distributional form as the prior. For example, a normal prior is the conjugate prior for the parameters in a normal likelihood with known variance, i.e., a likelihood that implies (conditionally) normally distributed data) Bayesian regression model with conditional prior $\beta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2 100)$ and without dropping any columns from Table 2 using the routine `runireg` in the R-package `bayesm` (Rossi et al. 2005) (The marginal prior for σ_ε^2 is inverse Gamma with 3 degrees of freedom and scale equal to the observed variance of y in each data set, i.e., the default in the R-package `bayesm`).

Table 5 reports the data generating true β -values, the mean of the OLS- and Bayes-estimates across 1000 data replications, as well as the corresponding standard deviations. The comparison between the data generating values and the mean of the

Table 5 Sampling experiment

	True values	OLS		Bayes	
		Mean	Standard deviation	Mean	Standard deviation
Constant	4.0	3.96	0.91	3.95	0.88
Visa	2.0	5.04	1.29	2.72	0.53
Discover	0.0	0.01	0.70	0.01	0.69
15%	1.0	4.02	1.33	1.70	0.54
12%	1.5	1.49	0.68	1.49	0.66
\$10	-1.0	-1.00	0.70	-0.98	0.68
\$20	-1.5	1.53	1.31	-0.77	0.55
\$2500	2.0	-0.98	0.68	1.33	0.38
\$5000	3.0	0	-	2.31	0.42

OLS-estimates clearly illustrates the bias analyzed theoretically earlier. The coefficients associated with “Visa,” “15% interest,” and “\$20 annual fee” are biased upwards by about a value of 3 which corresponds to the data generating preference contribution of $x_8 = 1$, i.e., “\$5,000 credit limit” which was dropped from estimation for identification. The coefficient associated with “\$2,500 credit limit” is biased downward by the same amount. Taking into account the standard deviations in parentheses, these biases appear to be statistically significant, despite the small samples of eight observations. In contrast, the mean of the Bayes-estimates for the same coefficients is much closer to the data generating values. In addition, the standard deviations show that especially the parameters affected by bias in the OLS-regression are estimated with more statistical precision in the Bayesian model.

The main difference between the classical OLS approach and the Bayesian approach here are the assumptions that enable the extraction of information from the data. While classical estimation requires prior information about how to reduce the dimensionality of the inferential problem to deliver estimates, the Bayesian approach allows us to retain the original dimensionality at the expense of assumptions that make regression parameters outside of some range very unlikely. In applications where the form and thus the dimensionality of the likelihood function derive from causal reasoning, i.e., theory, the Bayesian approach thus facilitates inference without having to compromise on what is the core of existing beliefs about the DGP in response to data deficiencies.

The rapidly developing field of machine learning provides alternative approaches to flexibly “regularize” a likelihood function (see e.g., Hastie et al. 2001). On a formal level, the regularization techniques employed in machine learning can be re-expressed as prior assumptions about parameters or likely model structures. And while the machine learning approach may have advantages in applications where the analyst has minimal to no prior knowledge about the DGP, the Bayesian approach excels when such knowledge is available.

The prior employed in our illustrative example certainly is closer to a common sense understanding of preferences for credit cards than the model implied by deleting x_8 (“\$5,000 credit limit”), or any other likelihood-identified model obtained

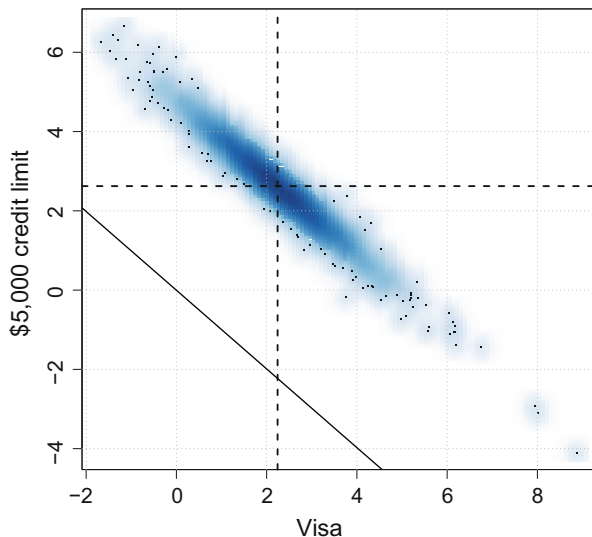
by deleting covariates in this example. However, it is still in the spirit of regularization without much attention to details and incidentally essentially corresponds to a ridge-regression approach (Hoerl and Kennard 1970).

To illustrate further, Fig. 1 depicts the joint posterior of coefficients associated with “Visa” and “\$5,000 credit limit” obtained from one of the 1000 simulated data sets. It illustrates a strong one-to-one trade-off between the “Visa” and “\$5000 credit limit” coefficients (compare the 45-degree downward sloping solid line through the origin). When the draw of the “Visa”-coefficient suggests an exceedingly positive preference for Visa relative to the baseline brand Mastercard, the “\$5,000 credit limit” -coefficient suggests a pronounced distaste for the \$5000 credit limit relative to the baseline, and vice versa. Without the prior, this distribution would collapse to a line with equal support for all coefficients from $(\beta_{\text{Visa}} = -\infty, \beta_{\$5,000} = \infty)$ to $(\beta_{\text{Visa}} = \infty, \beta_{\$5,000} = -\infty)$, and consequently zero support for any finite set of coefficients. This line is the graphical analogue to nonidentifiability. The prior essentially allows for point identification by concentrating posterior support away from the endpoints $(-\infty, \infty)$ and $(\infty, -\infty)$. I believe that essentially everybody would view this as a reasonable assumption after pondering combinations of, say “infinite” preference for Visa with “infinite” distaste for a credit limit of \$5000.

A more elaborate prior could, for example, harness the (weak) prior preference ordering of the levels of interest rate, annual fee, and credit limit, or specific knowledge about the person rating the credit cards (see e.g., Allenby et al. 1995).

Finally, many marketing applications such as, for example, conjoint experiments or the analysis of scanner panel data are characterized by a collection of small data sets that individually are similarly problematic as the one corresponding to Table 2. In such settings, so-called hierarchical Bayes models are useful. Hierarchical Bayes

Fig. 1 Posterior correlation of the “Visa” and the “\$5,000 credit limit” coefficient in one simulated data set



models learn the form of the prior to apply to each individual data set from the collection of data sets. In a hierarchical model, the prior that regularizes each individual level likelihood is therefore itself an object of statistical inference (see e.g., Lenk et al. 1996).

Even in settings where a data set formally identifies the parameters in a likelihood function Bayes theorem (Eq. 1) implies that the prior distribution will “bias away” the posterior from the information in the data. At least in small samples or generally in the context of data that does not contain much information about target parameters, the optimal Bayes action (see Eq. 2) may thus be different from the action that only conditions on likelihood information. And often analysts trained in classical frequentist statistics point out that an objective assessment of, for example, the statistical relevance of a parameter is no longer possible once a subjectively formulated prior enters the inferential procedure.

This criticism is certainly valid. However, the quest for objective inference comes at the price of not being able to use some data sets at all, or only subject to assumptions that likely are less defensible or further removed from a common understanding of the DGP than can be incorporated in a prior distribution. Furthermore, when only finite amounts of data are available, the frequentist assessment of statistical uncertainty in estimates or about models often relies on large sample asymptotic arguments in all but simple linear models. Large sample asymptotic arguments are certainly objective but may or may not hold in a particular application that has to rely on finite data.

Finally, the posterior distribution from priors that have positive support over the entire support of the parameter space as defined by the likelihood function, i.e., are neither degenerate or constrained, will converge to the maximum likelihood estimate as the data become more and more informative. In this sense, priors that are neither degenerate nor constrained result in large sample consistent inferences.

Bayesian Estimation

For the purpose of inference given a particular Bayesian model, knowledge of the marginal likelihood $p(\mathbf{y})$ is not required, because as long as $p(\mathbf{y})$ is finite and positive, we have

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (4)$$

i.e., the posterior distribution is proportional to the product of the likelihood times the prior. This proportionality follows from elementary probability calculus upon recognizing that the product of likelihood times the prior defines the joint density of the data \mathbf{y} and parameters $\boldsymbol{\theta}$, i.e., the conditional distribution of $\boldsymbol{\theta}$ given the data \mathbf{y} is proportional to the joint distribution of parameters and the data.

Another way to appreciate this proportionality is to think about the graphical representation of the posterior distribution of a scalar parameter. It is obvious that the linear scaling of the y -axis in this graph does not matter for *relative* probability

statements of the form $p(\theta_i|\mathbf{y})/p(\theta_j|\mathbf{y})$, because any finite multiplicative constant would cancel from this ratio. For the same reason, posterior Bayesian inference given a model is invariant to rescaling the likelihood, the prior, or both by multiplicative constants. Similarly, the *relative* expected loss from two actions a_k and a_l given a particular model $\mathcal{L}(a_k|\mathbf{y})/\mathcal{L}(a_l|\mathbf{y})$ does not depend on multiplicative constants. However, to compute the expected loss in Eq. 2, we need absolute probability statements about θ , i.e., we need to normalize the product $c_1 p(\mathbf{y}|\theta) c_2 p(\theta)$, where c_1 and c_2 are arbitrary positive “rescaling” constants.

I first discuss two examples where it is relatively obvious how to compute the normalizing constant $\int c_1 p(\mathbf{y}|\theta) c_2 p(\theta) d(\theta)$ in closed form. When the normalizing constant is available in closed form, the posterior $p(\theta|\mathbf{y})$ will usually be in the form of a known distribution. For known distributions, random number generators are implemented as part of statistical programming languages such as, for example, R or can be easily constructed. Based on $r = 1, \dots, R$ draws from such a random number generator, we can approximate the posterior expected loss in Eq. 2 to an arbitrary degree of precision and for arbitrarily complicated nonlinear loss-functions as

$$\mathcal{L}(a|\mathbf{y}) \approx \frac{1}{R} \sum_{r=1}^R \mathcal{L}(a, \theta^r), \quad \theta^r \sim p(\theta|\mathbf{y}), \tag{5}$$

because $\lim_{R \rightarrow \infty} \frac{1}{R} \sum_{r=1}^R \mathcal{L}(a, \theta^r) = \int \mathcal{L}(a, \theta) p(\theta|\mathbf{y}) d\theta$ by the law of large numbers provided that $\mathcal{L}(a|\mathbf{y})$ is known to be finite (Compare this to the definition of the (posterior) mean, i.e., $\int \theta p(\theta|\mathbf{y}) d\theta$ and its estimator from a sample $\theta^1, \dots, \theta^r, \dots, \theta^R$, i.e., $\frac{1}{R} \sum_{r=1}^R \theta^r$). This condition will always hold if the loss function evaluates to finite values over the definitional range of θ , formally $-\infty < \min_{\theta} (\mathcal{L}(a, \theta)) \leq \max_{\theta} (\mathcal{L}(a, \theta)) < \infty$, or more generally if nonfinite $\mathcal{L}(a, \theta)$ is an event of probability measure zero.

I then move to models where the posterior distribution cannot be computed in closed form and introduce Gibbs sampling facilitated by data augmentation and the Metropolis-Hastings algorithm as solutions to Bayesian inference in this case.

Examples of Posterior Distributions in Closed Form

Beta-binomial model. Consider a Bernoulli experiment that yields identically, independently (*iid*) distributed observations y_i taking one of two values, say “1” and “0” with probabilities θ and $1 - \theta$. Repeating the Bernoulli experiment n times results in $s = \sum_{i=1}^n y_i$ “1 s” and $n - \sum_{i=1}^n y_i$ “0 s”. The probability of observing s in n trials given θ is then

$$p(s|n, \theta) = \binom{n}{s} \theta^s (1 - \theta)^{n-s} = \frac{\Gamma(n + 1)}{\Gamma(s + 1) \Gamma(n - s + 1)} \theta^s (1 - \theta)^{n-s} \tag{6}$$

where Γ is the Gamma-function (The relation $\Gamma(n + 1) = n!$ provides some useful intuition for the Gamma-function).

As we will see, a convenient prior for the unobserved $p(y_i = 1) = \theta$ is in the form of a Beta density:

$$p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \tag{7}$$

The parameters a and b can be interpreted as the number of “1 s” and “0 s” in a hypothetical prior experiment and serve to express prior beliefs about θ . However, all real valued $a, b > 0$ result in proper priors for the probability θ over its definitional range, i.e., $\int_0^1 p(\theta|a,b) d\theta = 1$. For example, setting both a and b equal to 1 yields the uniform density over the unit interval expressing the absence of prior knowledge about what θ -values are more likely than others. Setting a and b equal to the same value larger than 1 yields a density that in the limit of $a, b \rightarrow \infty$ degenerates to a point mass at 0.5, which corresponds to various degrees of prior belief strength about θ being equal to 0.5. The mean and mode of the Beta density are given by $a/(a+b)$ and $(a-1)/(a+b-2)$. Therefore, $a > b$ ($a < b$) expresses prior beliefs that $\theta > 0.5$ ($\theta < 0.5$). Finally, for $0 < a, b < 1$, the Beta density takes a bathtub shape that piles up mass at the borders of the parameter space 0 and 1.

Conditional on the data y_1, \dots, y_n , the binomial coefficient that forms the first factor in Eq. 6 is a fixed constant. Similarly, the normalizing constant of the Beta density, i.e., the first factor on the right hand side of Eq. 7 is fixed for a given choice of a, b .

Defining $c_1 = (\Gamma(n+1))^{-1} \Gamma(s+1) \Gamma(n-s+1)$ and $c_2 = (\Gamma(a+b))^{-1} \Gamma(a) \Gamma(b)$ and making use of the proportionality in Eq. 4, we thus have

$$p(\theta|a, b, s, n) \propto c_1 p(s|n, \theta) c_2 p(\theta|a, b) \tag{8}$$

$$\propto \theta^s (1-\theta)^{n-s} \theta^{a-1} (1-\theta)^{b-1} = \theta^{s+a-1} (1-\theta)^{n-s+b-1}$$

Comparing the rightmost expression in Eqs. 8 to 7, we see that this product is in the form of a (non-normalized) Beta density with parameters $\tilde{a} = s + a$ and $\tilde{b} = n - s + b$, and therefore

$$p(\theta|a, b, s, n) = \frac{\Gamma(\tilde{a} + \tilde{b})}{\Gamma(\tilde{a})\Gamma(\tilde{b})} \theta^{\tilde{a}-1} (1-\theta)^{\tilde{b}-1} \tag{9}$$

The fact that the posterior distribution in Eq. 9 is of the same known distributional form as the Beta-prior makes the Beta-prior very convenient in the context of a binomial likelihood function. Technically, the Beta-prior is the conjugate prior to the binomial likelihood.

Moving from Eqs. 6 and 7 to Eq. 8 we dropped all multiplicative constants from the likelihood and the prior that do not depend on θ and then normalized the result from Eq. 8 to arrive at Eq. 9. As discussed following Eq. 4 above, we can do so for the purpose of inference given a particular model that consists of a specific likelihood function and prior. I will address the role of these model-specific constants in the context of formal comparisons between different models further below.

Finally, a useful exercise for first time acquaintances with Bayesian inference is to simulate binomial data, for example, using R’s `binom` command, or simply by

making up n and s , and then simulate from the posterior in Eq. 9 using R's `rbeta` command for different specifications of a and b . Observe how the posterior changes as you use more or less (informative) data and more or less informative priors.

Another intellectually useful exercise is to think about different finite amounts of Bernoulli data that either consists of only “1 s” (or only “0 s”). Clearly, the maximum-likelihood estimate of the data generating probability is one (zero) in this case, and a purely data-based assessment of uncertainty in this estimate is impossible. A question at the core of statistical decision theory then is the following: Is a decision maker better off taking the maximum likelihood probability estimate of one (zero) for granted, or should he rather base his decisions on a proper posterior distribution? (Obtained using a proper prior distribution with positive support over the uniform interval.) A general answer to this question, which we will not attempt to prove here, is that any proper prior will translate into better decisions than taking the maximum likelihood estimate for granted. The only exception is the case where prior knowledge itself implies a deterministic process.

Normal-Normal model. The second example is a normal regression likelihood with a known observation error variance coupled with a normal prior for the regression coefficients. This example is of limited direct practical value. However, it showcases another important conjugate relationship. Moreover, this model serves as a useful building block for Bayesian inference in the binomial probit model discussed later, and numerous other models. Consider the following regression model and implied likelihood function

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i \quad \varepsilon_i \sim iid \mathcal{N}(0, 1)$$

$$p(y_1, \dots, y_n) = \frac{1}{\sqrt{2\pi}^n} \prod_{i=1}^n \exp\left(-\frac{1}{2} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2\right), \quad (10)$$

and a multivariate normal prior distribution for the k regression coefficients corresponding to the entries in \mathbf{x}_i , i.e.,

$$p(\boldsymbol{\beta} | \boldsymbol{\beta}^0, \boldsymbol{\Sigma}^0) = (2\pi)^{-k/2} |\boldsymbol{\Sigma}^0|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}^0)' (\boldsymbol{\Sigma}^0)^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}^0)\right). \quad (11)$$

Defining $\mathbf{y} = (y_1, \dots, y_n)'$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ the posterior distribution is then proportional to (see Eq. 4):

$$p(\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\beta}^0, \boldsymbol{\Sigma}^0) \propto \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}^0)' (\boldsymbol{\Sigma}^0)^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}^0)\right) \prod_{i=1}^n \exp\left(-\frac{1}{2} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2\right)$$

$$\propto \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X} + (\boldsymbol{\Sigma}^0)^{-1}) (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right) \exp\left(-\frac{\tilde{s}}{2}\right)$$

$$\propto \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X} + (\boldsymbol{\Sigma}^0)^{-1}) (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right), \quad (12)$$

where

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})^{-1} (\mathbf{X}'\mathbf{y} + (\Sigma^0)^{-1}\beta^0) \tag{13}$$

$$\tilde{s} = (\mathbf{y} - \mathbf{X}'\tilde{\beta})'(\mathbf{y} - \mathbf{X}'\tilde{\beta}) + (\tilde{\beta} - \beta^0)'(\Sigma^0)^{-1}(\tilde{\beta} - \beta^0) \tag{14}$$

See Rossi et al. (2005) or Zellner (1971) for the details of the transformations in Eq. 12 and note that the posterior mean $\tilde{\beta}$ in Eq. 13 will converge to the ordinary least squares or maximum likelihood estimate as the sample size (the information in the data) increases, for all nondegenerate prior settings (i.e., $|\Sigma^0| > 0$). Smaller (larger) variances in Σ^0 put more (less) prior weight behind the prior guess β^0 . For a well-defined ordinary least squares estimate $\hat{\beta}$ (a well-defined inverse of $\mathbf{X}'\mathbf{X}$), we can write Eq. 13 as

$$\begin{aligned} \tilde{\beta} &= (\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})^{-1} (\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + (\Sigma^0)^{-1}\beta^0) \\ &= \tilde{\beta} = (\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})^{-1} (\mathbf{X}'\mathbf{X}\hat{\beta} + (\Sigma^0)^{-1}\beta^0) \end{aligned}$$

which illustrates that the posterior mean $\tilde{\beta}$ is a weighted convex combination of the ordinary least squares or maximum likelihood estimate $\hat{\beta}$ and the prior mean β^0 , where the weights are information from the data in $\mathbf{X}'\mathbf{X}$ and the amount of prior information $(\Sigma^0)^{-1}$, respectively. Thus, the posterior mean will be somewhere “in between” the ordinary least squares estimate and the prior mean.

When combining the normal likelihood (Eq. 10) with the normal prior (Eq. 11) in Eq. 12, we dropped the multiplicative constants $\frac{1}{\sqrt{2\pi}}$ and $(2\pi)^{-k/2}|\Sigma^0|^{-1/2}$ from the likelihood and the prior, respectively. Again, this is fine as long we are only interested in inference given this specific model. Upon recognizing that the last line of Eq. 12 is the so-called kernel of a multivariate normal distribution (the kernel of a distribution drops all factors that do not directly depend on both unobserved parameters and the data or variables the distribution is for) and thus using

$$\begin{aligned} \int \exp\left(-\frac{1}{2}(\beta - \tilde{\beta})'(\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})(\beta - \tilde{\beta})\right) d\beta &= \\ &= (2\pi)^{k/2}|\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1}|^{-1/2} \end{aligned} \tag{15}$$

we obtain the joint posterior distribution of the k regression coefficients in closed form:

$$\begin{aligned} p(\beta|\mathbf{y}, \beta^0, \Sigma^0) &= (2\pi)^{-k/2}|\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1}|^{1/2} \exp\left(-\frac{1}{2}(\beta - \tilde{\beta})'(\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})(\beta - \tilde{\beta})\right) \\ &= \mathcal{N}\left(\beta|\tilde{\beta}, (\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})^{-1}\right) \end{aligned} \tag{16}$$

We can directly sample from this distribution using, for example, the command `rmvnorm` in the R-package `mvtnorm` (Genz et al. 2018) or the faster version `rmvnorm` available in the R-package `mvnfast` (Fasiolo 2016). The `bayesm` (Rossi et al. 2005) routine corresponding to this model is `breg`.

Posterior Distributions Not in Closed Form

Next, I discuss the model defined by the combination of a binomial probit likelihood and a multivariate normal prior for the regression coefficients (see Eq. 11). Bayesian inference for this model is relatively much more challenging than for the two models discussed already because the normalizing constant of the posterior distribution is not available in closed form. The binomial probit likelihood is, similar to the binomial likelihood in Eq. 6, a DGP for independently distributed observations y_i taking one of two values, say “1” and “0”. The probit likelihood defines the probability of observing $y_i = 1$ as a function of covariates \mathbf{x}_i and (probit-)regression parameters $\boldsymbol{\beta}$ as follows:

$$p(y_i = 1|\boldsymbol{\beta}) = \Phi(\mathbf{x}'_i\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'_i\boldsymbol{\beta}} \mathcal{N}(z|0,1) dz \quad (17)$$

$$p(y_i = 0|\boldsymbol{\beta}) = \Phi(-\mathbf{x}'_i\boldsymbol{\beta}) = \int_{\mathbf{x}'_i\boldsymbol{\beta}}^{\infty} \mathcal{N}(z|0,1) dz \quad (18)$$

Thus, observations $\mathbf{y} = (y_1, \dots, y_n)'$ are not identically distributed but provide information about $\boldsymbol{\beta}$ exchangeably. “Exchangeably” essentially means that we don’t need to keep track of the order or sequence of the data for proper inference. Exchangeability here is a consequence of conditional independence given the data generating parameters and observed covariates (see e.g., Bernardo and Smith 2001), conditional on covariates $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$. The data \mathbf{y} then have probit likelihood:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}) &= \prod_{i=1}^n \left(\Phi(\mathbf{x}'_i\boldsymbol{\beta}) \right)^{y_i} \left(\Phi(-\mathbf{x}'_i\boldsymbol{\beta}) \right)^{1-y_i} \\ &= \prod_{i=1}^n \left(\int_{-\infty}^{\mathbf{x}'_i\boldsymbol{\beta}} \mathcal{N}(z|0,1) dz \right)^{y_i} \left(\int_{\mathbf{x}'_i\boldsymbol{\beta}}^{\infty} \mathcal{N}(z|0,1) dz \right)^{1-y_i} \end{aligned} \quad (19)$$

By Eq. 4, the posterior distribution of $\boldsymbol{\beta}$ is proportional to:

$$p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\beta}^0, \boldsymbol{\Sigma}^0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)'(\boldsymbol{\Sigma}^0)^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)\right) \prod_{i=1}^n \left(\Phi(\mathbf{x}'_i\boldsymbol{\beta}) \right)^{y_i} \left(\Phi(-\mathbf{x}'_i\boldsymbol{\beta}) \right)^{1-y_i} \quad (20)$$

As already mentioned, the normalizing constant of the right hand side in Eq. 20 cannot be computed in closed form and we thus cannot derive the posterior

distribution directly, unlike in the previous examples. I will introduce Gibbs sampling as one solution to Bayesian inference in this model. To this end, an alternative interpretation of the probit likelihood suggested by the integral on the right hand side of Eq. 17 will be useful. Taking advantage of the symmetry of the normal distribution, we can rewrite:

$$p(y_i = 1 | \boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'_i \boldsymbol{\beta}} \mathcal{N}(z|0,1) dz = \int_0^{\infty} \mathcal{N}(z|\mathbf{x}'_i \boldsymbol{\beta}, 1) dz \tag{21}$$

$$p(y_i = 0 | \boldsymbol{\beta}) = \int_{\mathbf{x}'_i \boldsymbol{\beta}}^{\infty} \mathcal{N}(z|0,1) dz = \int_{-\infty}^0 \mathcal{N}(z|\mathbf{x}'_i \boldsymbol{\beta}, 1) dz \tag{22}$$

and interpret the binomial probit model as a random utility model in which latent utilities z_i are independently normally distributed with means $\mathbf{x}'_i \boldsymbol{\beta}$ and standard deviations equal to 1. A latent utility draw z_i from $\mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$ larger than 0 generates an observed $y_i = 1$ and a draw smaller than 0 an observed $y_i = 0$, i.e., $y_i = \mathbf{1}(z_i > 0)$. This is exactly equivalent to generating a y -observation using the probability in Eq. 21 as the parameter of a Bernoulli distribution (Draw a random uniform number u from the interval $[0, 1]$, e.g., using `runif(1)` in R and compare to the probability in Eq. 21. Set $y_i = 1$ ($y_i = 0$) when u is smaller (larger) than this probability or use the R-command `rbinom`) because, e.g., $\lim_{R \rightarrow \infty} \frac{1}{R} \sum_{r=1}^R \mathbf{1}(z^r > 0) = E_z \mathbf{1}(z > 0) = \int_0^{\infty} \mathcal{N}(z|\mathbf{x}'_i \boldsymbol{\beta}, 1) dz$.

If we had access to the latent utilities $\mathbf{z} = (z_1, \dots, z_n)'$ that generated the observed binomial data $\mathbf{y} = (y_1, \dots, y_n)'$, we could comfortably rely on the closed form results in Eq. 16 for Bayesian inference. Conditional on the data generating \mathbf{z} , we would in fact learn more about the regressions coefficients than we ever could from the corresponding \mathbf{y} .

Conversely, if we knew the regression coefficients $\boldsymbol{\beta}$ that generated the data, we could make an informed guess about the corresponding data generating \mathbf{z} . Based on the \mathbf{y} -data, we know that \mathbf{z} that correspond to observed 1's must have been larger than zero and those corresponding to observed 0's smaller than zero. Based on $\boldsymbol{\beta}$ and the random utility interpretation of the probit likelihood, we know that the z_i came independently from $\mathcal{N}(\mathbf{x}'_i \boldsymbol{\beta}, 1)$. Putting these insights together, we arrive at the following conditional distribution for a z_i corresponding to observed $y_i = 1$, and that for a z_j corresponding to observed $y_j = 0$ given $\boldsymbol{\beta}$:

$$p(z_i | \boldsymbol{\beta}, y_i = 1) = \frac{\mathcal{N}(z_i | \mathbf{x}'_i \boldsymbol{\beta}, 1) \mathbf{1}(z_i > 0)}{\int_0^{\infty} \mathcal{N}(z | \mathbf{x}'_i \boldsymbol{\beta}, 1) dz} = \mathcal{TN}(z_i | \mathbf{x}'_i \boldsymbol{\beta}, 1, 0, \infty) \tag{23}$$

$$p(z_j | \boldsymbol{\beta}, y_j = 0) = \frac{\mathcal{N}(z_j | \mathbf{x}'_j \boldsymbol{\beta}, 1) \mathbf{1}(z_j < 0)}{\int_{-\infty}^0 \mathcal{N}(z | \mathbf{x}'_j \boldsymbol{\beta}, 1) dz} = \mathcal{TN}(z_j | \mathbf{x}'_j \boldsymbol{\beta}, 1, -\infty, 0) \tag{24}$$

Here $\mathbf{1}(\cdot)$ is an indicator function that evaluates to one if its argument is true and else to zero, and $\mathcal{TN}(a,b,c,d)$ is short for a normal distribution with mean a , variance

b , truncated below c , and above d . We can simulate from these distributions using a trick known as the inverse CDF-transformation (see e.g., Rossi et al. 2005), or rely on the command `rtruncnorm` in the R-package `truncnorm` (Mersmann et al. 2018) which builds on Geweke (1991).

Based on the results in Eq. 16 the conditional distribution of β given the \mathbf{z} and the \mathbf{y} is:

$$p(\beta|\mathbf{z},\mathbf{y},\beta^0, \Sigma^0) = p(\beta|\mathbf{z},\beta^0,\Sigma^0) = \mathcal{N}\left(\tilde{\beta}, \left(\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1}\right)^{-1}\right), \quad (25)$$

where

$$\tilde{\beta} = \left(\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1}\right)^{-1} \left(\mathbf{X}'\mathbf{z} + (\Sigma^0)^{-1}\beta^0\right) \quad (26)$$

Note that once we condition on the \mathbf{z} in Eq. 25, the \mathbf{y} are no longer required as conditioning argument. A particular set of \mathbf{z} transmits all the information, and in fact more information than contained in the \mathbf{y} , to β (I will discuss general rules for the derivation of conditional distributions later and for now concentrate on what can be achieved based on conditional distributions).

Gibbs sampler. Our goal is thus to derive the marginal posterior distribution $p(\beta|\mathbf{y}, \beta^0, \Sigma^0)$ that is free from the extra, but virtual information about β that comes with each particular set of \mathbf{z} we may condition on in Eq. 25. However, as we already know, this posterior is not available in closed form. A convenient solution to this problem is the Gibbs sampler. The Gibbs sampler allows us to generate draws from $p(\beta, \mathbf{z}|\mathbf{y}, \beta^0, \Sigma^0)$ based on knowledge of $p(\mathbf{z}|\mathbf{y}, \beta) = \prod_{i=1}^n p(z_i|\beta, y_i)$ and $p(\beta|\mathbf{z}, \beta^0, \Sigma^0)$, i.e., conditional distributions only. Once we have draws from $p(\beta, \mathbf{z}|\mathbf{y}, \beta^0, \Sigma^0)$, each draw of β in that sample is a draw from our target distribution $p(\beta|\mathbf{y}, \beta^0, \Sigma^0)$ (Recall that the joint distribution $p(\beta, \mathbf{z}|\mathbf{y}, \beta^0, \Sigma^0)$ can be decomposed into the product of the marginal distribution $p(\beta|\mathbf{y}, \beta^0, \Sigma^0)$ and the conditional distribution $p(\mathbf{z}|\mathbf{y}, \beta)$ by elementary probability calculus. If we have access to a sample from the joint distribution, drawing a β with no regard to the companion \mathbf{z} and then looking at the companion \mathbf{z} in the sample is equivalent to drawing from $p(\beta|\mathbf{y}, \beta^0, \Sigma^0)$ and then from $p(\mathbf{z}|\mathbf{y}, \beta)$).

The Gibbs sampler is an application of the fact that the joint distribution $p(\beta, \mathbf{z}|\mathbf{y}, \beta^0, \Sigma^0)$ is uniquely determined by corresponding complete sets of conditional distributions (Besag 1974). The correspondence between the conditional distributions $p(\beta|\mathbf{z}, \beta^0, \Sigma^0)$ and $p(\mathbf{z}|\mathbf{y}, \beta)$ and the joint posterior distribution is illustrated in Eq. 27 which is an instance of the Hammersley-Clifford theorem. For clarity of notation, I abbreviate the subjective prior parameters β^0, Σ^0 to “•” in the following.

$$\begin{aligned} p(\beta, \mathbf{z}|\mathbf{y}, \bullet) &= p(\beta|\mathbf{z}, \bullet)p(\mathbf{z}|\mathbf{y}, \bullet) \\ &= p(\beta|\mathbf{z}, \bullet) \left(\int \frac{p(\beta|\mathbf{z}, \bullet)}{p(\mathbf{z}|\mathbf{y}, \beta)} d\beta \right)^{-1} \end{aligned} \quad (27)$$

Proof:

$$\begin{aligned}
 p(\mathbf{z}|\mathbf{y},\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{y},\bullet) &= p(\boldsymbol{\beta}|\mathbf{z},\bullet)p(\mathbf{z}|\mathbf{y},\bullet) \\
 \frac{p(\boldsymbol{\beta}|\mathbf{y},\bullet)}{p(\mathbf{z}|\mathbf{y},\bullet)} &= \frac{p(\boldsymbol{\beta}|\mathbf{z},\bullet)}{p(\mathbf{z}|\mathbf{y},\boldsymbol{\beta})} \\
 \int \frac{p(\boldsymbol{\beta}|\mathbf{y},\bullet)}{p(\mathbf{z}|\mathbf{y},\bullet)} d\boldsymbol{\beta} &= \int \frac{p(\boldsymbol{\beta}|\mathbf{z},\bullet)}{p(\mathbf{z}|\mathbf{y},\boldsymbol{\beta})} d\boldsymbol{\beta} \\
 \frac{1}{p(\mathbf{z}|\mathbf{y},\bullet)} &= \int \frac{p(\boldsymbol{\beta}|\mathbf{z},\bullet)}{p(\mathbf{z}|\mathbf{y},\boldsymbol{\beta})} d\boldsymbol{\beta}
 \end{aligned} \tag{28}$$

Based on $r = 1, \dots, R$ draws from $p(\boldsymbol{\beta}|\mathbf{z}, \bullet)$, we can therefore estimate the *marginal* distribution:

$$p(\mathbf{z}|\mathbf{y}) = \left(\int \frac{p(\boldsymbol{\beta}|\mathbf{z},\bullet)}{p(\mathbf{z}|\mathbf{y},\boldsymbol{\beta})} d\boldsymbol{\beta} \right)^{-1} \approx \left(\frac{1}{R} \sum_{r=1}^R \frac{1}{p(\mathbf{z}|\mathbf{y},\boldsymbol{\beta}^r)} \right)^{-1} \tag{29}$$

and thus compute the *joint* distribution $p(\boldsymbol{\beta}, \mathbf{z}|\mathbf{y}, \bullet)$ based on only knowledge of the *conditional* distributions $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\beta})$ and $p(\boldsymbol{\beta}|\mathbf{z}, \bullet)$. The Gibbs sampler which builds on this fundamental relationship proceeds as follows:

1. Based on a starting value for $\boldsymbol{\beta}$ draw \mathbf{z} from $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\beta})$ as given in Eqs. 23 and 24.
2. Use the most recent draw of \mathbf{z} as conditioning argument in $p(\boldsymbol{\beta}|\mathbf{z},\bullet)$ (Eq. 25) and draw a new $\boldsymbol{\beta}$.
3. Use the most recent draw of $\boldsymbol{\beta}$ as conditioning argument in $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\beta})$ (Eqs. 23 and 24) and draw new \mathbf{z} .
4. Return to step 2, until completing R cycles through step 2 and step 3, and then stop.

Each completed cycle through steps 2 and 3 delivers a pair $(\boldsymbol{\beta}, \mathbf{z})^r$ where $r = 1, \dots, R$ indexes the cycle or iteration number of the Gibbs-sampler. Under rather general conditions for the conditional distributions involved, these pairs will represent draws from the joint distribution after some initial iterations, and independent of the choice of starting value. The initial iterations serve to “make the Gibbs sampler forget” the arbitrary starting value in step 1 above. This is often referred to as the “burn-in” period of the Gibbs-sampler. Intuitively, the choice of starting value does not matter, because the Gibbs sampler will forget it, no matter which value was chosen (However, the choice of starting value may influence how many iterations it takes before the Gibbs sampler converges, i.e., delivers pairs $(\boldsymbol{\beta}, \mathbf{z})^r$ in proportion to their joint posterior density in a finite sample of R draws. Another practical concern for the choice of starting values is the numerical stability of the techniques used to draw from the conditional distributions).

Steps 2 and 3 above are often referred to as “blocks of the sampler.” Note that step 2 itself consists of n -subblocks that each draw from the conditional

distribution of a particular z_i . However, because all z_i are conditionally independent, i.e., $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^n p(z_i|\boldsymbol{\beta}, y_i)$ (see Eqs. 23 and 24), step 2 effectively draws from the joint conditional posterior distribution of \mathbf{z} . Similarly, step 3 draws from the joint conditional posterior distribution of all elements in $\boldsymbol{\beta}$.

To further strengthen the intuition for the Gibbs sampler, it is useful to think about each iteration as an exploration of the joint distribution in some neighborhood defined by the respective conditioning arguments. By the notion of sampling and updating of conditioning arguments, the Gibbs sampler is, however, not going to stay in this neighborhood but will move away from it and eventually return.

Each time it returns to some fixed neighborhood of $\boldsymbol{\beta}$ -values, for example, it will do so from a different constellation of \mathbf{z} . Returns from \mathbf{z} -constellations that are closer to this $\boldsymbol{\beta}$ -neighborhood in the sense of Eqs. 25 and 26 will occur more often than returns from \mathbf{z} -constellations that are further away. Thus, looking at pairs $(\boldsymbol{\beta}, \mathbf{z})^r$ in this neighborhood, it is impossible to distinguish between moves “from $\boldsymbol{\beta}$ to \mathbf{z} ” and moves “from \mathbf{z} to $\boldsymbol{\beta}$,” and this will be true of every $\boldsymbol{\beta}$ -neighborhood and \mathbf{z} -neighborhood supported by the posterior distribution. In addition, by successively sampling from conditional distributions which are, by definition, proportional to the joint distribution, the Gibbs sampler is going to spend relatively more (fewer) iterations in areas of higher (lower) density under the joint distribution.

In other words, successive pairs $(\boldsymbol{\beta}, \mathbf{z})^1, \dots, (\boldsymbol{\beta}, \mathbf{z})^r, \dots, (\boldsymbol{\beta}, \mathbf{z})^R$ produced by iterations of the Gibbs sampler are locally dependent in the sense that pairs produced in successive iterations are more similar to each other than pairs produced further apart from each other, where distance is measured in iteration counts of the Gibbs sampler. However, all pairs provide exchangeable information about the joint posterior distribution. We can therefore use the output from the Gibbs sampler to approximate posterior expected loss (see Eq. 5) and any aspect of the posterior distribution we may be interested in by the corresponding expectation using the Gibbs output. For example, the posterior probability that a particular regression coefficient is larger than zero, i.e., $P(\beta_k > 0|\mathbf{y}, \bullet) = \int_0^\infty p(\beta_k|\mathbf{y}, \bullet)$ would be estimated from the Gibbs output as $\frac{1}{R} \sum_{r=1}^R \mathbf{1}(\beta_k^r > 0)$. Note that we control the degree of accuracy of these approximations by the length of the Gibbs sample R .

The particular Gibbs sampler described here is implemented as routine `rbprobitGibbs` in the R-package `bayesm` (Rossi et al. 2005) and dates back to Albert and Chib (1993). The routine comes with an example that illustrates input and output (Another `bayesm` routine, `rbiNormGibbs`, nicely illustrates how the Gibbs sampler explores a two-dimensional joint distribution by successively sampling from the corresponding two conditional distributions).

Data augmentation. In this application of the Gibbs sampler, the interest really is on the marginal posterior distribution of probit regression coefficients, i.e., $p(\boldsymbol{\beta}|\mathbf{y}, \bullet)$, and Gibbs sampling from the joint posterior distribution of $\boldsymbol{\beta}$ and \mathbf{z} is just a means to obtaining the marginal distribution of interest. Drawing from $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\beta})$ is therefore referred to as “data augmentation” in the literature. Data augmentation often helps transform Bayesian inference problems that involve “unknown” distributions, i.e., distributions without a normalizing constant in closed form, into problems that only

involve sampling from distributions with known normalizing constants through conditioning. Canonical examples for the successful application of this technique are the multinomial probit model (McCulloch and Rossi 1994), the multivariate probit model Edwards and Allenby (2003), mixture models (see e.g., Allenby et al. 1998; Frühwirth-Schnatter et al. 2004; Lenk and DeSarbo 2000; Otter et al. 2004), and hierarchical models in general.

From the perspective of Gibbs sampling, there is no distinction between (unobserved) aspects of the data, unobserved parameters, or any unobservable we can derive a conditional distribution for, within the confines of the Bayesian model under investigation. However, before one gets too excited about the possibilities of *inference* about any unobservable, it is useful to reflect about how much we can learn about β and \mathbf{z} from the data in this example.

While it is possible to attain perfect posterior knowledge about β in this model in the limit of an infinitely large sample, it is impossible to ever learn the particular set of \mathbf{z} 's that generated the data. This information is lost forever when moving from the data generating \mathbf{z} to the observed \mathbf{y} based on the indicator function $y_i = \mathbf{1}(z_i > 0)$. We have one observation y_i to learn about each z_i . This observation only set identifies z_i , i.e., indicates if $z_i < 0$ or $z_i > 0$. In addition $\mathcal{N}(\mathbf{x}'_i\beta, 1)$ which can be viewed as a hierarchical prior for the z_i cannot degenerate, i.e., cannot deliver a perfect prediction by the definition of the probit likelihood. Any finite valued $\mathbf{x}'_i\beta$ allows for $y_i = 1$ and $y_i = 0$, even if one of the two outcomes is extremely unlikely.

As such, we are severely limited in what we can learn about the data generating \mathbf{z} no matter how many probit observations become available or what subjective prior parameters β^0 and Σ^0 we use. Thus, it is generally useful to distinguish between unobservables that can be consistently estimated in a particular model and unobservables that cannot, before further using the output from the Gibbs sampler. Here “consistently” means that we can think of amounts of data, i.e., likelihood information, or a subjective prior setting that translates into a degenerate posterior distribution which concentrates all its mass in one point. For example, it would be foolish to believe that using the posterior distribution of \mathbf{z} could somehow further improve decisions informed by the data \mathbf{y} and the model at hand, which depend on $p(\beta|\mathbf{y})$, only.

Blocking. One could replace step 2 in the Gibbs sampler above by a Gibbs cycle through the full conditional distributions of each element β_k in β , i.e., $p(\beta_k | \beta_{-k}, \mathbf{z}, \bullet)$, where β_{-k} is short for all but the k -th element (These conditional densities are easily derived from the joint conditional normal distribution in Eq. 16 using linear regression theory).

Because any corresponding complete set of conditional distributions uniquely determines the joint distribution, this alternative sampler again delivers draws from the *same* joint posterior distribution $p(\beta, \mathbf{z}|\mathbf{y}, \bullet)$. However, the local dependence between successive pairs $(\beta, \mathbf{z})^1, \dots, (\beta, \mathbf{z})^r, \dots, (\beta, \mathbf{z})^R$ produced by iterations of this alternative Gibbs sampler is relatively higher. This is because two successive cycles through $p(\beta_k | \beta_{-k}, \mathbf{z}, \bullet)$ for all k -elements deliver draws of β that are more similar in expectation than two draws from $p(\beta|\mathbf{z}, \bullet)$, which are independently distributed.

Replacing a cycle like that through $p(\beta_k | \beta_{-k}, \mathbf{z}, \bullet)$ for all k -elements by a direct draw from the corresponding conditional joint distribution, in this case $p(\boldsymbol{\beta} | \mathbf{z}, \boldsymbol{\beta}^0, \boldsymbol{\Sigma}^0)$, in a Gibbs sampler is referred to as “blocking,” or “grouping” (e.g., Chen et al. 2000). In general, blocked Gibbs samplers deliver more additional information about the posterior distribution per incremental iteration than unblocked samplers, which is intuitive considering direct *iid*-sampling from the joint posterior distribution as the theoretical limit of blocking. As such, blocked samplers also deliver pairs $(\boldsymbol{\beta}, \mathbf{z})'$ in proportion to their joint posterior density in a finite sample based on fewer iterations, converge faster from arbitrary starting values.

Another technical aspect is the order in which to successively draw from the blocks of a Gibbs sampler. The theory of Gibbs sampling implies that the order does not matter and in fact a random ordering is easiest to motivate theoretically (see e.g., Roberts 1996, p. 51). However, in our particular example, repeated draws from step 2, i.e., $p(\boldsymbol{\beta} | \mathbf{z}, \bullet)$, or step 3, i.e., $p(\mathbf{z} | \mathbf{y}, \boldsymbol{\beta})$, without switching to the respective other block in between are a perfect waste of time because these draws are conditionally *iid*. Furthermore, randomly switching to step 2 before updating *all* elements of \mathbf{z} in step 3 is inefficient because step 2 pools information across all \mathbf{z} . The updated pooled information is then “redistributed” across all \mathbf{z} when drawing from $p(\mathbf{z} | \mathbf{y}, \boldsymbol{\beta})$ in step 3.

Conditional posterior distributions. Next I show how to derive the full conditional distributions that define the Gibbs sampler for the probit model above (see also Gilks 1996). Recall that by specifying a prior distribution and a likelihood function, we implicitly specify the joint distribution of unobservables and the data (see Eq. 1). Starting from the joint distribution of the data and unobservables in our example, i.e.,

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\beta} | \boldsymbol{\beta}^0, \boldsymbol{\Sigma}^0) = p(y_1, \dots, y_n, z_1, \dots, z_n, \beta_1, \dots, \beta_K | \boldsymbol{\beta}^0, \boldsymbol{\Sigma}^0)$$

we can derive any conditional distribution of interest using elementary probability calculus. Omitting the conditioning arguments $\boldsymbol{\beta}^0$ and $\boldsymbol{\Sigma}^0$ for clarity of notation we have for example

$$\begin{aligned} p(z_1 | y_1, \dots, y_n, z_2, \dots, z_n, \beta_1, \dots, \beta_K) \\ = \frac{p(y_1, \dots, y_n, z_1, \dots, z_n, \beta_1, \dots, \beta_K)}{\int p(y_1, \dots, y_n, z_1, \dots, z_n, \beta_1, \dots, \beta_K) dz_1} \end{aligned} \tag{30}$$

which does not look simple or useful yet. However, based on an understanding of how the model operates as a DGP, we can greatly simplify this expression. It is in this sense that Bayesian inference exactly reverses the steps that we believe generated the data.

Recall the latent utility interpretation of the probit likelihood function. Given $\boldsymbol{\beta}$, latent utilities \mathbf{z} are generated independently from $\mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n) = \prod_{i=1}^n \mathcal{N}(\mathbf{x}'_i\boldsymbol{\beta}, 1)$. Then the signs of the elements in \mathbf{z} independently determine the data \mathbf{y} according to indicator functions $y_i = \mathbf{1}(z_i > 0)$ for all $i = 1, \dots, n$. Based on this understanding

of the conditional independence relationships in the DGP, we can rewrite and simplify Eq. 30 as follows:

$$\begin{aligned}
 & p(z_1|y_1, \dots, y_n, z_2, \dots, z_n, \beta_1, \dots, \beta_K) \\
 &= \frac{p(\beta_1, \dots, \beta_K) \prod_{i=1}^n p(y_i|z_i)p(z_i|\beta_1, \dots, \beta_K)}{\int p(\beta_1, \dots, \beta_K) \prod_{i=1}^n p(y_i|z_i)p(z_i|\beta_1, \dots, \beta_K) dz_1} \\
 &= \frac{p(\beta_1, \dots, \beta_K) \prod_{i=2}^n p(y_i|z_i)p(z_i|\beta_1, \dots, \beta_K)p(y_1|z_1)p(z_1|\beta_1, \dots, \beta_K)}{p(\beta_1, \dots, \beta_K) \prod_{i=2}^n p(y_i|z_i)p(z_i|\beta_1, \dots, \beta_K) \int p(y_1|z_1)p(z_1|\beta_1, \dots, \beta_K) dz_1} \\
 &= \frac{p(y_1|z_1)p(z_1|\beta_1, \dots, \beta_K)}{\int p(y_1|z_1)p(z_1|\beta_1, \dots, \beta_K) dz_1} = \frac{p(y_1|z_1)p(z_1|\beta_1, \dots, \beta_K)}{p(y_1|\beta_1, \dots, \beta_K)} \\
 &\propto p(y_1|z_1)p(z_1|\beta_1, \dots, \beta_K) \propto p(z_1|y_1, \beta_1, \dots, \beta_K)
 \end{aligned} \tag{31}$$

The last line in Eq. 31 follows from the fact that both y_1 and β_1, \dots, β_K are conditioning arguments, i.e., fixed (for the moment). A useful interpretation of the final result, and in fact a way to derive the result almost instantly, is that the (conditional) posterior of z_1 is proportional to the “likelihood” of z_1 i.e., $p(y_1|z_1) = \mathbf{1}(z_1 > 0)^{y_1} \mathbf{1}(z_1 < 0)^{1-y_1}$ times a “prior probability” of z_1 , i.e., $p(z_1|\beta_1, \dots, \beta_K) = \mathcal{N}(z_1|\mathbf{x}'_1\boldsymbol{\beta})$. In other words, the (conditional) posterior is proportional to the probability of everything that directly depends on z_1 , i.e., the probability of z_1 ’s “children,” times the probability of z_1 given everything z_1 directly depends on, i.e., z_1 ’s “parents.” (The terminology “children” and “parents” is owed to the representation of joint distributions and their conditional independence relationships in the form of directed acyclic graphs (see e.g., Pearl 2009, p. 12))

Using the same logic, we can derive the full conditional density of, e.g., the first element in $\boldsymbol{\beta}$:

$$\begin{aligned}
 & p(\beta_1|y_1, \dots, y_n, z_1, \dots, z_n, \beta_2, \dots, \beta_K) \\
 &= \frac{p(\beta_1, \dots, \beta_K) \prod_{i=1}^n p(y_i|z_i)p(z_i|\beta_1, \dots, \beta_K)}{\int p(\beta_1, \dots, \beta_K) \prod_{i=1}^n p(y_i|z_i)p(z_i|\beta_1, \dots, \beta_K) d\beta_1} \\
 &= \frac{p(\beta_2, \dots, \beta_K) \prod_{i=1}^n p(y_i|z_i)p(\beta_1|\beta_2, \dots, \beta_K) \prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K)}{p(\beta_2, \dots, \beta_K) \prod_{i=1}^n p(y_i|z_i) \int p(\beta_1|\beta_2, \dots, \beta_K) \prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K) d\beta_1} \\
 &= \frac{p(\beta_1|\beta_2, \dots, \beta_K) \prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K)}{\int p(\beta_1|\beta_2, \dots, \beta_K) \prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K) d\beta_1} \\
 &= \frac{p(\beta_1|\beta_2, \dots, \beta_K) \prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K)}{\prod_{i=1}^n p(z_i|\beta_2, \dots, \beta_K)} \\
 &\propto p(\beta_1|\beta_2, \dots, \beta_K) \prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K) \propto p(\beta_1|z_1, \dots, z_n, \beta_2, \dots, \beta_K)
 \end{aligned} \tag{32}$$

Therefore, the full conditional posterior of β_1 does not depend on the observed data \mathbf{y} , conditional on \mathbf{z} . Again we find that the conditional posterior is proportional to the product of the (conditional) prior $p(\beta_1|\beta_2, \dots, \beta_K)$ times the “likelihood,” i.e., the probability of everything that directly depends on β_1 in the DGP, i.e., $\prod_{i=1}^n p(z_i|\beta_1, \dots, \beta_K)$. Note both factors in this product involve normal distributions, and drawing all elements of $\boldsymbol{\beta}$ jointly from $p(\beta_1, \dots, \beta_K|z_n, \dots, z_n)$, as in Eq. 25 is simple if the joint prior distribution of $\boldsymbol{\beta}$ is multivariate normal.

Bayesian prediction. We just saw that in a Bayesian model conditional posterior distributions derive from the joint density of the data and the parameters defined by the Bayesian model, i.e., the combination of a likelihood function with a prior distribution for its parameters. Now consider the problem of making predictions from the perspective of expanding this joint density to include the unobserved data response y^u . In the context of our exemplary Bayesian model, we move from $p(\mathbf{y}, \mathbf{z}, \boldsymbol{\beta})$ to $p(y^u, z^u, \mathbf{y}, \mathbf{z}, \boldsymbol{\beta})$ noting that the former is obtained from the latter by integration with respect to (y^u, z^u) .

$$\begin{aligned}
 & p(y^u, z^u | y_1, \dots, y_n, z_1, \dots, z_n, \beta_1, \dots, \beta_K) \\
 &= \frac{p(\beta_1, \dots, \beta_K) \prod_{i=1}^n p(y_i | z_i) p(z_i | \beta_1, \dots, \beta_K) p(y^u | z^u) p(z^u | \beta_1, \dots, \beta_K)}{\int p(\beta_1, \dots, \beta_K) \prod_{i=1}^n p(y_i | z_i) p(z_i | \beta_1, \dots, \beta_K) p(y^u | z^u) p(z^u | \beta_1, \dots, \beta_K) d(y^u, z^u)} \\
 &= \frac{p(y^u | z^u) p(z^u | \beta_1, \dots, \beta_K)}{\int p(y^u | z^u) p(z^u | \beta_1, \dots, \beta_K) d(y^u, z^u)} \\
 &= p(y^u | z^u) p(z^u | \beta_1, \dots, \beta_K)
 \end{aligned} \tag{33}$$

For predicting a pair y^u, z^u conditional on $\boldsymbol{\beta}$, we are thus back at data generation, i.e., get a draw z^u from $\mathcal{N}(z^u | (\mathbf{x}^u)' \boldsymbol{\beta})$ and determine y^u according to the sign of z^u . The predictive probability $p(y^u = 1 | \boldsymbol{\beta})$ can be simulated as $\frac{1}{R} \sum_{r=1}^R \mathbf{1}((z^u)^r > 0)$ or computed using Eq. 17.

However, predictions conditional on a particular value of $\boldsymbol{\beta}$ are rarely of interest or relevant because, with finite data and nondegenerate priors, $\boldsymbol{\beta}$ will only be known up to a posterior distribution. As a consequence, $p(y^u, \mathbf{y}, \boldsymbol{\beta}) \neq p(y^u, \mathbf{y})$ where the latter is defined as $\int p(y^u, \mathbf{y}, \boldsymbol{\beta}) d\boldsymbol{\beta}$ which in turn is defined as $\int p(y^u, z^u, \mathbf{y}, \mathbf{z}, \boldsymbol{\beta}) d(z^u, \mathbf{z}, \boldsymbol{\beta})$. The corresponding predictive probability marginalized with respect to latent utility z^u and parameters $\boldsymbol{\beta}$, i.e., $p(y^u = 1 | \mathbf{y})$ can be simulated as:

$$p(y^u = 1 | \mathbf{y}) \approx \frac{1}{R} \sum_{r=1}^R \mathbf{1}((z^u)^r > 0), \quad (z^u)^r \sim \mathcal{N}(z^u | (\mathbf{x}^u)' \boldsymbol{\beta}^r) \tag{34}$$

or more efficiently as:

$$p(y^u = 1 | \mathbf{y}) \approx \frac{1}{R} \sum_{r=1}^R \Phi((\mathbf{x}^u)' \boldsymbol{\beta}^r) \tag{35}$$

in the sense that the approximation to $p(y^u = 1|\mathbf{y})$ in Eq. 35 delivers the same accuracy as that in Eq. 34 based on relatively smaller R . The sample $(\beta^1, \dots, \beta^R)$ to be averaged over is obtained by Gibbs sampling from the posterior distribution $p(\beta|\mathbf{y})$. Note that because of the nonlinearity of the probit likelihood, $p(y^u = 1|\mathbf{y}) \neq p(y^u = 1|\hat{\beta})$ where $\hat{\beta}$ is some point estimate. Specifically, probabilities larger (smaller) than 0.5 will be over- (under-) estimated if posterior uncertainty in β is ignored.

To better appreciate this generally important point, it is useful to simulate probit data following the example given with the `rbprobitGibbs` routine in the R-package `bayesm`, to sample from the corresponding posterior using `rbprobitGibbs`, and then to simulate and compare predictions for different \mathbf{x}^u as explained above. For a comparison with predictions at a frequentist point, estimate the R-command `glm(..., family=binomial(link="probit"),...)` is useful.

Conditional posterior distributions in hierarchical models. Hierarchical models estimate a distribution of response coefficients, e.g., $\{\beta_i\}_{i=1}^N \sim p(\{\beta_i\}_{i=1}^N|\tau)$ from a collection of $i = 1, \dots, N$ time series $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)'$ where $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,t}, \dots, y_{i,T_i})'$. $P(\{\beta_i\}_{i=1}^N|\tau)$ forms a *hierarchical* prior distribution. The difference to a purely subjective prior distribution is that the sample of time series observations contains likelihood information about parameters τ that index the hierarchical prior. In other words, upon placing a subjective prior distribution on τ , the likelihood information contained in the collection of time series will update this prior distribution to the posterior distribution $p(\tau|\mathbf{Y})$.

It should be noted that in these models, marginal posteriors for individual level coefficients, i.e., $p(\beta_i|\mathbf{Y})$ will be biased or “shrunk” towards the hierarchical prior distribution for T_i relatively small or, more precisely, limited individual level likelihood information in $p(\mathbf{y}_i|\beta_i)$ relative to the information about β_i in the hierarchical prior. And it is precisely this situation that motivates the use of hierarchical models in the first place.

However, parameters τ indexing the hierarchical prior can be estimated consistently, and in many marketing applications where the behavior of the particular consumers in the estimation sample is just a means to learning about optimal actions in the population these consumers belong to, $p(\tau|\mathbf{Y})$ is the main target of inference.

The currently popular algorithms for Bayesian inference in a hierarchical model take advantage of the following decomposition of the joint distribution of the data and the parameters which is characteristic, if not definitive of a hierarchical model:

$$p(\mathbf{Y}, \{\beta_i\}_{i=1}^N, \tau) = p(\mathbf{Y}|\{\beta_i\}_{i=1}^N)p(\{\beta_i\}_{i=1}^N|\tau)p(\tau) \tag{36}$$

An important consequence of this decomposition is that, by the rules developed earlier, the conditional posterior distribution of τ does not involve the data \mathbf{Y} as conditioning argument:

$$p(\tau|\mathbf{Y}, \{\beta_i\}_{i=1}^N) = p(\tau|\{\beta_i\}_{i=1}^N) \propto p(\{\beta_i\}_{i=1}^N|\tau)p(\tau) \tag{37}$$

For many popular and useful choices of $p(\boldsymbol{\tau})$, Eq. 37 results in a conjugate update, i.e., a conditional distribution in the form of known distribution we can directly sample from. Perhaps the most prominent example is the model that takes $p(\{\boldsymbol{\beta}_i\}_{i=1}^N | \boldsymbol{\tau}) = \prod_{i=1}^N \mathcal{N}(\boldsymbol{\beta}_i, \overline{\boldsymbol{\beta}}, \mathbf{V}_\beta)$ and uses a so-called Normal-Inverse Wishart prior for $p(\overline{\boldsymbol{\beta}}, \mathbf{V}_\beta)$ that is sometimes rather confusingly referred to as “the H (ierarchical)B(ayes)-model.” Examples are the routines `rhierBinLogit`, `rhierLinearModel`, `rhierMnlRwMixture`, and `rhierNegbinRw`, in the R-package `bayesm` (Rossi et al. 2005) that implement this hierarchical prior (or its finite mixture generalization in the case of `rhierMnlRwMixture`) for collections of time series of binomial logit, linear, multinomial logit, and negative binomial observations, respectively.

One interpretation of this approach towards inference for the parameters in the hierarchical prior is that it relies on the so-called random effects $\{\boldsymbol{\beta}_i\}_{i=1}^N$ as augmented data, similar to the augmentation of latent utilities in the probit model discussed earlier. Different authors have argued that this approach may be sub-optimal depending on the amount of likelihood information at the individual level and the amount of unobserved heterogeneity in $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N)$ (see e.g., Chib and Carlin 1999; Frühwirth-Schnatter et al. 2004). However, practical alternative approaches that apply beyond the special case of conditionally normal individual level likelihood functions coupled with a (conditionally) normal hierarchical prior have yet to be developed.

In the common situation where $p(\mathbf{Y} | \{\boldsymbol{\beta}_i\}_{i=1}^N) = \prod_{i=1}^N p(\mathbf{y}_i | \boldsymbol{\beta}_i)$ and similarly $p(\{\boldsymbol{\beta}_i\}_{i=1}^N | \boldsymbol{\tau}) = \prod_{i=1}^N p(\boldsymbol{\beta}_i | \boldsymbol{\tau})$, we obtain the following conditional posterior distribution for $\boldsymbol{\beta}_i$.

$$p(\boldsymbol{\beta}_i | \mathbf{y}_i, \boldsymbol{\tau}) \propto p(\mathbf{y}_i | \boldsymbol{\beta}_i) p(\boldsymbol{\beta}_i | \boldsymbol{\tau}) \quad (38)$$

$p(\boldsymbol{\beta}_i | \boldsymbol{\tau})$ acts as a usually rather informative prior for $\boldsymbol{\beta}_i$ here. However, as already discussed $\boldsymbol{\tau}$ is not subjectively set but estimated from the data.

For many individual level likelihood functions of interest in marketing, and perhaps most prominently so for the multinomial logit likelihood, the product on the right hand side of Eq. 38 does not translate into a known distribution. A solution to generating draws from distributions with unknown normalizing constants, the Metropolis-Hastings algorithm is discussed next. Finally, if sampling from the distribution in Eq. 38 is computationally expensive, the combination of Eqs. 37 and 38 suggests scope for parallel sampling from the latter for $i = 1, \dots, N$ and then feeding back the updated $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N)$ as conditioning arguments into Eq. 37 and so on.

Metropolis-Hastings. The Gibbs sampler solves the problem posed by a (joint) posterior distribution with unknown normalizing constant if there is a corresponding set of conditional posterior distributions with known normalizing constants. The Gibbs sampler is extremely powerful and in some sense universal if one is content with approximations to the posterior on a discrete grid (Ritter and Tanner 1992). However, a general technique to sample from distributions with unknown normalizing constants known as the Metropolis-Hastings (MH)

algorithm further substantially facilitates real world applications of Bayesian inference. A practically important example in marketing is Bayesian inference for models defined by type-I extreme value error (TIEV) likelihoods, e.g., logit-models, coupled with normal prior distributions for the (regression) coefficients in the likelihood.

The MH-sampler generates a *dependent* sample from some posterior $p(\theta|\mathbf{y})$ according to the following transition rule:

$$\alpha = \min\left(1, \frac{p(\mathbf{y}|\theta^*)p(\theta^*)q(\theta^r)}{p(\mathbf{y}|\theta^r)p(\theta^r)q(\theta^*)}\right), \quad \theta^* \sim q \tag{39}$$

$$p(\theta^{r+1}|\mathbf{y},\theta^r) = \begin{cases} \alpha & \theta^{r+1} = \theta^* \\ 1 - \alpha & \theta^{r+1} = \theta^r \end{cases} \tag{40}$$

On iteration r , the MH-sampler thus transitions from the current “state” or parameter value θ^r to a new state θ^* with probability α . With probability $1 - \alpha$, the current state at iteration $r + 1$ equals that at iteration r , i.e., $\theta^{r+1} = \theta^r$ (Compute α according in Eq. 39, preferably on the log-scale, exponentiate, and compare the result to a draw u from a standard uniform distribution. If $u < \alpha$ move to θ^* , else stay at θ^r , to obtain θ^{r+1}). The so-called candidate value or state θ^* is sampled from the known “candidate generating” or “proposal” density q . Note that the unknown normalizing constant $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$ cancels from Eq. 39.

A remarkable property of this transition rule is that it defines a Markov chain or process with invariant or stationary distribution equal to the posterior distribution $p(\theta|\mathbf{y})$. (A Markov process is a stochastic process in which the future, i.e., the $(r + 1)$ -th value only depends on the value attained in the r -th iteration. All values taken before at the $(r - 1)$ -th, $(r - 2)$ -th, and so on iteration are irrelevant for predicting or generating the $(r + 1)$ -th value.) In practice, this implies that subject to rather weak conditions for the proposal density q , repeated application of the transition rule in Eq. 40 eventually delivers draws from the posterior distribution of the model under investigation, independent of the choice of initial or starting value $\theta^{r=0}$. In other words, after discarding, say the first b values $\theta^1, \dots, \theta^r, \dots, \theta^b$ generated by b applications of Eq. 40 starting from θ^0 , we can use the remaining $R - b$ draws as a representative sample of the posterior distribution.

To better appreciate this point, define the parameter space countably such that we can replace integration by summing over (a potentially infinite number of) countable sets (This is a technicality to avoid measure theoretic complications associated with events “of probability measure zero,” and without loss of generality. The event that a continuous parameter takes a particular value, for example, is an event of probability measure zero because any ε -environment around that value – no matter how small – contains uncountably infinitely many values), and consider a condition known as “detailed balance”:

$$\begin{aligned}
p(\boldsymbol{\theta}_i|\mathbf{y})q(\boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}_j) &= p(\boldsymbol{\theta}_j|\mathbf{y})q(\boldsymbol{\theta}_i)\alpha(\boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}_i) \\
p(\boldsymbol{\theta}_i|\mathbf{y})q(\boldsymbol{\theta}_j) \times & p(\boldsymbol{\theta}_j|\mathbf{y})q(\boldsymbol{\theta}_i) \times \\
\min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_i)}{p(\mathbf{y}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_j)}\right) &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_j)}{p(\mathbf{y}|\boldsymbol{\theta}_j)p(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_i)}\right) \quad (41) \\
\min(p(\boldsymbol{\theta}_i|\mathbf{y})q(\boldsymbol{\theta}_j), p(\boldsymbol{\theta}_j|\mathbf{y})q(\boldsymbol{\theta}_i)) &= \min(p(\boldsymbol{\theta}_j|\mathbf{y})q(\boldsymbol{\theta}_i), p(\boldsymbol{\theta}_i|\mathbf{y})q(\boldsymbol{\theta}_j))
\end{aligned}$$

where the last line establishes that the first two equalities hold. Now rewrite the first line of Eq. 41 as follows:

$$\frac{p(\boldsymbol{\theta}_i|\mathbf{y})}{p(\boldsymbol{\theta}_j|\mathbf{y})} = \frac{q(\boldsymbol{\theta}_i)\alpha(\boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}_j)} \quad (42)$$

Equation 42 makes apparent that the probability of proposing and accepting the move from $\boldsymbol{\theta}_j$ to $\boldsymbol{\theta}_i$ relative to the probability of proposing and accepting the reverse move in the MH algorithm is equal to the ratio of posterior probabilities of the respective target values. Because Eq. 42 holds for all $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \Theta$ where Θ is the parameter space defined by the model under investigation, we have:

$$\begin{aligned}
\sum_{\boldsymbol{\theta}_i} \frac{p(\boldsymbol{\theta}_i|\mathbf{y})}{p(\boldsymbol{\theta}_j|\mathbf{y})} &= \sum_{\boldsymbol{\theta}_i} \frac{q(\boldsymbol{\theta}_i)\alpha(\boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}_j)} \\
p(\boldsymbol{\theta}_j|\mathbf{y}) &= \left(\sum_{\boldsymbol{\theta}_i} \frac{q(\boldsymbol{\theta}_i)\alpha(\boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}_j)} \right)^{-1} \quad (43) \\
\sum_{\boldsymbol{\theta}_i} p(\boldsymbol{\theta}_j|\mathbf{y}) &= \sum_{\boldsymbol{\theta}_j} \left(\sum_{\boldsymbol{\theta}_i} \frac{q(\boldsymbol{\theta}_i)\alpha(\boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}_j)} \right)^{-1} \\
&= 1
\end{aligned}$$

Equation 43 makes intuitive that the collection of moves away from $\boldsymbol{\theta}_j$ and moves returning to $\boldsymbol{\theta}_j$ by the MH sampler eventually represent the posterior support for $\boldsymbol{\theta}_j$ and, because this holds for all values $\boldsymbol{\theta}_j$, the entire posterior support. The “eventual” part of this statement comes from the fact that we may start off the sampler at a parameter value $\boldsymbol{\theta}_j = \boldsymbol{\theta}^0$ in a region of the parameter space Θ with extremely small posterior probability, i.e., in some extreme tail of the posterior distribution. As the MH sampler perhaps very slowly navigates the posterior, i.e., using many iterations depending on the proposal density q , moving into regions of the parameter space with higher posterior support, the draws along the path to that region over-represent the posterior support for these draws in any finite MH sample. This explains why the first b -iterations of the MH sampler that deliver the sequence $\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^r, \dots, \boldsymbol{\theta}^b$ from the arbitrary initial starting value $\boldsymbol{\theta}^0$ need to be discarded as burn-in for the sequence $\boldsymbol{\theta}^{b+1}, \dots, \boldsymbol{\theta}^{b+r}, \dots, \boldsymbol{\theta}^R$ to be representative of the posterior distribution.

Convergence. Unfortunately, there is no simultaneously practical and reliable way to assess the length of the burn-in sample b . I strongly recommend that users of

so-called Monte-Carlo-Markov-Chain (MCMC) techniques that encompass the Gibbs sampler, the MH sampler, as well as collections and combinations, these techniques always take the time to check the convergence behavior of a particular algorithm using simulated data, no matter if the algorithm was designed by someone else or is being newly developed, coded from scratch. In this process, three additional advantages emerge from working with simulated data. First, it forces the researcher to be absolutely clear about his understanding of the data generating process. Second, it delivers an understanding of what informative and less informative data are. Third, it helps with assessing the influence of subjective prior choices.

The investigation of convergence behavior relies on time-series plots of posterior quantities of interest where “time” is measured in iterations of the MCMC sampler. We want these time series plots to look stationary, at least after projecting to the loss from different actions. In other words, at least times series plots of $\mathcal{L}(a, \theta^r)$ need to have converged to stationary sequences over the first b iterations of the sampler. Obviously, the series of $\mathcal{L}(a, \theta^r)$ will converge if the series of parameter draws θ^r converges. However, it sometimes may be easier to assess convergence in $\mathcal{L}(a, \theta^r)$ than in θ^r because the latter often is a high-dimensional object in applied work. In addition, strong posterior dependence between elements of the parameter vector θ may mask convergence to a stable predictive distribution. Interesting examples are “fundamentally over-parameterized” models in the sense that even an infinite amount of data only likelihood-identifies lower dimensional projections of the parameters (see e.g., McCulloch and Rossi 1994; Edwards and Allenby 2003; Wachtel and Otter 2013) (As discussed in section “Bayesian Essentials” above, a proper prior distribution effectively guarantees that the posterior distribution is proper, independent of what can be identified from the likelihood). However, strong posterior dependence between elements of θ^r is not limited to fundamentally over-parameterized models.

If a MCMC explores the posterior distribution quickly (“mixes well”), it will yield a representative sample of the posterior distribution in fewer iterations than a MCMC that explores the posterior distribution more slowly (“does not mix well”). The mixing-behavior of a MCMC has implications for the required length of the burn-in sample b . If a chain mixes well, we can choose vastly different starting values and we will quickly lose the ability to distinguish among chains that use different starting values based on summaries of draws. The information in the draws from the posterior all chains converge to will swamp the initial differences between chains. Reliable formal tests of convergence implemented in the R-package CODA (Plummer et al. 2006), for example, build on this idea. However, when a chain mixes well, the researcher will (almost always) see this when exploring the posterior sample generated by the MCMC graphically. And because chains that mix well converge quickly, this limits the need for formal testing. In applied work, it thus is a priority to make sure that the MCMC employed mixes well. This brings us back to the role of simulated data in the development and testing of numerically intensive inference routines such as MCMC. I will give practical examples further below.

Construction of proposal densities The proposal density q needs to be known in the sense that we need to generate draws from it. In general, we also need to be able

to evaluate the proposal density, i.e., to compute $q(\boldsymbol{\theta})$ when computing α in Eq. 39. However, normalizing constants can be omitted because they cancel from the ratio in α . The best proposal density possible is the posterior distribution itself. Setting $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y})$ the acceptance probability α becomes

$$\begin{aligned} \alpha &= \min \left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)p(\boldsymbol{\theta}'|\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')p(\boldsymbol{\theta}^*|\mathbf{y})} \right) \quad \boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|\mathbf{y}) \\ &= \min \left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')p(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)p(\mathbf{y})} \right) \\ &= \min \left(1, \frac{p(\mathbf{y})}{p(\mathbf{y})} \right) \\ &= 1 \end{aligned} \tag{44}$$

However, the reason for using the MH sampler in the first place is that we cannot directly sample from the posterior distribution (Note that one can think of the Gibbs sampler as a cycle through MH steps with conditional proposal densities equal to the conditional posterior distributions). Nevertheless, it is sometimes possible to construct proposal densities as close approximations to the posterior distributions. An example is the routine `rmnl1ndepMetrop` in the R-package `bayesm` (Rossi et al. 2005) that uses a normal approximation to the likelihood to construct a multivariate t -distributed proposal centered at a penalized maximum likelihood estimate.

An obvious requirement for the proposal density is that the parameter set over which the proposal density q has positive support Θ_q is equal to, or a superset of the parameter set over which the posterior distribution has positive support, i.e., $\Theta_{p(\boldsymbol{\theta}|\mathbf{y})} \subseteq \Theta_q$. If the proposal density q is such that parameter values that have positive support under the posterior distribution can never be reached, an MH sampler using this proposal density cannot possibly deliver draws that are representative of the posterior distribution.

Conversely, if the proposal density extends beyond the support of the posterior, i.e., $\Theta_{p(\boldsymbol{\theta}|\mathbf{y})} \subset \Theta_q$, proposals to move into a region of the parameter space that is not supported under the posterior will simply be rejected. The corresponding acceptance probability α is equal to zero (see Eq. 39).

A related, less obvious but nevertheless practically important requirement for the proposal density is that it should have more mass in its tails relative to the posterior distribution. The reason is that a concentrated proposal density may effectively fail to navigate the entire posterior distribution in a way similar to a proposal that is only defined over a subset of the parameters space. A tricky aspect of thin tailed proposal densities, and concentrated in an area where the posterior distribution is relatively flat, is that time series plots of any finite number of MH draws may fail to indicate that the sampler has not converged, i.e., the plots may indicate convergence over a range of parameters that is not representative of the entire posterior distribution.

A simple recipe to specifying a proposal that necessarily has more mass in the tails relative to the posterior distribution is to define q as a random walk (RW), i.e., $\boldsymbol{\theta}^* = \boldsymbol{\theta}' + \boldsymbol{\epsilon}$ with $q(\boldsymbol{\epsilon})$ defined such that $q(\boldsymbol{\epsilon}) = q(-\boldsymbol{\epsilon})$ for all $\boldsymbol{\theta}^* \in \Theta_q$. This recipe

works for continuous and discrete distributions, and both for multivariate and univariate posterior distributions, in principle. Based on a RW proposal, the MH acceptance probability α simplifies to

$$\begin{aligned}
 \alpha &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^r)}{p(\mathbf{y}|\boldsymbol{\theta}^r)p(\boldsymbol{\theta}^r)q(\boldsymbol{\theta}^*)}\right), \boldsymbol{\theta}^* = \boldsymbol{\theta}^r + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim q \\
 &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^r + \boldsymbol{\epsilon})p(\boldsymbol{\theta}^r + \boldsymbol{\epsilon})q(\boldsymbol{\theta}^* - \boldsymbol{\theta}^r)}{p(\mathbf{y}|\boldsymbol{\theta}^* - \boldsymbol{\epsilon})p(\boldsymbol{\theta}^* - \boldsymbol{\epsilon})q(\boldsymbol{\theta}^r - \boldsymbol{\theta}^*)}\right) \\
 &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^r + \boldsymbol{\epsilon})p(\boldsymbol{\theta}^r + \boldsymbol{\epsilon})q(\boldsymbol{\epsilon})}{p(\mathbf{y}|\boldsymbol{\theta}^* - \boldsymbol{\epsilon})p(\boldsymbol{\theta}^* - \boldsymbol{\epsilon})q(-\boldsymbol{\epsilon})}\right) \\
 &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta}^r)p(\boldsymbol{\theta}^r)}\right)
 \end{aligned}
 \tag{45}$$

However, in many applications, the dimensionality of the parameter space is too large for a RW proposal that attempts to move all parameters simultaneously in one “big” MH step to work well. Conditional independence relationships in the DGP can be exploited to break one big MH step into a collection of MH steps of smaller dimensionality following the same logic that we used earlier to decompose the joint posterior distribution into a set of more manageable conditional posterior distributions for the Gibbs sampler.

In fact, the MH sampler delivers draws from conditional posterior distributions automatically if we propose to only change an individual element of the parameter vector, say θ_k :

$$\begin{aligned}
 \alpha &= \min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}_{-k}^r, \theta_k^*)p(\boldsymbol{\theta}_{-k}^r, \theta_k^*)q(\boldsymbol{\theta}^r)}{p(\mathbf{y}|\boldsymbol{\theta}^r)p(\boldsymbol{\theta}^r)q(\boldsymbol{\theta}_{-k}^r, \theta_k^*)}\right), \theta_k^* \sim q(\theta_k|\boldsymbol{\theta}_{-k}) \\
 &= \min\left(1, \frac{p(\theta_k^*|\mathbf{y}, \boldsymbol{\theta}_{-k}^r)q(\boldsymbol{\theta}_{-k}^r|\boldsymbol{\theta}_{-k}^r)q(\boldsymbol{\theta}_{-k}^r)}{p(\theta_k|\mathbf{y}, \boldsymbol{\theta}_{-k}^r)q(\theta_k^*|\boldsymbol{\theta}_{-k}^r)q(\boldsymbol{\theta}_{-k}^r)}\right) \\
 &= \min\left(1, \frac{p(\theta_k^*|\mathbf{y}, \boldsymbol{\theta}_{-k}^r)q(\boldsymbol{\theta}_{-k}^r|\boldsymbol{\theta}_{-k}^r)}{p(\theta_k|\mathbf{y}, \boldsymbol{\theta}_{-k}^r)q(\theta_k^*|\boldsymbol{\theta}_{-k}^r)}\right)
 \end{aligned}
 \tag{46}$$

The second line in Eq. 46 follows from the application of Bayes’ theorem (see Eq. 1 and note that normalizing constants $\int p(\mathbf{y}|\boldsymbol{\theta}_{-k}^r, \theta_k)p(\boldsymbol{\theta}_{-k}^r, \theta_k)d\theta_k$ cancel) and the decomposition of the joint proposal density into a conditional times a marginal. However, it is wasteful not to exploit conditional independence relationships that often vastly simplify the computation of the ratio in Eq. 46 for particular conditional posterior distributions (see e.g., the conditional posterior distribution in Eq. 31).

Moreover, unobservables that are conditionally independent a posteriori should always be drawn in separate MH steps, upon introducing the respective conditioning argument. It would be wasteful to constrain the sampler to either accept a joint move

of all these unobservables to the respective candidate values or to reject the entire move and to repeat all respective values from iteration r . The conditional posterior $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^n p(z_i|\boldsymbol{\beta}, y_i)$ discussed earlier in the context of the binomial probit likelihood serves as an example.

The practical advantage of working with full conditional distributions as the basis for MH-RW sampling is that the proposal densities $q_k(\epsilon)$ are univariate. As a consequence, we only need to determine the concentration of these distributions around $\epsilon_k = 0$, which corresponds to $\theta_k^* = \theta_k^r$. When attempting to make multivariate proposals with the goal to move more than one element of the parameter vector in one step, a simple multivariate RW proposal of the form $q(\boldsymbol{\epsilon}) = \prod_{k=1}^K q(\epsilon_k)$ may suggest moves into directions with minimal support under the posterior which will result in $\boldsymbol{\theta}^{r+1} = \boldsymbol{\theta}^r$ for many iterations. Thus, setting up an MCMC as a repeated cycle through conditional MH steps facilitates the definition of suitable proposal densities. This is analogous to conditioning leading to known distributions in the Gibbs sampler, which can be viewed as a special case of MH sampling (see Eq. 44).

For continuous parameters the default choice for $q_k(\epsilon)$ is $N(0, \sigma_k^2)$ where the parameter σ_k^2 is subject to “tuning” by the analyst. For an integer parameter $\epsilon = (\eta + 1)s$ could be used, where η is distributed Poisson with tuning parameter λ , and s takes values from $\{-1, 1\}$ with probability 0.5 (For strictly categorical parameters with no ordering among their values, the notion of a random walk is not defined. However, because of the finite prior support of such parameters, it is possible to use discrete uniform proposal distributions. Because all values have the same probability under a uniform distribution, the proposal distributions again cancel from the ratio in the acceptance probability α).

The tuning parameter implicitly specifies an average size of ϵ and thus an average distance between θ_k^* and θ_k^r (also known as the step-size of the proposal distribution), ϵ small in absolute value result in θ_k^* close to θ_k^r that are more likely accepted, i.e., $\theta_k^{r+1} = \theta_k^*$ than ϵ large in absolute value that will more likely result in $\theta_k^{r+1} = \theta_k^r$ when applying Eq. 40. If the number of total iterations R to run the MH sampler were of no concern, any setting of the tuning parameters that results in nondegenerate $q_k(\epsilon)$ would result in valid posterior inferences based on applications of Eq. 40.

However, both ϵ that are too small on average and ϵ that are too large on average will result in MH samplers that require a larger number of total iterations R to deliver the same *amount* of information about the posterior distribution than “optimally sized” ϵ . The situation is analogous to studying a population based on sampling. Larger samples result in more reliable inference and some sampling techniques result in higher statistical efficiency than others based on the same number of observations. Here, the population is the posterior distribution, the proposal density plays the role of the sampling plan, and importantly the sample size R is under our control, within the limits set by computational speed and time.

When the tuning parameter is set such that ϵ is too small on average, the MH sampler will explore the posterior in local neighborhoods extensively and navigate the entire posterior over many, many small steps creating “large swings” such that time series plots look like those of financial indices that can move into one direction

for extended periods of times, in this case potentially for tens of thousands of iterations. The consequence is that the chain may appear as if it does not converge to a stationary distribution at all.

When the tuning parameter is set such that ϵ is too large on average, the chain will remain at the same value for many iterations and may fail to move at all, i.e., never accept to set $\theta_k^{r+1} = \theta_k^*$. However, if it at least moves sometimes, such a chain will arrive at a region of relatively large posterior support in large jumps and tend to stay there. In that sense, ϵ that are too large – provided that the chain moves at all – are the lesser evil. However, any reliable statements about posterior uncertainty based on a finite number of MH draws require decently tuned proposal densities. In practice, some experimentation is required that again is supported by the analysis of simulated data.

To illustrate, I simulated 500 observations from a binomial-probit model with data generating parameter vector $\beta = (-3, 2, 4)$. The first coefficient is an intercept and the remaining two are slope coefficients for two randomly uniformly distributed covariates (see the [Appendix](#) for the corresponding R-script). The script calls a simple, stylized RW-MH-sampler for a binomial probit model coupled with a multivariate normal prior for the probit coefficients implemented in plain R (see the function `rbprobitRWMetropolis` in the [Appendix](#)).

I ran the MCMC for 200,000 iterations using a weakly informative prior and initializing the chain at $\beta^{r=0} = (0,0,0)$. Fig. 2 shows MCMC-traces of β for four different $q(\epsilon) = \prod_{k=1}^K N(0, \sigma_k^2)$, i.e., $\sigma_k = 0.001$, $\sigma_k = 0.005$, $\sigma_k = 0.2$, and finally $\sigma_k = 3$ for all $k = 1,2,3$. These step-sizes translate into average acceptance rates α of RW-proposals of 99%, 97%, 25%, and 0.05% (see Eq. 45). The black, red, and green MCMC-traces correspond to the first, second, and third element of the parameter vector, respectively.

The top-left plot in Fig. 2 depicts the MCMCs that use the smallest step-size investigated here. It presents an example of an MCMC-trace from a sampler that has not converged to delivering samples from the posterior distribution. All three traces exhibit a trend away from zero over the entire course of the 200,000 iterations the sampler was run. Looking at the y-axis, we see that the individual traces are nowhere near the data generating values and reflective of the starting values, even in the last iteration. In an application to real data, we would not now what the data generating parameter values are to compare. However, upon seeing something similar to the top-left plot, we would conclude that the sampler has not converged to a stationary distribution yet. Thus, summaries of the full set or any subset of the 200,000 draws in the top-left plot do not represent the posterior distribution.

The traces in the top-right plot are with a step-size σ_k that is five times larger than that in the top-left plot. We see that the three traces appear to converge to stationarity around iteration 50,000 or so, and we could use summaries of the last 150,000 draws to learn about the posterior distribution. With an even larger $\sigma_k = 0.2$, convergence to the stationary distribution is much quicker (see the bottom-left plot). Finally, when we use $\sigma_k = 3$, the largest MH step-size investigated here, we see that the MCMC relatively quickly jumps into the neighborhood of the data generating β , but sticks to the same parameter value, often for thousands of iterations.

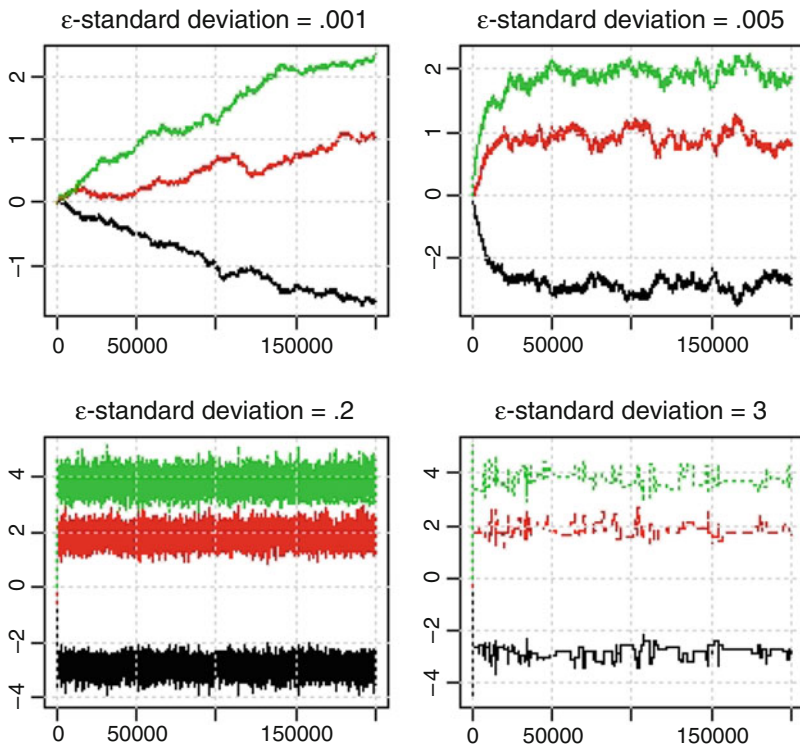


Fig. 2 MH-sampling – different step-sizes, 200,000 iterations

From theory we have that all four MCMC chains investigated here will eventually represent the posterior distribution $p(\beta|y)$ equally well, when run for an infinite number of iterations. The concept of an infinite number of iterations is not helpful in practice. However, to illustrate convergence of traces even with poorly tuned MH-steps, I ran each chain for 400,000 more iterations. Figure 3 depicts MCMC-traces obtained by stringing together the first 200,000 iterations from Fig. 2 with the subsequent 400,000 for a total of 600,000 iterations. It can be seen that all four MH-samplers converge eventually, even the sampler that uses $\sigma_k = 0.001$.

However, convergence of the MCMC to its stationary distribution is a necessary but not a sufficient criterion for high-quality inferences about the posterior distribution based on any finite sample of MCMC draws. To illustrate this point, Fig. 4 zooms into the last 50,000 iterations of the 600,000 total iterations from each sampler. Intuitively, the collection of draws in the bottom-left contain most information about the posterior, followed by that in the top-right. It is harder to order the collection of draws in the top-left and the bottom-right according to their information content by visual inspection.

Table 6 summarizes the traces depicted in Fig. 4 numerically, i.e., the last 50,000 draws from each chain. We see reasonable agreement between the chains operating with step-sizes of 0.005, 0.02, and 3 in terms of posterior means. However, the

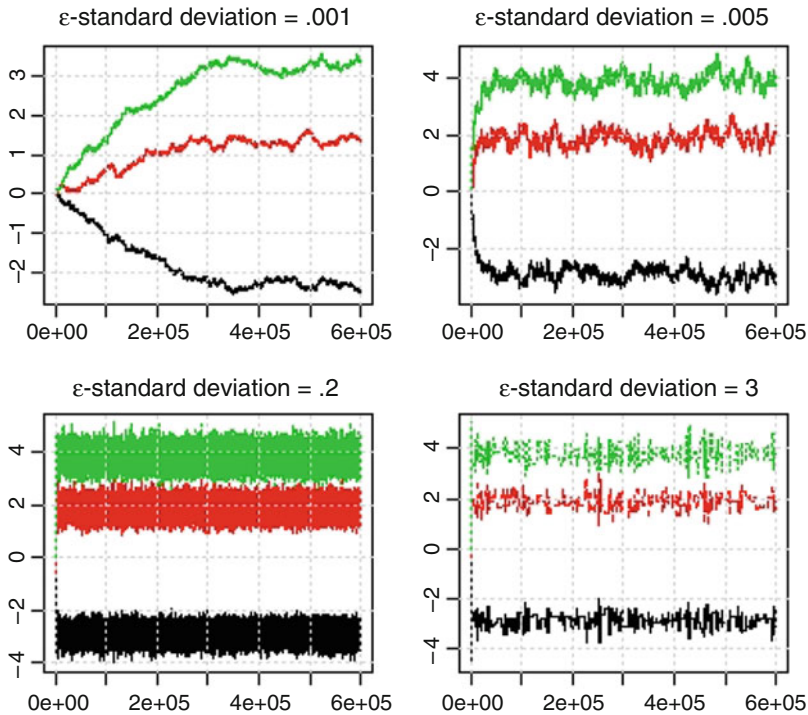


Fig. 3 MH-sampling – different step-sizes, 600,000 iterations

chains with step-sizes of 0.005 and 3 underestimate the posterior standard deviations relative to that with a step-size of 0.2 based on the last 50,000 draws (The posterior standard deviations of MCMC draws measure the posterior uncertainty in the knowledge about the parameters to be estimated. The analogy to frequentist standard errors applies. However, while reasonable estimates of frequentist standard errors maybe hard to come by in finite samples, posterior standard deviations are well defined automatically by virtue of using proper priors. In addition, based on a sample from the posterior distribution, posterior standard deviations of functions of parameters are easily computed as the standard deviation of functional values computed at each draw from the posterior). The chain with step-size 0.001, which required about 350,000 draws to converge to stationarity (see the top-left plot in Fig. 3), results in different means and dramatically smaller posterior standard deviations when looking at the last 50,000 draws.

Table 7 reports analogous summaries, but now based on the last 250,000 draws (compare Fig. 3). Based on these five times larger samples from the posterior, we see reasonable agreement between chains with step-sizes 0.005, 0.2, and 3 both in terms of posterior means and posterior standard deviations. This again illustrates that MCMC will “always work,” if we only run the chains for long enough. However, it also illustrates that some MCMCs deliver more information about the posterior holding the number of iterations fixed than others, and that a valid MCMC chain can be practically useless if it explores the posterior too slowly.

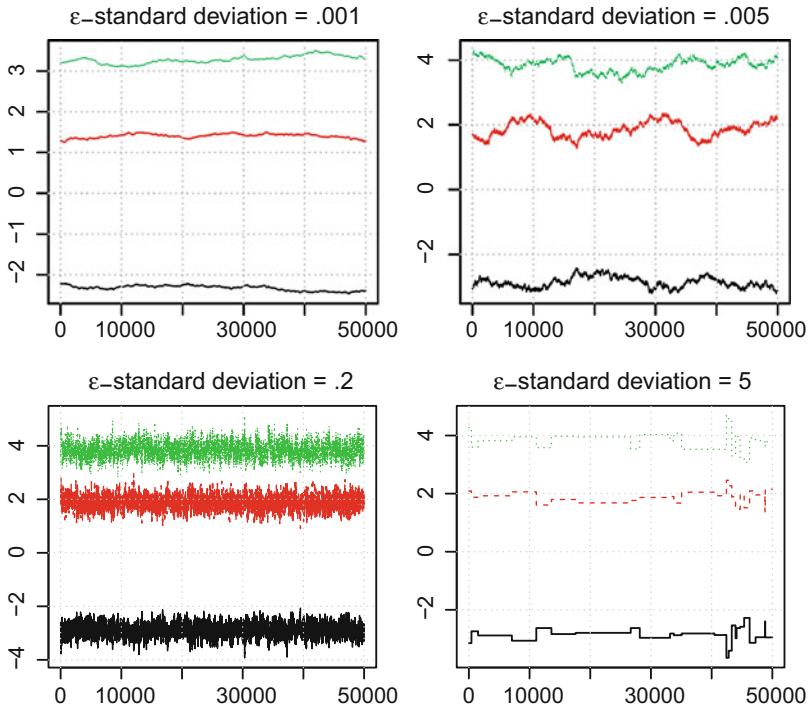


Fig. 4 MH-sampling – different step-sizes, last 50,000 of 600,000 iterations

Table 6 Posterior means and standard deviations from the last 50,000 iterations

Step-size	Mean			Standard deviation		
	β_0	β_1	β_2	β_0	β_1	β_2
0.001	-2.32	1.41	3.28	0.06	0.05	0.09
0.005	-2.86	1.85	3.84	0.16	0.25	0.21
0.2	-2.89	1.89	3.83	0.24	0.26	0.30
0.3	-2.85	1.87	3.82	0.17	0.18	0.25

Table 7 Posterior means and standard deviations from the last 250,000 iterations

Step-size	Mean			Standard deviation		
	β_0	β_1	β_2	β_0	β_1	β_2
0.001	-2.28	1.34	3.23	0.10	0.10	0.13
0.005	-2.92	1.90	3.89	0.23	0.27	0.29
0.2	-2.88	1.88	3.83	0.24	0.26	0.30
3	-2.90	1.94	3.82	0.23	0.23	0.32

Finally, I illustrate the notion of exploring the posterior distribution more quickly (more efficiently) and more slowly (less efficiently) by comparing the RW-MH-chains with step-sizes 0.005, 0.2, and 3 to each other, and to posterior draws from the Gibbs-sampler that relies on data-augmentation discussed earlier (`rbprobitGibbs` in the

R-package `bayesm`). I focus on the first slope coefficient (the red trace in the figures above), and compute means and standard deviations from batches of 1000 consecutive draws starting from iteration 50,001 until iteration 600,000. The histograms in Figs. 5 and 6 summarize the resulting distributions of 550 ($= (600,000-50,000)/1000$) batch means and batch standard deviations for RW-MH-chains with step-sizes 0.005, 0.2, and 3 and the Gibbs-sampler.

Assuming the true posterior standard deviation (from a hypothetical infinite run of the MCMC) to be about 0.26 (see Table 7), we would expect the batch means to be distributed normally around the true mean with standard deviation $.26/\sqrt{1000} \approx .008$ simply because we cannot learn the exact mean of a non-degenerate posterior distribution from a finite sample. This translates into a $5\text{-}\sigma$ interval around the mean with a length of about 0.08. Any excess variation in batch means is evidence of the inefficiency of the employed sampling technologies relative to a hypothetical *iid*-sampler. From the x -axes in Fig. 5, we can see that batch means are distributed much more widely. Intuitively, a single 1000-iterations batch from each of the MCMCs is less informative about the posterior (more likely to summarize information from only parts of the posterior) than 1000 draws from a hypothetical *iid*-sampler. In addition, if someone had to bet on the inference from a randomly

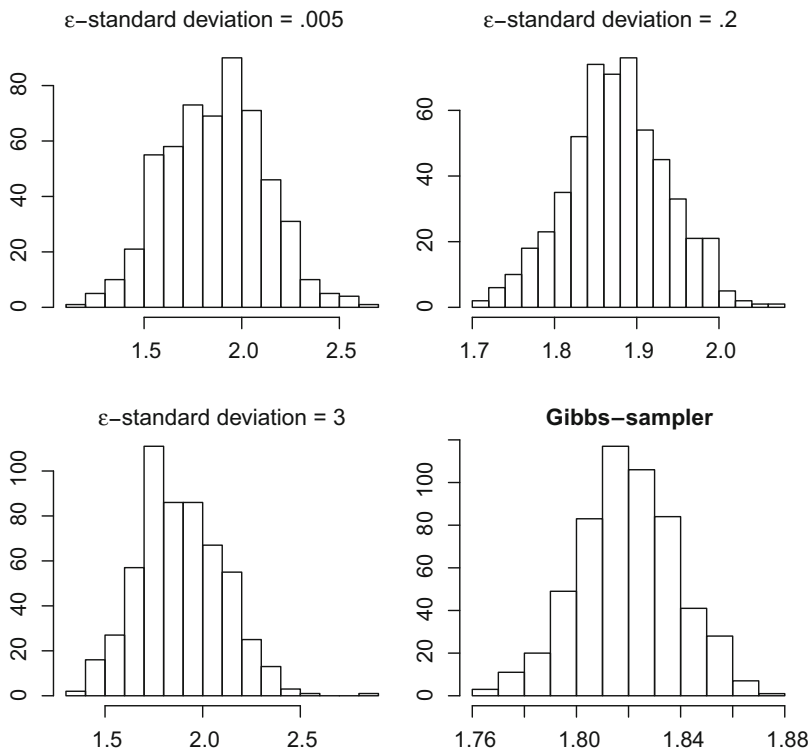


Fig. 5 MH-sampling – distribution of batch means from different step-sizes compared to Gibbs-sampling

drawn single batch, he would prefer a draw from the Gibbs-sampler (in the bottom-right), followed by a draw from the RW-MH-sampler with step-size 0.2 (in the top-right). The decision between step-sizes 0.005 and 3 is less clear. However, from the wider distribution of batch means, it is obvious that MCMCs with these step-sizes explore the posterior less efficiently.

Finally, the batch standard deviations in Fig. 6 again identify the Gibbs-sampler as most efficient, followed by the RW-MH-chain with step-size 0.2. A randomly drawn batch of 1000 consecutive draws from these samplers is likely to yield a posterior standard deviation close to the posterior standard deviation estimated from all $600,000 - 50,000 = 550,000$ draws. In addition, the top-left plot in Fig. 6 demonstrates that each and every single 1000 consecutive iterations batch from the chain with step-size 0.005 substantially underestimates the posterior standard deviation. In contrast, the chain with the (too) large step-size of 3 often suggests no posterior uncertainty at all – when no proposal is accepted in the batch – but does not uniformly underestimate the posterior standard deviation. This again suggests that chains with step-sizes that are too small are potentially more misleading than chains with step-sizes that are too large.

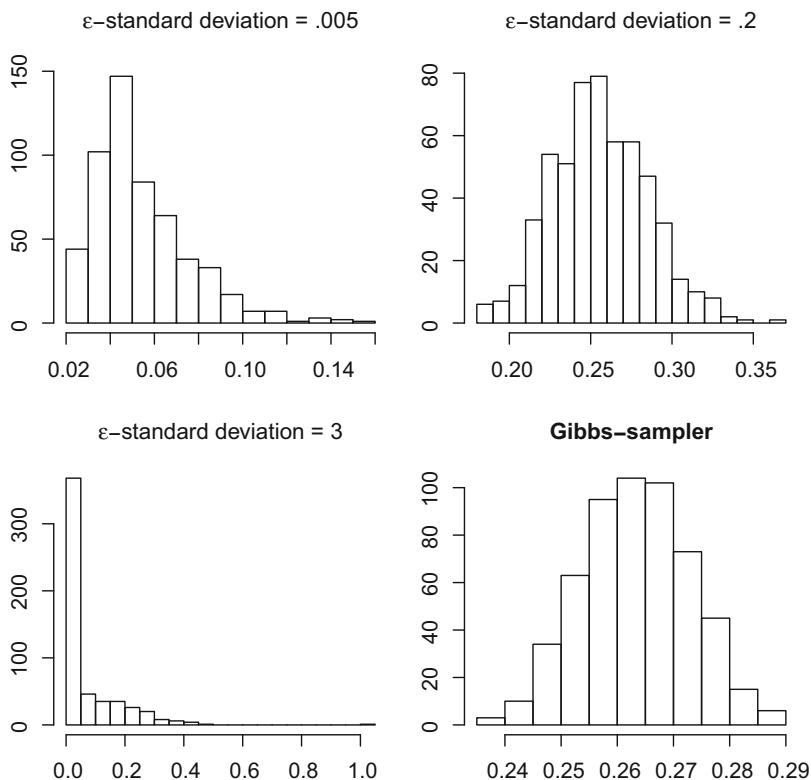


Fig. 6 MH-sampling – distribution of batch standard deviations from different step-sizes compared to Gibbs-sampling

The examples discussed here nicely showcase that the emphasis in applied work should be on using, devising sampling schemes that mix well, before even considering the formal assessment of convergence. In a sense, it is almost always obvious from a graphical inspection of MCMC-trace plots whether a sampler that mixes well has converged or not.

For first time acquaintances with MH sampling I suggest the following additional coding exercises to develop an intuition for MH-sampling based on personal experience:

1. Change the function `rbprobitRWMetropolis` in the [Appendix](#) to cycling through MH-steps that update individual elements of the parameter vector one at a time from their conditional posterior distributions. Experiment with tuning RW-proposals for each element of the parameter vector independently.
2. Obtain a copy of the “plain R” version of `rbprobitGibbs` (version 2.2–5 of `bayesm` available from the CRAN-archives), replace the part that generates latent utilities \mathbf{z} in line 141 with RW-MH steps, and verify with simulated data that this new algorithm works. The setup is generally interesting, because it is a toy version of a hierarchical model with MH-updates at the lower level and conjugate updates of parameters that form the hierarchical prior.
3. Modify this sampler such that you propose candidate values z_i^* from their (hierarchical) prior distribution $\mathcal{N}(\mathbf{x}; \boldsymbol{\beta}, 1)$. Note that the proposal and the prior distribution will cancel from the ratio in the MH-acceptance probability α . You will likely see that this sampler does not converge to a posterior distribution $p(\boldsymbol{\beta} | \mathbf{y})$ anywhere near the data generating values, even though the time series of $\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^r, \dots, \boldsymbol{\beta}^R$ suggests immediate convergence and superior mixing! This is an example of the drawbacks of a (collection of) proposal densities that do not have enough mass in their tails.

Recent developments. An important recent development in the context of making numerically intensive Bayesian analysis more practical is the No U-turn Sampler (NUTS) by Hoffman and Gelman (2014) which is a self-tuning Hamiltonian-Monte-Carlo sampler (see e.g., Neal 2011). This technique has been implemented in `Stan` (Carpenter et al. 2017) which interfaces with many popular software environments including R, Python, Matlab, and Stata, for example.

The basic principle of Hamiltonian-Monte-Carlo (HMC) is to leverage Hamiltonian dynamics for a more effective exploration of the posterior. In physics, Hamiltonian dynamics describe the change in location and momentum of an object by differential equations. The solutions to the differential equations yield the location and the momentum of an object at any particular point in time.

In HMC, the locations correspond to value of the q -element parameter vector to be estimated. Each location is associated with a potential energy and the statistical analogue is the negative of log-posterior evaluated these values (Thus, the posterior mode is the point of lowest potential energy we would gravitate to in the absence of “extra” kinetic energy that enables movements away from this point). The analogue to the momentum comes from expanding the parameter space by p additional

parameters (where $p = q$), the negative log-density of which is the statistical analogue of kinetic energy (Thus, again the mode of this density is the point (the momentum vector) with the lowest kinetic energy). Usually, these additional parameters are assumed to be standard normally distributed. However, it should be noted that the p additional parameters and their density are purely technical devices to complete the Hamiltonian. Similarly, proposal distributions in the context of MH-sampling are technical devices to accomplish MH-sampling.

The algorithm first draws a p -element “momentum” vector from standard normal distributions. The momentum vector both defines the direction of the movement away from the current location (parameter value), and the maximum distance that can be realized, as explained next. HMC obeys the principle that the total energy, i.e., the sum of the potential and the kinetic energy is constant in the closed system described by the Hamiltonian, when deriving a new location (and a new momentum) at any point in time (see Eq. 2.12 in Neal 2011). Here time refers to some arbitrary time point after the onset of the momentum that generates a movement away from the current location.

The location-change is a function of the change in kinetic energy and the momentum-change a function of the change in potential energy. Note that the change in potential energy corresponds to the gradient of the negative log-posterior, and the change in kinetic energy to the gradient of the negative log-density of auxiliary momentum variables respectively, in statistical applications. If the differential equations describing the change in position and momentum could be solved exactly, one could solve for the location that is furthest away from the current location that can be reached in the direction of the current draw of the momentum, given its associated kinetic energy, define this as the new location, draw a new momentum vector, and so on.

It is useful to contemplate how such a procedure would explore the posterior. With a fixed distribution of momentum vectors (and corresponding kinetic energies), it would tend to move away more slowly from a pronounced posterior mode, i.e., in smaller steps in expectation, because of the steep increase in potential energy (defined as the negative of the log-posterior) around this mode. Here, the expectation is with respect to the fixed distribution of momentum vectors (and corresponding kinetic energies). Only outlying momentum vectors would supply sufficient kinetic energy to move far into directions of (much) higher potential energy. Conversely, it would tend to move more quickly, i.e., in larger steps in expectation, through areas of high-potential energy (small values of the log-posterior), and in the direction of low potential energy, in expectation. It is therefore somewhat intuitive that such a procedure would result in direct draws from the posterior that could represent the posterior effectively based on a relatively small number of draws. In contrast to RW-MH-sampling, the distance between two successive draws from this procedure would automatically reflect the concentration of the posterior at every value of the parameter space.

However, in practice, the solutions to the differential equations defining the Hamiltonian dynamics need to be approximated in discretized time. Again, time here refers to the time after the onset of the momentum that generates a movement away from the current location, i.e., the current parameter value. A discrete approximation that can be tuned to high accuracy (relative to the exact solution) is leapfrog integration. At each iteration of the HMC, L leapfrog steps that each correspond to a

discrete time step of length ϵ are performed. Ideally, the number of steps L and the length of each step ϵ are chosen so that the new location (a new parameter value) is as far away as possible from the current parameter value, given the current draw of the p-element momentum vector and its associated kinetic energy, while keeping the approximation error low. Any remaining approximation error is controlled in a MH-step that compares the value of the Hamiltonian at the new position and the momentum at this position to the value of the Hamiltonian at the old position and the momentum vector that initiated the movement to the new position (In other words, the potential energy at the new location and the (remaining) kinetic energy are compared to the potential energy at the old location and the kinetic energy that brought about the movement to the new location). By the law of conservation of energy in the closed system described by the Hamiltonian, the Hamiltonian would evaluate to the same value if the discrete time approximation were exact.

NUTS automatically tunes L , ϵ , and additional parameters that rescale the kinetic energy in different dimensions of the log-posterior to arrive at a highly effective HMC-sampler that does not normally require user intervention. Thus, the researcher can fully concentrate on specifying the model, i.e., the likelihood and the prior, knowing that high quality numerical inference from the implied posterior is available through NUTS. A limitation is that the gradient of the log-posterior needs to be defined, which excludes discrete variables as direct objects of inference. However, in many models, discrete latent variables are introduced as augmented data, such as in models defining a discrete mixture of distributions. In these cases, NUTS could be used to sample from the posterior marginalized with respect to discrete latent variables. Based on the marginal posterior, the posterior distribution of discrete latent variables can be easily derived.

To numerically illustrate the performance of NUTS, I revisit the binomial probit example discussed earlier. I run the NUTS implemented in `Stan` for 600,000 iterations and compute 550 batch means and batch standard deviations of the first slope coefficient (the red trace in Figs. 2 to 4) from the last 550,000 iterations (see Fig. 7). A comparison between Fig. 7 and Figs. 5 and 6 shows that a randomly drawn batch of 1000 consecutive iterations from NUTS is likely to be a better representation of the posterior than a randomly drawn batch of 1000 consecutive iterations from the samplers discussed earlier, including the Gibbs-sampler. However, it should be noted that each NUTS-iteration is more computationally intensive than one iteration of the MH-sampler investigated. The computational intensity of Gibbs-sampling relative to NUTS in this model depends on the sample size, where larger samples are likely to favor NUTS because of the need to augment latent utilities for all observations when Gibbs-sampling.

Model comparison

In the introduction, I mentioned the possibility of determining the dimensionality of a flexibly formulated model using the Bayesian approach. I also alluded to the possibility of making comparisons across different models for the *same* data, where models may arbitrarily differ in terms of likelihood functions, prior

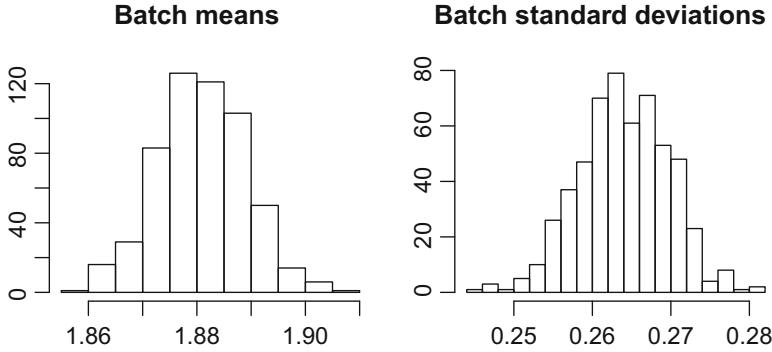


Fig. 7 No U-Turn-sampling – distribution of batch means and batch standard deviations

specifications, or both. Here, I will briefly describe the basic principles to this end. Specifically, I will show how the Bayesian approach can deliver consistent evidence for a more parsimonious model. As usual, consistency means convergence to the data generating truth as the sample size increases (When the set of models compared does not contain the model that in fact corresponds to the data generating truth, consistency means convergence to the model that is closest to the data generating truth in a predictive sense).

This contrasts with the classical frequentist approach, where we can only “fail to reject” relatively simpler descriptions of the world, i.e., more parsimonious theories and models in comparison to more complex models. I personally see this as a drawback of the classical frequentist approach because theory aimed at understanding the underlying causal mechanisms of observed associations generally thrives on establishing that particular (direct) causal effects do not exist.

The Bayesian approach towards comparing between two or more alternative models builds – as one may expect – on Bayes’ theorem. Consider a set of models $\mathcal{M}_1, \dots, \mathcal{M}_K$ formulated for the same observed data \mathbf{y} . Note that this encompasses the possibility that models use different sets of covariates, different likelihood functions, different priors, or may be calibrated including additional or even different data \mathbf{y}' , as long as they define a predictive density for the *same* \mathbf{y} (For example, Otter et al. (2011) show how to derive a marginal likelihood for demand data in a model that specifies a joint density for supply side variables (that enter the demand model as conditioning arguments) and demand data). Bayesian model comparisons then rest on the posterior probabilities of a model given the (focal) data (Eq. 47).

$$\Pr(\mathcal{M}_j|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{M}_j)\Pr(\mathcal{M}_j)}{\sum_{k=1}^K p(\mathbf{y}|\mathcal{M}_k)\Pr(\mathcal{M}_k)} \tag{47}$$

Here $\Pr(\mathcal{M}_k)$ is the subjective prior probability that model k is the true model which is often chosen to be $1/K$ in the absence of better knowledge, and $p(\mathbf{y}|\mathcal{M}_k)$ is the so-called marginal likelihood of the data given model k defined as $\int p_k(\mathbf{y}|\boldsymbol{\theta})p_k(\boldsymbol{\theta}) d\boldsymbol{\theta}$. The subscript k indicates that the likelihood and the prior and thus the “content”

of θ can be model dependent. If $\Pr(\mathcal{M}_k)$ can be reduced to one and the same constant for all models under consideration, this constant can obviously be ignored in Eq. 47. Then, the comparison between any two models k and j in the set can be based on so-called Bayes' factors, defined as ratios of marginal likelihoods (Eq. 47).

$$BF_{k,j} = \frac{p(\mathbf{y} | \mathcal{M}_k)}{p(\mathbf{y} | \mathcal{M}_j)} \tag{48}$$

By convention, Bayes Factors larger 3 count as weak but sufficient evidence in favor of the model in the numerator; Bayes Factors larger 20 count as strong evidence (Kass and Raftery 1995). I will comment more on this convention later.

For example, it would be perfectly alright to compare model k with marginal likelihood $\int p_k(\mathbf{y}|\theta)p_k(\theta)d\theta$ to a model j that introduces observed conditioning arguments (predictors, covariates) \mathbf{X} , i.e., $\int p_j(\mathbf{y}|\mathbf{X}, \theta)p_j(\theta)d\theta$, or to include model i that uses additional data \mathbf{y}' for calibration in the comparison, based on $\int p_i(\mathbf{y}|\theta)p_i(\theta|\mathbf{y}')d\theta$ (However, note that $\int p_i(\mathbf{y}, \mathbf{y}'|\theta)p_i(\theta)d\theta \neq \int p_i(\mathbf{y}|\theta)p_i(\theta|\mathbf{y}')d\theta$. The former is a marginal likelihood for the data $(\mathbf{y}, \mathbf{y}')$ and not for the data \mathbf{y} . Marginal likelihoods for different models can only be directly compared as long as they pertain to the *same* data). A useful intuition for marginal likelihoods is that they reduce radically different, per se incomparable "stories" about what may have generated the data to densities for the data, which are directly comparable in the same way as we can compare predictions for the *same* event completely independent of the considerations that gave rise to the prediction.

However, we still need to establish the intuition for how Bayesian model comparisons can possibly consistently support the more parsimonious model. I will do this by returning to the regression example from Eq. 10. Recall that we were able to derive the posterior distribution analytically in this example (see Eq. 16). Exploiting this fact, we obtain an analytical expression for the marginal likelihood of the data under this model as follows:

$$\begin{aligned} p(\mathbf{y}|\beta^0, \Sigma^0) &= \frac{p(\mathbf{y}|\beta)p(\beta|\beta^0, \Sigma^0)}{p(\beta|\mathbf{y}, \beta^0, \Sigma^0)} \\ &= \frac{\frac{1}{\sqrt{2\pi}} \prod_{i=1}^n \exp\left(-\frac{1}{2}(y_i - x'_i\beta)^2\right) (2\pi)^{-k/2} |\Sigma^0|^{-1/2} \exp\left(-\frac{1}{2}(\beta - \beta^0)'(\Sigma^0)^{-1}(\beta - \beta^0)\right)}{(2\pi)^{-k/2} |\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1}|^{1/2} \exp\left(-\frac{1}{2}(\beta - \tilde{\beta})'(\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1})(\beta - \tilde{\beta})\right)} \\ &= |\Sigma^0|^{-1/2} |\mathbf{X}'\mathbf{X} + (\Sigma^0)^{-1}|^{-1/2} \exp\left(\frac{\tilde{\delta}}{2}\right) \end{aligned} \tag{49}$$

Here, we exploited the fact $p(\mathbf{y}|\beta^0, \Sigma^0)p(\beta|\mathbf{y}, \beta^0, \Sigma^0) = p(\mathbf{y}|\beta)p(\beta|\beta^0, \Sigma^0)$, by elementary rules of probability. Also note that with the intent to eventually compare across models defined by different likelihoods and priors, we kept track of all normalizing constants that we conveniently ignored before, when deriving the posterior distribution in Eq. 16. Specifically, we previously ignored the factors

$1/\sqrt{2\pi}$ and $(2\pi)^{-k/2}|\Sigma^0|^{-1/2}$ in the likelihood $p(\mathbf{y}|\boldsymbol{\beta})$ and the prior $p(\boldsymbol{\beta}|\boldsymbol{\beta}^0, \Sigma^0)$, respectively.

Recall that \tilde{s} in the last line of Eq. 49 is a deterministic function of the subjective prior parameters $\boldsymbol{\beta}^0, \Sigma^0$, and the data \mathbf{y} (see Eq. 14). For all nondegenerate prior choices, \tilde{s} is going to be dominated by the term $(\mathbf{y} - \mathbf{X}'\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}'\tilde{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\beta}}$ converges to the maximum likelihood or ordinary least squares estimate as more data become available (assuming regular \mathbf{X}).

Now consider the comparison between two models. Model \mathcal{M}_0 happens to employ the p -column \mathbf{X} -matrix that collects all covariates that systematically influenced \mathbf{y} , when the data was generated – the true model. Model \mathcal{M}_1 uses a model matrix that features the same p covariates in \mathbf{X} plus s additional covariates in \mathbf{X}^s that did not contribute to the variation in \mathbf{y} , when the data was generated. The Bayes' factor $BF_{0,1}$ is then:

$$BF_{0,1} = \frac{p(\mathbf{y}|\mathcal{M}_0)}{p(\mathbf{y}|\mathcal{M}_1)} = \frac{|\Sigma_0^0|^{-1/2}|\mathbf{X}'\mathbf{X} + (\Sigma_0^0)^{-1}|^{-1/2} \exp\left(\frac{-\tilde{s}_0}{2}\right)}{|\Sigma_1^0|^{-1/2}|(\mathbf{X}, \mathbf{X}^s)'(\mathbf{X}, \mathbf{X}^s) + (\Sigma_1^0)^{-1}|^{-1/2} \exp\left(\frac{-\tilde{s}_1}{2}\right)} \tag{50}$$

where Σ_0^0 and Σ_1^0 are of dimension $p \times p$ and $(p + s) \times (p + s)$, respectively. In the limit of more and more data, \tilde{s}_0 and \tilde{s}_1 will converge to the same value, as the data determine that the elements in $\tilde{\boldsymbol{\beta}}_1$ that correspond to \mathbf{X}^s are equal to zero. Then, the limit of the ratio in Eq. 50 only depends on:

$$\begin{aligned} & (|(\mathbf{X}, \mathbf{X}^s)'(\mathbf{X}, \mathbf{X}^s) \parallel \mathbf{X}'\mathbf{X}|^{-1})^{1/2} \\ & = (n^{p+s}|n^{-1}(\mathbf{X}, \mathbf{X}^s)'(\mathbf{X}, \mathbf{X}^s) \parallel n^{-1}\mathbf{X}'\mathbf{X}|^{-1}n^{-p})^{1/2} \\ & \approx n^{s/2} \end{aligned}$$

which is easily seen to converge to infinity in the limit of more and more data (larger n), for regular $(\mathbf{X}, \mathbf{X}^s)$ (The expressions $n^{-1}(\mathbf{X}, \mathbf{X}^s)'(\mathbf{X}, \mathbf{X}^s)$ and $n^{-1}\mathbf{X}'\mathbf{X}$ define covariance matrices that will converge to fixed matrices in the sample size n for covariates with finite variance). Thus, the Bayes' factor can in fact produce infinitely strong evidence for the more parsimonious model, if it is the data generating mechanism.

If in contrast \mathcal{M}_1 were the true model, or just closer to the truth in this case, the coefficients in $\tilde{\boldsymbol{\beta}}_1$ that correspond to \mathbf{X}^s do not converge to zero. As a consequence, \tilde{s}_0 would grow faster in n than \tilde{s}_1 , and $SF_{0,1}$ would converge to zero (Note that $\exp\left(\frac{-\tilde{s}_0 + \tilde{s}_1}{2}\right) = \exp\left(n\frac{-\tilde{s}_0/n + \tilde{s}_1/n}{2}\right)$ converges to zero faster than $n^{s/2}$ grows because of the exponential function, where $-\tilde{s}_0/n + \tilde{s}_1/n$ converges to the true difference in average squared errors between \mathcal{M}_1 and \mathcal{M}_0). Thus, the Bayes' factor can both produce increasing evidence for the more parsimonious model, when the constraints imposed by this model hold exactly, and increasing evidence against it, when they do

not (consider $BF_{1,0}$ instead of $BF_{0,1}$ in this case) (In this case, the conventional classifications of weak and strong evidence in favor of the model in the numerator of the Bayes' factor often align with the usual cut-off values for rejecting a more constrained model based on p-values). In contrast, p-values can reliably reject a parsimonious model but are incapable of producing increasing evidence for such a model. By construction, the probability of rejecting a true, more parsimonious model in favor of a larger, over-parameterized model is equal to the chosen significance level data in repeated applications of the frequentist testing procedure, and independent of the sample size (the amount of information in the data).

Numerical Illustrations

A Brief Note on Software Implementation

Researcher interested in adopting the Bayesian approach nowadays have quite some choice regarding different software and available implementations of the Bayesian approach. More recently, established products for data analysis such as *SPSS*, *STATA*, or *SAS* have started to include options for Bayesian estimation of well established “standard” statistical models such as ANOVA and generalized linear regression models (Advanced users can certainly use these tools to estimate “their own” models too, and *STATA* specifically emphasizes this possibility). In contrast, *WINBUGS* is an example of an attempt to automate Bayesian inference, with the idea that the user should be able to exclusively concentrate on the specification of a model – likely outside of the set of “standard” statistical models implemented elsewhere – aided by a graphical user interface.

Much if not the vast majority of “Bayesian-papers” published in marketing to this day have relied on coding up the model and the (invariably) MCMC-routine to perform Bayesian inference “from scratch,” starting with some example code and taking advantage of components that repeat themselves across different models, e.g., conditionally conjugate updating of parameters indexing hierarchical priors. The programming languages used in this context include compiled languages such as C or Fortran, and interpreted languages such as Matlab, R, and Gauss. Here, the former are by construction less interactive when coding and the latter slower in the execution of code “that works.” Recently, Rcpp (Eddelbuettel and François 2011; Eddelbuettel 2013) emerged as an extremely useful compromise between the speed of compiled and the coder-friendliness of interpreted languages.

I am currently relying heavily on Rcpp in my own research. However, I view the advent of the No U-turn Sampler (NUTS) by Hoffman and Gelman (2014) as implemented in Stan (Carpenter et al. 2017) as a major breakthrough towards the goal of focusing on the specification of innovative models (almost) exclusively.

A Hierarchical Bayesian Multinomial Logit Model

At least in marketing, no treatment of Bayesian modeling would be complete without illustrating the benefits from a hierarchical Bayesian model in the context of large N , small T data. I consider the stylized case of multinomial logit choice from choice sets with two inside alternatives, say brands A and B , and an outside option with expected utility normalized to zero. The utility of the two inside alternatives stems, in addition to alternative specific constants, from a uniformly distributed covariate x , i.e., $U_{Ait} = \beta_{Ai} + \beta_F x_{Ait} + \varepsilon_{Ait}$ and $U_{Bit} = \beta_{Bi} + \beta_F x_{Bit} + \varepsilon_{Bit}$. Here, $i = 1, \dots, N$ indexes heterogeneous individuals and $t = 1, \dots, T$ choice occasions. Population preferences are distributed according to:

$$\beta_i = \begin{pmatrix} \beta_{Ai} \\ \beta_{Bi} \\ \beta \end{pmatrix} \sim \mathcal{N} \left(\begin{bmatrix} .3 \\ -2 \\ -1 \end{bmatrix}, \begin{bmatrix} 3 & -2.99 & 0 \\ -2.99 & 3 & 0 \\ 0 & 0 & .1 \end{bmatrix} \right)$$

Thus, brand A is slightly preferred to the outside good on average, whereas brand B is less attractive than the outside good to the average consumer in this market. However, there is a fair amount of heterogeneity in brand preferences in this market. For example, about 12.4% of consumers in this market prefer brand B to the outside good and around 43% prefer the outside good to brand A at $x = 0$. Moreover, consumers that have an above average preference for brand A are likely to have a below average preference for brand B in this market, as per the strongly negatively correlated brand coefficients in the population ($\rho = -.997$). The tastes for the covariate x are relatively more homogenous and only consumers in the extreme tail of the preference distribution exhibit a higher preference for larger values of x in this population. I simulate $N=2,000$ individuals from this population and have each individual make $T = 5$ choices from complete sets that randomly vary in the x -values for brands A and B , both across $t = 1, \dots, T$ and $i = 1, \dots, N$. I use this data to calibrate a Bayesian hierarchical MNL-model. I rely on the default subjective prior distributions implemented in `bayesm`'s estimation routine `rhierMnIRwMixture` and run this RW-MH-sampler with automatic tuning of proposal densities for 100,000 iterations saving every 10th draw (in `bayesm`: `R = 100,000`, `keep = 10`). The complete posterior is a 6009-dimensional object (3 means plus 3 variances plus 2 covariances plus 2000 times 3 individual level random effects). Because of the high dimensionality of the posterior, saving every draw from a long MCMC run can easily produce an object that taxes a computer's RAM heavily. Saving every `keep`-th draw increases the information content in a posterior sample limited by a computer's RAM. For a maximum number of draws than can be saved, we can increase the number of MCMC-iterations `R`, when we simultaneously increase the number of iterations between parameters to be saved (`keep - 1`). The information content in the resulting sample is increased because saved draws separated by `keep - 1` MCMC iterations will tend to be more independent from each other, replicate less of the information contained the preceding draw saved.

Figure 8 exhibits individual level posteriors for individuals 3, 99, and 2000 in our simulated panel data. For this purpose, I use the last 9000 draws of the 10,000 draws I saved. The three rows in Fig. 8 correspond to β_A , β_B , and β , respectively. Each

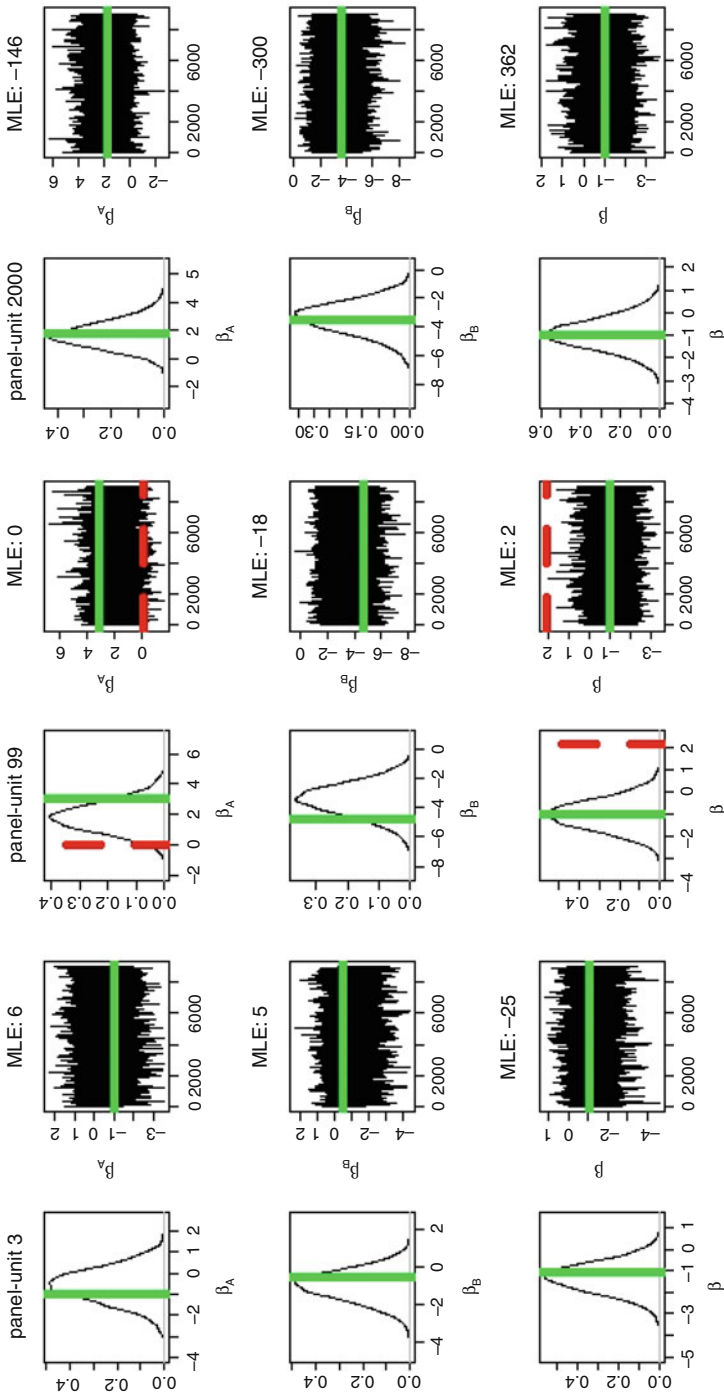


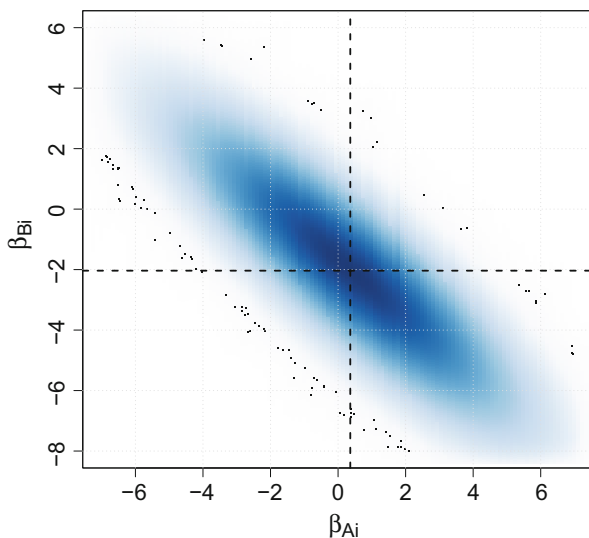
Fig. 8 Individual level posterior distributions

individual's posterior is depicted in two different ways in two adjacent columns of Fig. 8, each. The first column summarizes marginal posterior densities using density plots obtained by Gaussian-kernel-smoothing. The second column shows the MCMC-trace plot of draws underlying the density plots in the respective first column. Green solid bars indicate data generating parameter values. Red-dashed bars indicate individual level maximum likelihood estimates obtained from numerical maximization of the likelihood function using the R-function `optim` (I report numerical estimates regardless of the existence of a finite maximum likelihood estimate given the 5 choices from a specific individual and the corresponding design matrices). If no red-dashed bar is showing, this indicates that the maximum likelihood estimate falls outside of the range of parameter values plotted. To still give an impression of maximum likelihood estimates, the MCMC-trace plots in the respective second columns have the maximum likelihood estimates in the title.

Looking at the maximum likelihood estimates and comparing them to the green bars, we can see that they are extremely inaccurate. Clearly, individual level posterior inference benefits tremendously from the information in the hierarchical prior distribution that the model learns by pooling information across the 2000 consumers in our simulated short panel.

Finally, Fig. 9 illustrates how the hierarchical Bayesian MNL-model recovers the joint distribution of preferences for brands A and B in the population of consumers. We recognize the strongly negative relationship between preferences for brands A and B in the population (However, the posterior mean correlation of -0.88 (0.037) overestimates the data generating correlation of -0.997 , which can be traced back to the finite information in the data available for calibration and the subjective priors for population level parameters employed here. See the documentation of `rhierMnlRwMixture` for details). Thus, if a particular individual level likelihood is only informative about the preference for brand A (B), the corresponding preference

Fig. 9 Joint posterior distribution of $\{\beta_{Ai}\}$ and $\{\beta_{Bi}\}$



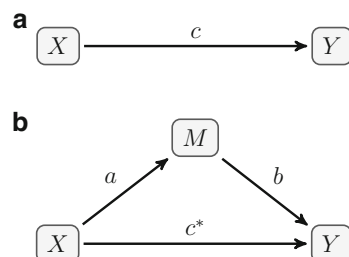
for brand B (A) can be inferred rather accurately from the hierarchical prior distribution. Dashed lines in Fig. 9 indicate posterior population means, that nicely recover the data generating values from the information in all $N \times T$ choice observations. The code to replicate this illustration is again available in the [Appendix](#).

Note that we estimated the model that was used to generate the data here. In applications, it is very likely that some or all subjective choices that go into the formulation of the model result in systematic differences from the data generating mechanism, including the choice of the hierarchical prior distribution that was (implicitly) chosen to be multivariate normal in this illustration. However, it is also clear that even misspecified hierarchical prior distributions can strike a beneficial bias-variance trade-off in applications where individual level maximum likelihood estimates are extremely noisy or may not exist at all. In fact, this bias-variance trade-off is at the source of the inroads Bayesian hierarchical models have made into applications in marketing. For a discussion of how to imbue hierarchical prior distributions with subjective knowledge about ordinal relationships, see Pachali et al. (2018).

Mediation Analysis: A Case for Bayesian Model Comparisons

In this section, we borrow from Otter et al. (2018). Mediation analysis has developed in psychology, as a tool to empirically establish the process by which an experimental manipulation brings about its effect on the dependent variable of interest. An important distinction in this context is that between full and partial mediation at a causal theory level. I will not discuss the related model specification questions here but focus on the fact that if an experimentally manipulated cause X and a measured consequence Y become independent when conditioned on a measured mediator M , evidence for (full) mediation is established (This is because conditional independence would only result in very particular essentially zero probability circumstances from models where full mediation is not the causal mechanism at work. Results that do not establish some form of conditional independence, which are often interpreted as “partial mediation,” actually are ambiguous with regarding their interpretation (Otter et al. 2018)). The original test for mediation proposed by Baron and Kenny (1986) builds on the connection between full mediation and conditional independence and tests conditional mean independence. Their test rests on the following set of regression equations, where t 's denote intercepts (see also Fig. 10).

Fig. 10 Mediation according to Baron and Kenny (Baron and Kenny 1986)



$$Y_i = t_1 + cX_i + \varepsilon_{Y,i} \quad (51)$$

$$M_i = t_2 + aX_i + \varepsilon_{M,i} \quad (52)$$

$$Y_i = t_3 + c^*X_i + bM_i + \varepsilon_{Y,i}^* \quad (53)$$

The first equation regresses Y on the randomly assigned experimental variable X . A statistically significant coefficient c establishes empirical support for the total effect from X to Y (see Fig. 10, Panel a). Because of random assignment of X , the coefficient c necessarily measures a causal effect. The second equation regresses M on X . A statistically significant coefficient a establishes empirical support for the effect from X to M that is again causal by experimental design. The third equation regresses Y on randomly assigned X and on observed M . Finding that the effect from X on Y vanishes, when conditioned on M (i.e., that there is no direct effect c^*), unequivocally establishes (full) mediation as the causal data generating model (see Fig. 10, Panel b) (In the limit of an infinite amount of data, the estimate of c^* will only converge to exactly zero under full mediation. The only alternative process that yields $c^* = 0$ in the limit features M as a joint cause of X and Y without another connection between X and Y . This process is ruled out a priori, when X is experimentally manipulated). Usually, empirical support for the hypothesis of $c^* = 0$ is established based on p-values larger than some subjectively chosen significance level. An obvious drawback of this approach is that p-values, by construction, fail to measure the strength of empirical support for conditional independence, which in turn establishes full mediation. Based on p-values, we can only “fail to reject” the null-hypothesis.

Next, I illustrate the differences between the classical and the Bayesian approach in the context of $c^* = 0$ using a sampling experiment. I thus consider the case of full mediation as DGP. Accordingly, I set $t_2 = t_3 = 1$, $a = 4$, $c^* = 0$, $b = 0.5$, and $\sigma_M = \sigma_{Y^*} = 1$ in Eqs. 52 and 53 and generate artificial data sets of different sizes: $N_1 = 50$, $N_2 = 200$, as well as $N_3 = 2000$ (X_i is drawn from a uniform distribution for each $i \in \{1, \dots, N\}$). I conduct 1000 replications for each data set size and compute Bayes’ factors defined as ratios of marginal likelihoods of the model $\mathcal{M}_0 : Y_i = t_3 + bM_i + \varepsilon_{Y,i}^*$ and the model $\mathcal{M}_1 : Y_i = t_3 + c^*X_i + bM_i + \varepsilon_{Y,i}^*$. Note that the former is more restricted than the latter and implies that the coefficient c^* is equal to zero in the latter model (see Otter et al. (2018) for the computational details and R-scripts).

Table 8 illustrates the distribution of estimated Bayes Factors over the 1000 simulation replications testing the hypothesis of $c^* = 0$. The results in Table 8 verify that the Bayes Factor correctly favors \mathcal{M}_0 over \mathcal{M}_1 for the vast majority of sampling replications. Importantly, this table also illustrates that the Bayes Factor provides increasingly stronger evidence for \mathcal{M}_0 (i.e., $c^* = 0$) as the sample size increases.

The classical testing framework based on p-values fails to measure the strength of evidence in favor of $c^* = 0$. In line with how they are defined, p-values are uniformly distributed over sampling replications in the interval of (0,1) (see Table 9). The probability of observing a p-value smaller than the specified significance level is

Table 8 Distribution of Bayes' factors in simulation

	Pr(BF) > 3	Pr(BF) > 20	Pr(BF) > 100
$N = 50$	0.94	0.04	0.00
$N = 200$	0.97	0.71	0.00
$N = 2000$	0.99	0.93	0.43

Table 9 Distribution of p-values in simulation

	Pr(p-value) > 0.01	Pr(p-value) > 0.05	Pr(p-value) > 0.10
$N = 50$	0.99	0.96	0.90
$N = 200$	0.99	0.96	0.90
$N = 2000$	0.99	0.96	0.90

equal to this level, and independent of the sample size, when the null-hypothesis is actually true. In contrast, the probability of obtaining a Bayes Factor larger than 20 in support of $c^* = 0$ increases in the sample size and, for example, approaches one for $N = 2000$ (see Table 8).

Thus, when the data generating process implies conditional independence ($c^* = 0$), the Bayesian approach is the superior measure of empirical evidence for this process, compared to the approach based on p-values.

Conclusion

Writing a chapter like this, one certainly involves many trade-offs. I have chosen to emphasize general principles of Bayesian decision-making and inference in the hope of interesting and exciting readers that have an inclination towards quantitative methodology and are serious about improving marketing decisions. The promise from a deeper appreciation of the Bayesian paradigm, both in terms of its foundations in (optimal) decision-making and in terms of its computational approaches are better tailored quantitative approaches that can be developed and implemented as required by a new decision problem, or for the purpose of extracting (additional) knowledge from a new data source.

A drawback of this orientation is that the plethora of existing models that are usefully implemented in a fully Bayesian estimation framework, including common place prior distributions, are not even enumerated in this chapter. However, I believe that the full appreciation of individual, concrete applications requires a more general understanding of the Bayesian paradigm. Once this understanding develops, that for different individual models follows naturally.

Cross-References

- ▶ [Finite Mixture Models](#)
- ▶ [Fusion Modeling](#)

- ▶ [Multilevel Modeling](#)
- ▶ [Panel Data Analysis: A Non-technical Introduction for Marketing Researchers](#)

Acknowledgments I would like to thank Anocha Aribarg, Albert Bemmaor, Joachim Büschken, Arash Laghaie, anonymous reviewers, the editors, and participants in my class on “Bayesian Modeling for Marketing” helpful comments and feedback. All remaining errors are obviously mine.

Appendix

MCMC for Binomial Probit Without Data Augmentation

Simulate data, call MCMC routine, plot MCMC-traces. This R-script sources a RW-MH-sampler for the binomial probit model (see the following script), simulates probit data, and runs the code with different step-sizes (standard deviations of ϵ).

```
# may need to install these packages first
library ( bayesm )
library ( latex2exp )

# needs to be in R's working directory
source ( ' rbprobitRWMetropolis .r' )

# function to simulate from binary probit
simbprobit = function ( X, beta ) {
y= ifelse ( (X*% beta + rnorm ( nrow (X))) < 0 , 0 , 1)
list ( X=X,y=y, beta = beta )
}

nobs =500 # number of simulated observations
X= cbind ( rep ( 1, nobs ), runif ( nobs ), runif ( nobs ))
beta =c( -3 , 2 , 4) # data generating parameters
nvar = ncol ( X)
simout = simbprobit ( X, beta )

# probit responses
y= simout $y

R =200000 # length of MCMC sample

# data list to passed to MCMC routine
Data = list ( X= simout $X,y= simout $y)

Mcmc = list ( R=R, keep =1)

# prior mean set to zero, prior variances set to 100
Prior = list ( betabar = double ( nvar ), A= diag ( rep (.01, nvar )))

out _1= rbprobitRWMetropolis ( Data =Data, Mcmc =Mcmc,
```

```

Prior =Prior, stepsize =.001)
out _2= rbprobitRWMetropolis ( Data =Data, Mcmc =Mcmc,
Prior =Prior, stepsize =.005)
out _3= rbprobitRWMetropolis ( Data =Data, Mcmc =Mcmc,
Prior =Prior, stepsize =.8)
out _4= rbprobitRWMetropolis ( Data =Data, Mcmc =Mcmc,
Prior =Prior, stepsize =3)

windows ()
par ( mfrow =c (2,2))
matplot ( out _1$ betadraw, type ='l',xlab = "", ylab = "",
main = TeX ('$\\cr epsilon $-standard\\deviation\\u=\\u.001 ')); grid ()
matplot ( out _2$ betadraw, type ='l',xlab = "", ylab = "",
main = TeX ('$\\cr epsilon $-standard\\deviation\\u=\\u.005 ')); grid ()
matplot ( out _3$ betadraw, type ='l',xlab = "", ylab = "",
main = TeX ('$\\cr epsilon $-standard\\deviation\\u=\\u.8 ')); grid ()
matplot ( out _4$ betadraw, type ='l',xlab = "", ylab = "",
main = TeX ('$\\cr epsilon $-standard\\deviation\\u=\\u3')); grid ()

```

MCMC function. The following function implements a simple RW-MH-sampler for the binomial probit model coupled with a multivariate normal prior. All regression parameters are updated simultaneously in one MH-step.

```

rbprobitRWMetropolis <- function (Data, Prior, Mcmc, stepsize )
{
require ( bayesm )
# because of the use of lndMnv to evaluate the log - density of a ...
# ... multivariate normal distribution

y = Data $y
nvar = ncol (X)
nobs = length (y)
betabar = Prior $ betabar
A = Prior $A
R = Mcmc $R
keep = Mcmc $ keep

betadraw = matrix ( double ( floor (R/ keep ) * nvar ), ncol = nvar )
loglike = double ( floor (R/ keep ) )
beta = c( rep (0, nvar ) )

priorcov = chol2inv ( chol (A))
rootp = chol ( priorcov )
rootpi = backsolve (rootp, diag ( nvar ))

# intialize log - likelihood at starting value
oldloglike =

```



```

sum ( pnorm (0, (X%% beta ) [ as. logical (y) ], 1, log .p= TRUE ))+
sum ( pnorm (0, (-X%% beta ) [!as. logical (y) ], 1, log .p= TRUE ))

# compute non - normalized log - posterior at starting value
oldlpost = oldloglike + lndMvn (beta, betabar, rootpi)

naccept = 0
for ( rep in 1:R) {
betac = beta + rnorm ( nvar ) * stepsize # random walk proposal

# compute probit log - likelihood at proposed value
cloglike =
sum ( pnorm (0, -(X%% betac ) [ as. logical (y) ], 1, log .p= TRUE ))+
sum ( pnorm (0, (X%% betac ) [!as. logical (y) ], 1, log .p= TRUE ))

# compute non - normalized log - posterior at proposed value
clpost = cloglike + lndMvn (betac, betabar, rootpi )

# compute log - ratio of non - normalized posterior at proposed ...
# ... and old value
ldiff = clpost - oldlpost
alpha = min (1, exp ( ldiff )) # acceptance probability
if ( alpha < 1) {
unif = runif (1)
}
else {
unif = 0
}
if ( unif <= alpha ) {
beta = betac
oldloglike = cloglike
oldlpost = clpost
naccept = naccept + 1
}

if ( rep %% keep == 0) {
mkeep = rep / keep
betadraw [mkeep, ] = beta
loglike [ mkeep ] = oldloglike
}
}

# betadraw is the matrix containing draws from the posterior
# rateaccept is the relative frequency of accpeting proposed moves ...
# ... from oldbeta to betac
# loglike is the log - likelihood ...
# ... evaluated at the current MCMC state ( beta )
return ( list ( betadraw = betadraw, mkeep =mkeep,

```

```
rateaccept = naccept /R, loglike = loglike ))
}
```

Stan probit definition file. This file that is called as `StanProbit.stan` by the R-script immediately below defines a binomial probit model with a multivariate normal prior for Stan. According to the model, the data are independently Bernoulli distributed with probabilities implied by the probit-link, parameters, and covariates.

```
data {
int N; // number of observations
int K; // number of covariates
int < lower =0, upper =1> y[N]; // information
matrix [N,K] X; // design matrix
}
parameters {
vector [K] beta ; // beta coefficients
}
model {
vector [N] mu;
beta ~ normal (0, 100);
mu = X* beta ;
for (n in 1:N) mu[n] = Phi (mu[n ]);
y ~ bernoulli (mu );
}
```

Calling Stan from R to estimate a binomial probit model. This R-script calls **Stan** to sample from the posterior of the binomial probit model coupled with a multivariate normal prior defined in the file above.

```
# may need to install the rstan package first
require ( rstan ) # load the rstan package
# see sripts above for nobs, nvar, simout objects
prob _ data = list (N=nobs ,K=nvar ,X= simout $X,y=as. vector (
simout $y))
rstan _ options ( auto _ write = TRUE )
options (mc. cores = parallel :: detectCores ())
stanfit _ probit = stan ( file =" StanProbit . stan ",data = prob _
data,
pars = c(" beta "), chains = 1,
iter = 600000, warmup = 1000)

# Make draws available for posterior analysis in R
out _ StanProbit = extract ( stanfit _ probit )
```

HB-Logit Example

This code generates MNL-data from a hierarchical model, estimates an HB-logit model, and compares selected individual level posteriors to the corresponding maximum likelihood estimates.

```

genXy <- function (betai ,p,T){
## generate multinomial logit choices
# alternative specific constants
# ... this assumes p=3 ( two inside brands, one outside choice )
X= kronecker ( rep (1,T), matrix (c(1 ,0 ,0 ,0 ,1 ,0) , ncol =(
length ( betai ) -1)))
# add the continuous covariate
X= cbind (X, runif (T*p))
index = seq (p,p*T,p)
X[index ,]=0 # outside good
Xbeta =t( matrix (X%*%betai , nrow =p))
index = cbind (1:T, max . col ( Xbeta ))
maxl = Xbeta [ index ]
logsumel = log ( rowSums ( exp (Xbeta - maxl ))) + maxl
logprob = matrix (Xbeta - logsumel , nrow =T)
y= double (T)
for (t in 1:T){
y[t]= sum ( cumsum ( exp ( logprob [t ,])) < runif (1))+1 ## draw
from the CDF of probs
}
return ( list (y=y,X=X))
}

p=3 # number of alterantives in each choice set
T=5 # number of repeated measurements, i.e., choice sets or choices

# generate panel data for MCMC analysis
N =2000 # number of individuals in the panel

# population mean preference
betap =c(.3 , -2 , -1)

# variance - covariance of preferences in the population
Vbeta = matrix (c(3 , -2.99 ,0 , -2.99 ,3 ,0 ,0 ,0 ,.1) , ncol =3)
# just for demonstration to make sure we all get ...
# ... the same result date and results
set . seed (66)
# draw individual specific preferences from MVNormal distribution
betai = betap +t( chol ( Vbeta ))%*% matrix ( rnorm (N* length (
betap )), ncol =N)
lgtdata <- vector (" list ", N)

```

```

T=5 # number of choices per individual

betaMLE = betai
betaMLE [ , ]=0

for (i in 1:N){
outgen = genXy ( betai [,i],p,T)
# For Bayesian analysis using rhierMnlRwMixture ...
# ... you need to organize your data in list format as ...
# ... in the command line below
# y :: vector of choice outcomes of length T or ...
# ... T_i in case different panel units provide different numbers
of choices
# X :: A (p*T) rows x length ( beta [,i]) columns model matrix ;
# the first ( second ) p rows correspond to the first ( second )
choice set, and so on.
# Each alternative is represented by one row in X.
# The numbers in y point to which 'row ' was chosen from a
particular choice set
lgtdata [[i ]]= list (y= outgen [[1]], X= outgen [[2]])
out = optim ( par = betai [,i], fn=l1MNL, gr=NULL, y= outgen [[1]],
X= outgen [[2]], p=p, hessian = FALSE, control = list ( fnscale =
-1))
betaMLE [,i]= out $ par # collect MLE estimates
}

# load the bayesm package into the workspace
# (if this gives you an error, ...
# ... you need to install the package first )
library ( bayesm )
# run the Bayesian hierarchical model
outMCMC = rhierMnlRwMixture ( Data = list (p=p, lgtdata = lgtdata
),
Prior = list ( ncomp =1), Mcmc = list (R =100000, keep =10))

# posterior of individual specific coefficients
betaimc = outMCMC $ betadraw

index =1001:10000

# may need to install this first
library ( latex2exp )

M=c ( 3 ,99 ,2000) # plot betai posterior for consumers in M
jpeg ( filename = " ILposteriors880 . jpg ", quality = 100 , width =
880 , height = 480)
# windows ()
par ( mfcol =c( length ( betap ), length (M)* 2))

```

```

for (i in M){
plot ( density ( betaimc [i ,1, index ]),
xlab = TeX ('$\cr beta_{A}$'), ylab = "\u", main = paste ("panel -
unit\u", i))
abline (v= betai [1,i], col ='green ', lwd =5, lty =1 )
abline (v= betaMLE [1,i], col ='red ', lwd =5, lty =2 )
plot ( density ( betaimc [i ,2, index ]),
xlab = TeX ('$\cr beta_{B}$'), ylab = "\u", main ="\u")
abline (v= betai [2,i], col ='green ', lwd =5, lty =1 )
abline (v= betaMLE [2,i], col ='red ', lwd =5, lty =2 )
plot ( density ( betaimc [i ,3, index ]),
xlab = TeX ('$\cr beta $'), ylab = "\u", main ="\u")
abline (v= betai [3,i], col ='green ', lwd =5, lty =1 )
abline (v= betaMLE [3,i], col ='red ', lwd =5, lty =2 )

plot ( betaimc [i ,1, index ], type ='l',
xlab ="\u", ylab = TeX ('$\cr beta_{A}$'), main = paste (" MLE :
\u", round ( betaMLE [1,i ])))
abline (h= betai [1,i], col ='green ', lwd =5, lty =1 )
abline (h= betaMLE [1,i], col ='red ', lwd =5, lty =2 )
plot ( betaimc [i ,2, index ], type ='l',
xlab ="\u", ylab = TeX ('$\cr beta_{B}$'), main = paste (" MLE :
\u", round ( betaMLE [2,i ])))
abline (h= betai [2,i], col ='green ', lwd =5, lty =1 )
abline (h= betaMLE [2,i], col ='red ', lwd =5, lty =2 )
plot ( betaimc [i ,3, index ], type ='l',
xlab ="\u", ylab = TeX ('$\cr beta $'), main = paste (" MLE : \u",
round ( betaMLE [3,i ])))
abline (h= betai [3,i], col ='green ', lwd =5, lty =1 )
abline (h= betaMLE [3,i], col ='red ', lwd =5, lty =2 )
}

```

References

- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679. <http://www.jstor.org/stable/2290350>
- Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32(2), 152–162. <http://www.jstor.org/stable/3152044>
- Allenby, G. M., Arora, N., & Ginter, J. L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, 35(3), 384–389. <http://www.jstor.org/stable/3152035>
- Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182. <https://doi.org/10.1037/0022-3514.51.6.1173>.

- Bernardo, J. M., & Smith, A. F. M. (2001). Bayesian theory. *Measurement Science and Technology*, 12(2), 221. <http://stacks.iop.org/0957-0233/12/i=2/a=702>.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 192–236. <http://www.jstor.org/stable/2984812>
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>. <https://www.jstatsoft.org/v076/i01>.
- Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. New York: Springer. <http://gateway.library.qut.edu.au/login?url=http://link.springer.com/openurl?genre=book&isbn=978-1-4612-1276-8>.
- Chib, S., & Carlin, B. P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing*, 9(1), 17–26. <https://doi.org/10.1023/A:1008853808677>.
- Eddelbuettel, D. (2013). *Seamless R and C++ integration with Repp*. New York: Springer.
- Eddelbuettel, D., & François, R. (2011). Repp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>. <http://www.jstatsoft.org/v40/i08/>
- Edwards, Y. D., & Allenby, G. M. (2003). Multivariate analysis of multiple response data. *Journal of Marketing Research*, 40(3), 321–334. <https://doi.org/10.1509/jmkr.40.3.321.19233>.
- Fasiolo, M. (2016). *An introduction to mvnfast. R package version 0.1.6*. <https://CRAN.R-project.org/package=mvnfast>
- Frühwirth-Schnatter, S., Tüchler, R., & Otter, T. (2004). Bayesian analysis of the heterogeneity model. *Journal of Business & Economic Statistics*, 22(1), 2–15. <https://doi.org/10.1198/073500103288619331>.
- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2018). *mvtnorm: Multivariate normal and t distributions*. <https://CRAN.R-project.org/package=mvtnorm>. R package version 1.0-8.
- Geweke, John. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In: E. M. Keramidas (Ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 571–578.
- Gilks, W. R. (1996). Full conditional distributions. In S. (Sylvia) Richardson, D. J Spiegelhalter, & W. R. (Walter R.) Gilks (Eds.), *Markov chain Monte Carlo in practice* (pp. 75–88). London/Melbourne: Chapman & Hall.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>.
- Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15, 1593–1623. <http://jmlr.org/papers/v15/hoffman14a.html>.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>.
- Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1), 93–119. <https://doi.org/10.1007/BF02294188>.
- Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2), 173–191. <https://doi.org/10.1287/mksc.15.2.173>.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage Publications. <https://uk.sagepub.com/en-gb/eur/regression-models-for-categorical-and-limited-dependent-variables/book6071>.
- McCulloch, R., & Rossi, P. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1–2), 207–240. <https://EconPapers.repec.org/RePEc:eee:econom:v:64:y:1994:i:1-2:p:207-240>.

- Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). *truncnorm: Truncated normal distribution*. <https://CRAN.R-project.org/package=truncnorm>. R package version 1.0-8
- Montgomery, A. L., & Bradlow, E. T. (1999). Why analyst overconfidence about the functional form of demand models can lead to overpricing. *Marketing Science*, 18(4), 569–583. <http://www.jstor.org/stable/193243>
- Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (Chap. 5). Chapman & Hall/CRC. <http://arxiv.org/abs/1206.1901>
- Orme, B. (2017). The CBC system for choice-based conjoint analysis. Technical Report. <https://sawtoothsoftware.com/download/techpap/cbctech.pdf>
- Otter, T., Tüchler, R., & Frühwirth-Schnatter, S. (2004). Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. *International Journal of Research in Marketing*, 21(3), 285–297. <https://doi.org/10.1016/j.ijresmar.2003.11.002>. <http://www.sciencedirect.com/science/article/pii/S0167811604000308>
- Otter, T., Gilbride, T. J., & Allenby, G. M. (2011). Testing models of strategic behavior characterized by conditional likelihoods. *Marketing Science*, 30(4), 686–701. <http://www.jstor.org/stable/23012019>
- Otter, T., Pachali, M. J., Mayer, S., & Landwehr, J. R. (2018). Causal inference using mediation analysis or instrumental variables – Full mediation in the absence of conditional independence. *Marketing ZFP*, 40(2), 41–57. <https://doi.org/10.15358/0344-1369-2018-2-41>.
- Pachali, M. J., Kurz, P., & Otter, T. (2018). How to generalize from a hierarchical model? Technical Report. <https://ssrn.com/abstract=3018670>
- Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). New York: Cambridge University Press.
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1), 7–11. <https://journal.r-project.org/archive/>.
- Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419), 861–868. <https://doi.org/10.1080/01621459.1992.10475289>.
- Robert, C. P. (1994). *The Bayesian choice: a decision-theoretic motivation*. New York: Springer.
- Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In S. (Sylvia) Richardson, D. J. Spiegelhalter, & W. R. (Walter R.) Gilks (Eds.), *Markov chain Monte Carlo in practice* (pp. 45–58). London/Melbourne: Chapman & Hall.
- Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, 15(4), 321–340. <https://doi.org/10.1287/mksc.15.4.321>.
- Rossi, P. E., Allenby, G. M., & McCulloch, R. E. (2005). *Bayesian statistics and marketing*. Chichester: Wiley.
- Wachtel, S., & Otter, T. (2013). Successive sample selection and its relevance for management decisions. *Marketing Science*, 32(1), 170–185. <https://doi.org/10.1287/mksc.1120.0754>.
- Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley.



Choice-Based Conjoint Analysis

Felix Eggers, Henrik Sattler, Thorsten Teichert, and Franziska Völckner

Contents

Introduction	782
Model	787
Utility Model	787
Choice Model	791
Procedure for Conducting Discrete Choice Experiments	791
Identification of Attributes and Attribute Levels	791
Creating the Experimental Design	793
Implementation into Questionnaire	796
Estimation	799
Advanced Estimation Techniques	808
Outlook	813
Appendix: R Code	813
References	816

Abstract

Conjoint analysis is one of the most popular methods to measure preferences of individuals or groups. It determines, for instance, the degree how much consumers like or value specific products, which then leads to a purchase decision. In particular, the method discovers the utilities that (product) attributes add to the

F. Eggers
University of Groningen, Groningen, The Netherlands
e-mail: f.egggers@rug.nl

H. Sattler · T. Teichert
University of Hamburg, Hamburg, Germany
e-mail: henrik.sattler@uni-hamburg.de; thorsten.teichert@uni-hamburg.de

F. Völckner (✉)
Department of Marketing and Brand Management, University of Cologne, Köln, Germany
e-mail: voelckner@wiso.uni-koeln.de

overall utility of a product (or stimuli). Conjoint analysis has emerged from the traditional rating- or ranking-based method in marketing to a general experimental method to study individual's discrete choice behavior with the choice-based conjoint variant. It is therefore not limited to classical applications in marketing, such as new product development, pricing, branding, or market simulations, but can be applied to study research questions from related disciplines, for instance, how marketing managers choose their ad campaign, how managers select internationalization options, why consumers engage in or react to social media, etc. This chapter describes comprehensively the "state-of-the-art" of conjoint analysis and choice-based conjoint experiments and related estimation procedures.

Keywords

Preference measurement · Choice experiments · Conjoint analysis · Conjoint measurement · Tradeoff analysis · Choice-based conjoint · Adaptive conjoint · Utility function · New product development · Revealed preference · Incentive-aligned mechanisms · Willingness-to-pay · Market simulation

Introduction

Assume that an electronics company wants to enter the market for ebook readers. The company has already developed a working prototype with the basic functionality. However, consumers did not yet consider buying this specific product according to a survey, but continue to buy a (more expensive) competitor's product instead. The manufacturer therefore would like to know which attributes of an ebook reader are valued by consumers and which specific attributes they need to improve. Given limited budgets, they can only modify their product in one or two attributes, depending on the manufacturing costs, so that they need to reveal which attributes are most important. Moreover, they would like to know how price-sensitive consumers are and how much they are willing to spend for an ebook reader. Finally, they also need an estimate of the achievable market share to reach the final decision if they should market their product or not.

These questions and related ones can be addressed with preference measurement. The aim of preference measurement is to discover the degree how much consumers like or value (i.e., derive a utility from) specific products, which then leads to a purchase decision. Conjoint analysis, as one of the most popular methods within preference measurement, assumes that products are attribute bundles. Accordingly, an ebook reader is considered as a bundle of screen technology, screen size, screen resolution, storage size, brand name, price, etc. The method tries to discover the utilities that each attribute (and attribute level, respectively) adds to the overall utility of the product by systematically varying specific levels of the attribute. It is a decompositional method, meaning that it elicits consumers' overall utilities for experimentally varied product concepts and then decomposes the overall utility into the attributes' utilities (so-called "partworth utilities" or just "partworths") via statistical procedures. In line with this description, the American Marketing

Association (2015) defines conjoint analysis as a “statistical technique in which respondents’ utilities or valuations of attributes are inferred from the preferences they express for various combinations of these attributes.”

As a result, conjoint analysis provides researchers with a utility function that translates the specific attribute levels of a product into consumers’ preferences. This utility function serves multiple purposes; it can explain consumers’ actual purchase decisions and predict their choices given changes to the product configuration, i.e., modification of attributes. In this regard, it is the basis for a multitude of relevant marketing applications, for example:

- New product development and innovation, e.g., which product concept will be preferred by consumers? (e.g., Page and Rosenbaum 1992; Urban and Hauser 1993)
- Pricing, e.g., how much are consumers willing to pay and how much are improvements in products attributes allowed to cost? (e.g., Miller et al. 2011)
- Branding, e.g., how much value can be attributed to the brand of a product? (e.g., Sattler 2005)
- Market segmentation, e.g., are there different market segments that differ in terms of certain preferred product attributes? (e.g., Teichert 2001b)
- Market scenarios, e.g., what is the effect of a new product entry on the market shares of the incumbents? (e.g., Burmester et al. 2016)

Conjoint analysis is not limited to applications in marketing, but can be generally applied when individuals need to make a decision regarding multiattributive objects. It is also a popular method in other areas, such as transportation (e.g., Hensher 1994), litigation (e.g., Eggers et al. 2016), agriculture (e.g., Lusk and Schroeder 2004), or health economics (e.g., De Bekker-Grob et al. 2012). Due to its broad area of applications, conjoint analysis has advanced to a widely respected method since its introduction into marketing in the 1970s. Overviews of its popularity can be found in Green and Srinivasan (1978, 1990) as well as in empirical studies conducted, for example, by Wittink et al. (1994), Voeth (1999), Sattler (2006), and Orme (2016).

Conjoint methods differ in terms of how the overall utilities are elicited. Traditional approaches use ratings of single product concepts (rating-based conjoint), ratings of pairs of products, or rankings of a selection of products (ranking-based conjoint). Currently, the most popular conjoint approach with over 80% of applications (Orme 2016) is based on choices among several product concepts, i.e., choice-based conjoint (CBC; also termed discrete choice experiments; Haaijer and Wedel 2003; Louviere and Woodworth 1983). Using choices as the dependent variable has become popular because they mimic consumers’ behavior when they are making purchase decisions.

Continuing the example case mentioned above, assume that the manufacturer of the ebook reader is currently producing a black ebook reader with a 6-in. E Ink display and 4 GB storage. They are exploring different options to improve their product, e.g., identified via qualitative research or pretests: (1) increasing the storage from 4 GB to 8 GB, (2) increasing the screen size from 6 to 7 in., or (3) changing the

Table 1 List of potential ebook readers (2^3 design)

Concept	Storage (GB)	Screen size (in.)	Color
1	4	6	Black
2	4	7	Black
3	4	6	White
4	4	7	White
5	8	6	Black
6	8	7	Black
7	8	6	White
8	8	7	White

case color from black to white. Accordingly, there are (2^3) eight different options they could potentially offer, resulting from the different combination of attribute levels (Table 1).

Although one could assume that more storage is better so that 8 GB models are preferred to 4 GB models, this is not necessarily true for screen size since consumers might either value a small (and less bulky) product or a larger (and more readable) screen. There is also no a priori preference order for color. Hence, it is not known beforehand which option would be the most preferred one. Moreover, it might not be profitable to offer an 8 GB model if the increase in preference, and therefore demand, is only marginal and does not justify the additional manufacturing costs. Thus, conjoint analysis is a suitable method to solve this decision problem.

Traditional conjoint analysis (e.g., rating-based conjoint) would present each of the products in Table 1 to a consumer in a survey and ask for his/her preference, e.g., on a rating scale from 0 (“not at all preferred”) to 10 (“very much preferred”). The partworth utilities for the attribute levels can then be derived by using the ratings as a dependent variable in a regression model in which the attribute levels serve as independent variables (e.g., as dummy variables). Although ratings can be considered an acceptable manifestation of preferences, they do not mimic consumers behavior in the marketplace. Moreover, it is often questionable how the ratings can be translated into actual choices (Teichert 2001a).

These issues are among the reasons why CBC approaches have become popular. They offer respondents a selection of product alternatives in a choice set (also called “choice task”) and ask for their most preferred option (Fig. 1). This procedure is repeated across multiple sequential choice sets, each presenting alternatives that are systematically varied by an experimental design. The decisions within a choice set often require a trade-off between attributes. For example, if a consumer prefers larger screens (as in option 1 in Fig. 1) and more storage (as in option 2), she/he needs to determine how important each of these attributes really is in order to reach a decision between option 1 and option 2, while also considering color. These decisions increase the realism of the tasks as trade-off decisions are very often required in the marketplace, e.g., when a higher quality is offered for a higher price. Another element that increases the realism of CBC is that it is possible to include a so-called no-choice option (also termed “none option” or “outside good”), which can be

Which of these ebook readers do you prefer?

Please assume that these two options do not differ in terms of other attributes, i.e., both option have a self-lit E Ink display with 758x1024 pixels resolution, WiFi, and 3 weeks battery life. They both support multiple formats (PDF, EPUB) and connect to major book distributors.

	Option 1	Option 2	Option 3
Storage:	4 GB	8 GB	I would not buy any of these
Screen size:	7 inch	6 inch	
Color:	White	Black	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 1 Exemplary choice set of a CBC experiment

chosen if none of the alternatives are acceptable. In this example, the no-choice option could also be termed, e.g., “With these options I would keep reading books on paper,” so that a threshold can be identified which indicates the utility that is needed to make consumers switch from traditional books to an ebook reader.

The higher degree of realism of CBC experiments leads to the expectation that CBC exhibits a higher validity compared to traditional, metric conjoint analysis. However, not all studies find significantly better results for CBC compared to traditional conjoint analysis, although the direction of the effects is as expected (Chakraborty et al. 2002; Elrod et al. 1992; Moore 2004; Moore et al. 1998; Vriens et al. 1998). A disadvantage of CBC experiments is that choices among alternatives are nominal and generate less information than, e.g., rating each alternative separately. Therefore, CBC requires collecting a multitude of sequential choice sets, which might invoke respondent fatigue and could serve as an explanation for those findings in which CBC is not predicting significantly better than rating or ranking-based conjoint.

The traditional conjoint approaches (e.g., rating and ranking-based conjoint) and CBC can be classified as static because they do not adapt to the responses that the consumer has given in the survey. To make the information collection more efficient, adaptive procedures dynamically adjust to the preferences of the respondents. They are typically based on a hybrid approach that combines a decompositional and a compositional method. Compositional approaches (e.g., the self-explicated method) ask respondents directly about their preference for attribute levels and the relative importance of the attributes, e.g., via rating scales (Srinivasan and Park 1997). This input can then be used as a first estimate of the consumer’s preferences in order to show product concepts in the conjoint procedure that are meaningful to the individual respondent or that generate most information about the respondent’s preferences. The rating-based Adaptive Conjoint Analysis (ACA, Johnson 1987) and Adaptive CBC (ACBC, Sawtooth 2014) follow this idea. Other adaptive approaches from the machine learning literature dynamically anticipate each respondent’s utility based on previous answers, i.e., either ratings (Toubia et al. 2003) or choices (Toubia et al. 2004, 2007). Hybrid individualized two-level CBC (HIT-CBC, Eggers and Sattler 2009) uses a compositional approach in order to ask for the best and worst levels for

each attribute and adjusts the CBC part to these two extreme levels only. Thus, it can be seen as a compositional approach in which the attribute importance is derived by a conjoint experiment.

In newer conjoint analysis approaches, respondents interact with each other, following the principles of barter markets (Ding et al. 2009), auctions (Park et al. 2008), or poker games (Toubia et al. 2012). Preferences can then be inferred from these transactions. Figure 2 summarizes the evolution of conjoint analysis approaches.

It should be noted that the above-mentioned example of ebook reader attributes is a very simple case that is used for illustration only. Typically, conjoint studies apply more complex scenarios with more attributes, including price, and additional levels per attribute. Therefore, as an extended example, we will introduce additional attribute levels and a fourth attribute: price. The list of attributes and levels for the extended example is given in Table 2. Because of the popularity of CBC approaches, the remaining chapters will focus on these approaches.

• Static

- Rating-/Ranking-based Conjoint (Srinivasan/Rao 1971)
- Choice-based Conjoint (CBC) (Louviere/Woodworth 1983)

• Adaptive

- Adaptive Conjoint Analysis (Johnson 1987)
- Adaptive CBC (Sawtooth 2014)
- Fast Polyhedral Adaptive Conjoint (Toubia et al. 2003, 2007)
- Hybrid Individualized Two-Level CBC (Eggers/Sattler 2009)

• Interactive

- Upgrading Auctions (Park/Ding/Rao 2008)
- Barter Markets (Ding/Park/Bradlow 2009)
- Conjoint Poker (Toubia et al. 2012)

Fig. 2 Evolution of conjoint analysis approaches

Table 2 Attributes and levels for the extended example

Attribute	Level 1	Level 2	Level 3	Level 4
Storage	4 GB	8 GB	16 GB	n.a.
Screen size	5 in.	6 in.	7 in.	n.a.
Color	Black	White	Silver	n.a.
Price	€79	€99	€119	€139

Model

Conjoint applications assume a (purchase) decision model in which consumer preferences, i.e., utilities, are the central element of the choice process. The assumption is that specific product attributes determine the individual utility evaluations and these, in turn, form the basis for the observed choice behavior (Fig. 3). This requires two interdependent models: a utility model and a choice model, which translates utilities into multinomial choices.

The literature on preference measurement or conjoint-related literature is often equivocal in their terminology. Throughout this chapter, we will use the following terminology (with alternative formulations noted in parentheses): We measure the utility (= preference, need, liking, worth, value) of a consumer (= respondent, individual, subject) for a specific product or service (= alternative, stimulus, object, option, profile) that consists of different attributes (= factors, dimensions), each having specific attribute levels (= characteristics, features).

Utility Model

The basis for the utility model in a choice context is random utility theory (RUT), which states that the overall utility U of consumer c for a product i is a latent construct that includes a systematic component V and an error component e , i.e., $U_{ci} = V_{ci} + e_{ci}$ (McFadden 1981; Walker and Ben-Akiva 2002). The stochastic error term catches all effects that are not accounted for and can include, e.g., respondent fatigue, omitted variables, biases in the data collection, or unaccounted heterogeneity (Louviere and Woodworth 1983).

The theory assumes that a consumer chooses the product from a set of alternatives that exhibits the highest utility. Since the overall utility is influenced by a stochastic component, it is only possible to state a probability that this consumer would choose the product. Consequently, the probability p that a consumer chooses product i from a set of products $S = \{i, j\}$ is (Train 2009):

$$p_i = p(U_i > U_j) = p(V_i - V_j > e_j - e_i) \quad (1)$$

According to Eq. (1) a consumer is more likely to choose product i if the utility of i is larger than the utility of j . This requires that there is a positive residual from the difference in systematic utilities and that this residual exceeds the influence of error. Consequently, only differences in product attributes are considered, e.g., if consumers need to choose between two ebook readers and both devices are black then

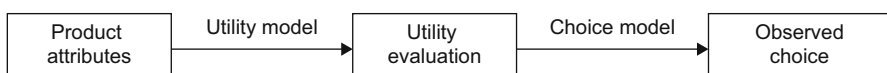


Fig. 3 Elements of a purchase decision model

color does not affect the decision. Generally, any constant value can be added to the utility functions and it will not affect the outcome, which is why choice-based utilities are interval-scaled and choice models do not have a general regression constant (constants, if any, need to be alternative-specific).

The systematic utility V represents the function that translates the product attributes and their levels into partworth utilities. The estimated utility V_i for a product i with N attributes can be divided into two subfunctions ψ and f_n as follows (Teichert 2001a):

$$V_i = \Psi[f_1(v_{1i}), f_2(v_{2i}), \dots, f_N(v_{Ni})] \tag{2}$$

with

v_{ni} : Partworth utility of attribute n in product i , $n = 1, 2, \dots, N$

f_n : Evaluation function of attribute n , $n = 1, 2, \dots, N$

ψ : Function to combine partworth utilities across attributes

Evaluation Function for Attribute Levels

The function f_n in Eq. 2 describes how levels of attribute n are evaluated. The basic idea is that at least one attribute level represents the ideal point for the consumer (or at least the most preferred level from the available attribute levels). Differences to this ideal point lead to a loss in utility. Figure 4 depicts three potential functional forms.

The vector model assumes that increasing (decreasing) the attribute level leads to a proportional positive (negative) effect in utility. Hence, the ideal point is positive (negative) infinity. This model would be appropriate when assuming, e.g., that increasing the screen size of an ebook reader from 5 to 6 in. leads to the same positive utility difference as upgrading the screen from 6 to 7 in. The vector model uses the actual numeric values of the attributes and just one utility parameter to represent the partworth utility:

$$v_{in} = \beta_n * X_{inm} \tag{3}$$

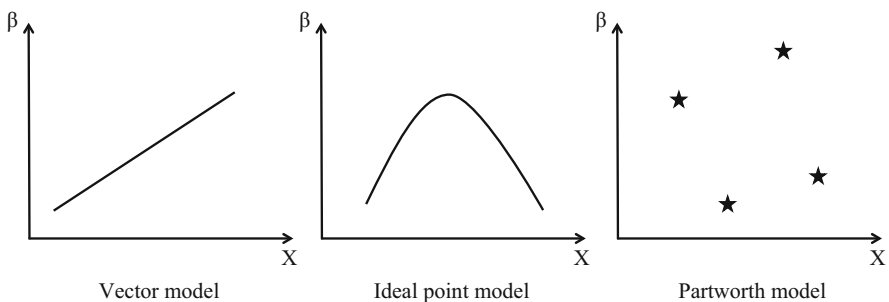


Fig. 4 Alternative functional forms for the evaluation of attribute levels

with,

v_{in} : partworth utility for attribute n in product i
 β_n : utility vector for attribute n
 X_{inm} : numeric value of level m of attribute n in product i

The ideal point model does not assume a linear slope of the utility function as the vector model but assumes diminishing (or increasing) marginal utilities. For example, although consumers might in general prefer larger screens for an ebook reader, very large sizes will become impractical so that utilities will decrease again when increasing the size from an (individually perceived) ideal point further. Likewise, when an ebook reader already has a very large storage, it can be expected that increasing the storage further leads to a diminishing marginal utility for the consumer. The ideal point model thus considers not only the numeric value of the attribute level, e.g., its screen size, but also its squared term:

$$v_{in} = \beta_{n1} * X_{inm} + \beta_{n2} * X_{inm}^2 \quad (4)$$

with,

v_{in} : partworth utility for attribute n in product i
 β_{n1} : utility vector for attribute n
 β_{n2} : utility vector for the squared value of attribute n
 X_{inm} : numeric value of level m of attribute n in product i

The partworth model estimates separate partworth utilities for each level of the attribute, i.e., there is no assumed functional relationship between the attribute levels. This model is required for qualitative, nominal attributes, e.g., color, but can also be applied to quantitative, numeric attributes. If the choice sets include a no-choice option, this option is also represented by a separate partworth that measures the attractiveness of not choosing any of the alternatives. The partworth model is typically based on dummy-coding (or effect-coding) techniques, which requires $M-1$ variables to represent an attribute with M levels:

$$v_{in} = \sum_{m=1}^{M-1} \beta_{nm} * X_{inm} \quad (5)$$

with,

v_{in} : partworth utility for attribute n in product i
 β_{nm} : partworth utility for level m of attribute n
 X_{inm} : dummy variable with value 1 if product i features level m of attribute n , otherwise 0

Regarding the number of parameters that these models require for the estimation, the vector model is the most parsimonious as it only uses one parameter per attribute.

The ideal point model is based on two parameters. The partworth model requires setting one attribute level as the reference level, which is left out of the estimation so that it requires $M - 1$ parameters.

The partworth model can be considered conservative since it does not require a prior specification or theory about the slope of the partworth utility function. If more than two attribute levels are present, it uses the most number of parameters and therefore provides the best model fit (by sacrificing degrees of freedom). It is therefore not surprising that the partworth model is predominantly used in conjoint analysis and is partly also considered as a constitutive element (Shocker and Srinivasan 1973).

Function to Combine Partworth Utilities Across Attributes

The function ψ in Eq. 2 determines how to combine partworth utilities across attributes. Conjoint analysis assumes a compensatory utility model. In a linear additive utility model, the overall systematic utility V_i of a product i is the sum of the partworth utilities v_{in} of its attributes $n = 1, \dots, N$:

$$V_i = \sum_{n=1}^N v_{in} \tag{6}$$

Complex functions can be modeled as extension to this base model, e.g., interaction effects between attributes. Interaction effects occur when the utility evaluation of one attribute level depends on the level of another attribute. For example, consumers might prefer a white color for ebook readers with large screens but black for readers with smaller screens.

Interaction effects can be modeled as additional effects in the linear additive base model by including separate partworth utilities for the cross product of two attributes. The overall utility for a product is then represented as the sum of the partworth utilities of both the main effects and the interaction effects:

$$V_i = \sum_{n=1}^N v_{in} + \sum_{m=1}^{M-1} \sum_{m'=1}^{M'-1} \beta_{nm,n'm'}^{IA} * X_{inm} * X_{in'm'} \tag{7}$$

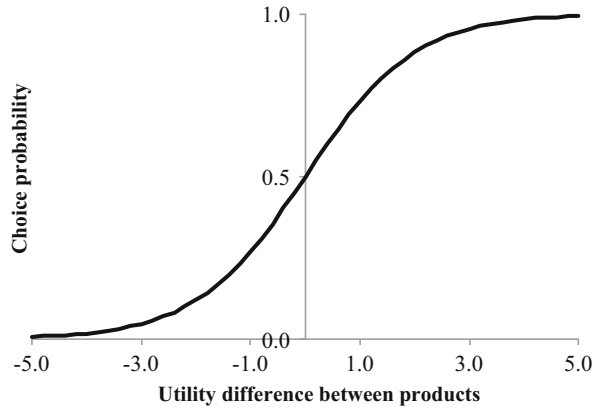
with,

$\beta_{nm,n'm'}^{IA}$: Interaction effect between level m of attribute n and level m' of attribute n' ;
 $m = 1, 2, \dots, M; m' = 1, 2, \dots, M'$

$X_{inm}; X_{in'm'}$: Dummy variable with value 1 if product i features level m (m') of attribute n (n'), otherwise 0

Interaction effects increase the complexity of a model. For this reason, they are predominantly added if theory or prior assumptions about them exist. However, being able to measure interaction effects with conjoint analysis is a major advantage compared to other survey techniques, e.g., compositional approaches.

Fig. 5 S-shaped function of the multinomial logit model



Choice Model

Choice models can be differentiated according to the assumptions about the stochastic error component (see Train 2009 for an overview). In most applications, the error is assumed to be independent and identically distributed (iid) as extreme value type, i.e., Gumbel. This assumption leads to a logistic distribution of the differences of error terms and the multinomial logit (MNL) model (McFadden 1981; Hensher and Johnson 1981; Louviere et al. 2000). Accordingly, choosing an object i from a choice set with S alternatives is represented by the MNL model in terms of choice probabilities p :

$$p(i|S) = \frac{\exp(V_i)}{\sum_{j \in S} \exp(V_j)} \quad (8)$$

The MNL model results in an S-shaped relationship between utility difference and choice probability (Fig. 5).

An alternative to the Gumbel distribution is the assumption of a normal distribution of the error term, which results in a multinomial probit model (Haaijer et al. 1998). The probit model requires multiple integrals and complex estimation procedures. Because of the compact form of the logit function (see Eq. 8), the MNL model is predominantly applied in CBC analyses (Haaijer and Wedel 2003).

Procedure for Conducting Discrete Choice Experiments

Identification of Attributes and Attribute Levels

The prerequisite – and most relevant step – for conducting conjoint analyses is to identify the relevant determinants of consumers' choices, i.e., product attributes and their levels. The selection of attributes and levels should reflect the products on the

marketplace and should affect consumers' preferences. Otherwise, the validity of the model can be questioned. In general, the selection of attributes has to fulfill the following requirements (Green and Srinivasan 1978, 1990; Orme 2002):

- Attributes should be relevant, i.e., they should influence consumers' utility. In order to identify relevant attributes qualitative surveys, e.g., focus groups or depth interviews can be used.
- Attributes should discriminate, i.e., they should be able to differentiate between the competitive offerings on the marketplace.
- The number of attributes should be manageable. CBC experiments typically use less than seven attributes. Using more attributes greatly increases the complexity of the experimental design and requires high cognitive capabilities of the respondents.
- Attributes should not be interrelated, i.e., they should measure independent aspects of the product. If attributes are interrelated, then certain combinations might be highly unrealistic and confusing to the respondents. However, if, e.g., higher storages typically go along with higher prices, it is possible to consider these attributes as independent and analyze "what-if" scenarios. It should be noted that this requirement does not preclude potential interaction effects, i.e., although the attributes are independent, it does not mean that the preferences for them are as well.

After setting the attributes, their levels need to be determined. Regarding the type and number of levels, the following requirements should be considered (Green and Srinivasan 1978, 1990; Orme 2002; Teichert 2001a):

- The levels should span a range that is larger than in reality, but not substantially, in order to be able to cover potential future scenarios.
- Levels that have an ambiguous meaning should be avoided. For example, instead of using levels "large" and "small" for screen size, it is better to use specific values because they are free from interpretation. Moreover, specific values allow using a vector or ideal point model for estimation.
- The number of levels should be kept low because the complexity of the experimental design will increase exponentially with more levels. Consider the example in Table 1 with $2^3 = 8$ combinations. If three levels per attribute were used instead there are already $3^3 = 27$ potential options. Conjoint experiments can consider complex designs, however, most applications use an average of three to four levels per attribute.
- When setting the number of attribute levels, it should also be considered if the linearity or nonlinearity of the utility function (e.g., an ideal point model) should be tested, which then requires at least three levels. For testing interaction effects, it would be preferable (but not required) to use just two levels in order to keep the number of interaction effect parameters low.
- The number of levels should be balanced across attributes. Otherwise, the number-of-levels effect can occur, which leads to an artificially higher relevance of attributes that have more levels (Eggers and Sattler 2009; Verlegh et al. 2002).

- Levels should be generally acceptable. Unacceptable levels would otherwise invalidate the assumed compensatory utility model.
- Attribute levels are assumed to be mutually exclusive. For example, if an attribute “extra features” is added to the ebook reader setup with the levels “waterproof” and “integrated music player,” the reader can only have one of these levels. If it is also interesting for the researcher to analyze preferences for both features in combination, this combination should be added as a separate level (an alternative would be to define each extra as a separate attribute with the levels “yes” and “no”).

Creating the Experimental Design

The experimental design determines which combinations of attribute levels are presented to the respondent as stimuli (factorial design) and how these stimuli are allocated to choice sets (choice design). It represents the independent variable matrix for the analysis. To estimate the main effects of the attributes – and potentially interaction effects between them – the experimental design needs to make sure that these effects can be identified.

Criteria to evaluate the efficiency of an experimental design are (Huber and Zwerina 1996):

- Balance, i.e., each attribute level is presented an equal number of times
- Orthogonality, i.e., attribute levels are uncorrelated
- Minimal overlap, i.e., alternatives within a choice set are maximally different
- Utility balance, i.e., alternatives within a choice set should be equally attractive so that there should not be dominated or dominating alternatives

Balance and orthogonality refer to the factorial design, while minimal overlap and utility balance relate to the choice design.

Factorial Design

The set of all potential stimuli, i.e., every combination of attribute levels, leads to a full factorial. With N attributes and M_1 levels for attribute 1, M_2 levels of attribute 2, and M_N levels of attribute N , the size of the full factorial consists of all permutations $M_1 * M_2 * \dots * M_N$. Table 1 shows a full factorial of the 2^3 design. Full factorials are always balanced, i.e., the attribute levels occur an equal number of times (here, four times), and orthogonal, i.e., each pair of attribute levels is balanced (here, each pair occurs twice).

A full factorial is only required if all main effects and all potential interaction effects should be estimated. The 2^3 design with three binary attributes A , B , C allows to estimate the three main effects, the three two-factor interaction effects ($A*B$, $A*C$, $B*C$), as well as the three-factor interaction ($A*B*C$). This is demonstrated in Table 2, in which the attribute levels are effect-coded (first level = 1, second level = -1). The interaction levels result from multiplying the levels of the

underlying main effect attributes. As can be seen, the resulting interaction levels are not identical to any other column, i.e., are independent, and are also balanced and orthogonal so that they can be identified.

Since the full factorial increases exponentially when more attributes and/or more attribute levels are added, its size quickly becomes hard to handle in an experimental survey. For example, the extended example with three three-level attributes and one four-level attribute consists of $3^3 * 4 = 108$ potential alternatives. Moreover, very often three-factor interaction effects can be neglected and not all two-factor interaction effects may be required. In general, smaller factorials, i.e., fractional factorials, still allow estimating main effects and selected interaction effects (Addelman 1962).

The idea of creating a fractional factorial design is demonstrated with an example. Consider that a fourth binary attribute D would be added to the simple example in Table 3. The full factorial would then increase to $2^4 = 16$ stimuli. A fractional design assumes that at least one of the interaction effects between the attributes A, B, and C would be zero so that it can be replaced with the main effect of D, e.g., $D = A*B$, i.e., each level of the interaction between A and B becomes the new level of D. The fractional factorial then consists of the 8 entries in Table 2 and columns A, B, C, as well as $D = AB$. The factorial was reduced to 8 stimuli, i.e., by 50% compared to the full factorial. Nevertheless, it is still able to identify all main effects, i.e., the design is still balanced and orthogonal. As a downside, however, the interaction effect between A and B cannot be estimated as it is confounded with the main effect of D.

Fractional factorials are documented for the most common experimental designs (e.g., Sloan 2015) or can be generated via software (e.g., SAS or SPSS). The efficiency of the fractional design can be tested easily by checking the correlation matrix of all assumed main and interaction effects. If there are no or only minor correlations, then the design is orthogonal and the parameters can be identified without bias.

For traditional rating- or ranking-based conjoint procedures, it is sufficient to evaluate the factorial design. CBC methods require an additional step of allocating alternatives of the factorial design to specific choice sets, i.e., to evaluate the choice design.

Table 3 Main and interaction effects of a full factorial 2^3 design

Stimulus	Main effect			Interaction effects			
	A	B	C	AB	AC	BC	ABC
1	-1	-1	-1	1	1	1	-1
2	-1	-1	1	1	-1	-1	1
3	-1	1	-1	-1	1	-1	1
4	-1	1	1	-1	-1	1	-1
5	1	-1	-1	-1	-1	1	1
6	1	-1	1	-1	1	-1	-1
7	1	1	-1	1	-1	-1	-1
8	1	1	1	1	1	1	1

Choice Design

Choice experiments require that the factorial is subdivided into choice sets with a selection of alternatives. Creating an optimal choice design involves complex algorithms based on combinatorics. For example, even with the simple example and a 2^3 full factorial, there are $\binom{8}{2} = 28$ different choice sets with two alternatives. The complexity increases with the size of the factorial, e.g., in the extended example there would be $\binom{108}{3} = 205,156$ potential choice sets of size three. The challenge lies in selecting those choice sets that provide the most information about the respondents' preferences. The efficiency criteria minimal overlap and utility balance help reducing the size of the list of potential choice sets (Huber and Zwerina 1996).

Minimal overlap requires that the alternatives within a choice set are maximally different, i.e., have different attribute levels (Sawtooth 1999). It is based on the idea that an attribute that exhibits the same level for each alternative within a set does not affect the choice (see Eq. 1). A choice design with minimal overlap can be created for the simple example when the first four entries in the full factorial in Table 1 are coupled with their fold-over, i.e., opposite level. Accordingly, concept 1 (4 GB, 6 in., black) would be coupled with concept 8 (8 GB, 7 in., white) to create one choice set; concept 2 would be coupled with concept 7, etc., so that in total four choice sets with minimal overlap are created.

The idea of selecting choice sets that are utility balanced is that alternatives are allocated to a choice set that are equally attractive (Huber and Zwerina 1996). Contrarily, a choice set that features a dominating or dominated alternative provides no new knowledge since the choice can be anticipated. However, dominating alternatives can only be identified if there is a priori knowledge about the respondents' preference structure or if respondents' preferences are anticipated during the experiment with adaptive conjoint approaches (see above).

Because of the complexity of creating an optimal choice design, computer algorithms are recommended. For example, SAS or Sawtooth offer algorithms to create optimal choice designs and analyze their efficiency.

Decision Parameters

Relevant decision parameters for the experimental design also concern the number of stimuli per choice set and the number of choice sets.

Each choice task should be manageable for the respondent, which favors showing only a few alternatives per set (Batsell and Louviere 1991). On the other hand, more alternatives increase the information of each choice. Therefore, two to five stimuli per choice set are most common (Meissner et al. 2016). Using eye-tracking data, Meissner et al. (2016) show that the number of alternatives also affects search patterns. It is therefore advisable to use a choice set size that is similar to the typical size of a consideration set when consumers make purchase decisions. In product categories in which consumers frequently have to choose from a multitude of alternatives, e.g., toothpaste in supermarkets, choice sets could also include a larger number of alternatives (Hartmann 2004). The

selection of the number of alternatives should also consider the number of attribute levels since using a number of alternatives that is a subset of the number of levels provides statistical benefits (Zeithammer and Lenk 2009).

Apart from the number of alternatives per choice set, the number of choice sets needs to be considered when selecting an optimal design. More choice sets lead to a higher reliability of the parameters. However, from a consumer perspective, more choice sets induce fatigue so that respondents tend to make more errors or even switch their decision strategy, e.g., focusing more on the price attribute (Johnson and Orme 1996), which is counterproductive. Consistently, results concerning the predictive validity depending on the number of choice sets indicate that the marginal benefit of additional choice sets declines (Sattler et al. 2004; Teichert 2001a). A review of articles published in the *Journal of Marketing Research* between 2000 and 2017 shows that most researchers make a compromise between statistical reliability and consumer fatigue so that most applications (14 out of 42) have used 11–15 sets. Slightly fewer studies (13 out of 42) have used ten sets or less. The number of applications decreases with more choice sets, i.e., nine studies used 16–20 choice sets, five applications 21–25 sets, and one study more than 25.

Implementation into Questionnaire

The implementation of the CBC experiment into a questionnaire requires decisions regarding the presentation of stimuli, integration of a no-choice option, collecting additional choices per choice set, applying incentive alignment mechanisms, and adding holdout choice sets.

Presentation of Stimuli

Most CBC interviews are computer-based since they facilitate handling complex experimental designs. Moreover, having more than two alternatives per choice set puts high cognitive burden on respondents, e.g., when described via telephone interviews. Computer-based interviews are beneficial because they allow implementing attribute levels or overall stimuli as multimedia information. Instead of using text only it is possible to depict the size of the ebook reader screens as a pictogram or to show actual ebook readers in different colors. When certain functionalities, e.g., page-turn effects, are included as attributes, these could be showcased with instructional videos (e.g., following the idea of information acceleration, Urban et al. 1996). Eggers et al. (2016) demonstrate that the more realistic the experiment can be made compared to what consumers see in the marketplace, i.e., investing in “craft,” the higher is the validity of the results, which might also change the managerial implications from the results compared to studies that rely on defaults, e.g., text-only descriptions of the stimuli.

No-Choice Option

An advantage of CBC experiments compared to metric (rating or ranking-based) conjoint analyses is that respondents can indicate that they prefer none of the

presented alternative. This none (or no-choice) option increases the realism since it does not force a decision if the alternatives are unacceptable so that consumers would not buy any of them or switch stores in reality (Haaijer et al. 2001). Recent approaches suggest asking for the no-choice option separately, i.e., sequentially after each choice set (“dual response none”; Brazell et al. 2006). In the dual response procedure respondents are first asked to select the most preferred option (excluding no-choice) in a forced-choice task and, sequentially, whether they would purchase the selected product concept in a second step (Brazell et al. 2006; Wlömert and Eggers 2016).

This procedure allows observing the preferred alternative even if it is not acceptable to be purchased. At the same time, consumers have no possibility to opt out of difficult decisions. Moreover, Wlömert and Eggers (2016) show that the increased salience of the no-choice option leads to more realistic predictions of adoption shares.

The no-choice option plays a central role when calculating (absolute) willingness-to-pay (see section “Market Simulations”). Implications from these analyses are limited if consumers show extreme response behavior and never or always choose the none option. To avoid these extremes, Gensler et al. (2012) present an adaptive approach that dynamically adjusts the price levels downwards whenever the respondent selected the no-choice option and upwards whenever the respondent selected an alternative. Schlereth and Skiera (2016) address this issue by proposing a separated adaptive dual response (SADR) procedure. They adjust the dual response procedure so that the forced choice and purchase question are not presented within the same task but are separated into sequential blocks. Presenting the block of forced choices first allows them to approximate the utility of the alternatives and adaptively select fewer, but more informative alternatives (not necessarily the alternatives selected in the forced choices) in the purchase questions thereafter.

Collecting Additional Choices per Choice Set

Recently, it was suggested to ask not only for the best option but also for the worst option in a so-called best-worst scaling (or MaxDiff) approach (Louviere et al. 2015; Sawtooth 2013). By assuming that worst choices are reversed best choices, both decisions measure the same construct, i.e., preferences. Stated differently, if β_{nb} represents the partworth utility for attribute n based on best choices and β_{nw} is the partworth utility for the same attribute based on worst choices then it can be assumed that $\beta_{nb} = -\beta_{nw}$. The choices can then be used to make the estimation more reliable since twice as many observations exist. Collecting more choices per set is not limited to best and worst decisions only. More choices can be used as separate dependent variables in order to explore different aspects of consumers’ preferences. An additional choice can be, e.g., “Which of these ebook readers would you buy for your partner?” which might explore consumers’ gift giving behavior. In a study by Kraus et al. (2015), the authors collected additional choices per set to analyze managers’ perception of risk and success of different internationalization strategies.

Figure 6 shows an example of a choice set that includes best and worst choices and a dual response no-choice option.

Which of these ebook readers is your most preferred option and which option is the least attractive?

Please assume that these options do not differ in terms of other attributes, i.e., all options have a self-lit E Ink display with 758x1024 pixels resolution, WiFi, and 3 weeks battery life. They support multiple formats (PDF, EPUB) and connect to major book distributors.

	Option 1	Option 2	Option 3
Storage:	4 GB	8 GB	16 GB
Screen size:	6 inch	7 inch	5 inch
Color:	Silver	Black	White
Prize:	€99	€119	€139
Best option:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Worst option:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Would you actually buy your most preferred option if it was available?

Yes

No

Fig. 6 Choice set with best and worst choices and dual response no-choice option

Incentive Alignment

Ding et al. (2005) introduced incentive alignment mechanisms to conjoint analysis. The basic idea of incentive-aligned (IA) mechanisms is to attenuate hypothetical bias by influencing the type of reward that is provided to respondents. Specifically, the reward is linked to the preferences the respondent expresses during the data collection.

Ding et al. (2005) implemented the IA mechanism by rewarding the respondent with the alternative that she/he selected in a randomly selected choice task (including the no-choice option). In this way, each choice might constitute the potential reward so that respondents are motivated to answer truthfully. If the study features a price attribute, then respondents are required to actually purchase the product for the price shown. Payment is typically achieved by providing the respondents with a budget. If the respondent selected the no-choice option, she/he gets the full budget as a monetary reward. If she/he selected a product for a price €X, she/he will get the actual product plus the remaining change (i.e., initial budget minus €X).

Ding (2007) proposed an alternative IA approach in which respondents are informed before completing the choice tasks that their choices will be used to infer their willingness-to-pay (WTP) for one specific product concept (see sections “Willingness-to-Pay” and “Market Simulations” for details about calculating WTP). Under this WTP-based mechanism, incentive alignment is achieved by obliging participants to purchase this specific product concept at a randomly drawn price if this random price is less or equal to the WTP inferred from the CBC experiment. This approach integrates the incentive compatible Becker-DeGroot-Marschak (BDM) auction procedure (Becker et al. 1964, see also Wertenbroch and Skiera 2002) with CBC analysis. Ding (2007) shows theoretically

that truthful answers constitute the Bayesian Nash equilibrium for participants in such applications as long as the respondents do not know the configuration of the product that is used as a reward prior to the study.

Dong et al. (2010) introduced and validated a third variant of IA conjoint experiments which involves predicting a rank ordering of the possible rewards based on estimated preferences. Eventually, the reward that is predicted to be ranked first is given to the respondent. Again, respondents are motivated to answer truthfully and keep the impact of error small in order to be rewarded with their most preferred product.

It has been shown that incentive-aligned (IA) data collection procedures substantially increase the predictive performance of conjoint choice experiments compared with traditional CBC analysis (Ding 2007; Ding et al. 2005; Dong et al. 2010) so that their application is recommended. However, one drawback of incentive alignment is that their application is limited to contexts where at least one concept of the research object can be rewarded after the experiment. This may not be feasible in many instances, for example, when the research object is an innovative product and not yet available on the market.

Holdout Choice Sets

A holdout choice set is a choice task that mimics a regular choice set but that is not used in the estimation. The answers given in the holdout choice set provide a benchmark for the (internal) predictive validity of the estimation results. The better the preference estimates are able to predict the actual choices made in the holdout sets the higher the predictive validity. Validity can be assessed with different measures. The hit rate compares on an individual level if the predicted most preferred alternative based on the estimates equals the alternative actually chosen in the holdout set, i.e., a hit meaning a correct prediction. The hit rate is then the mean value across all respondents. The mean absolute error (MAE), as an alternative measure among others, considers the absolute differences between predicted and actual choice shares for each alternative in the holdout set (e.g., Moore et al. 1998).

Estimation

Since choices from choice sets typically do not provide enough information to estimate reliable utilities at the individual level, they require some level of aggregation (see Frischknecht et al. 2014 for an alternative approach). The estimation procedure described here is based on the maximum likelihood procedure. It aggregates all choices from all respondents and produces one set of utilities that represent all consumers, i.e., it neglects consumer heterogeneity (see section “[Advanced Estimation Techniques](#)” for advanced estimation procedures without this assumption).

We will use the MNL model for describing the estimation in more detail. The estimates are based on the extended ebook reader example. The (simulated) data are based on 200 respondents who answered 10 choice sets, each showing three product alternatives plus a no-choice option.

Coding

The estimation of partworth utilities requires transforming the attribute levels according to a dummy (or effect) coding technique. When applying a partworth utility model to an attribute with M levels, $M - 1$ dummy-coded variables are needed to represent this attribute in the estimation. Each variable represents one attribute level and can take the values 1 or 0 depending on whether the attribute level was shown or not. The M^{th} attribute level (or any other level) is left out since it can be expressed as a linear combination of the other variables and cannot be estimated separately. The partworth utility of this reference level is set to 0. The partworth utilities of the remaining attribute levels need to be interpreted in relation to this level. Thus, it matters for the interpretation which level represents the reference.

Conjoint experiments are frequently coded using effect-coding. Effect-coded variables (Louviere et al. 2000), as an alternative to dummy-coding, are zero-centered so that the sum of partworth utilities across all levels of the attribute is zero, i.e., positive partworth utilities indicate higher preferences for that level compared to the average partworth utility across all levels of the attribute. Therefore, positive or negative values do not necessarily mean that these levels are perceived as positive or negative on an absolute level but only compared to the mean of the levels that were included in the experimental design. The reference level, which is left out of the estimation, can be recovered by calculating the partworth utility that is needed so that the average across all utilities is zero. Effect-coding therefore provides a partworth utility value for each attribute level, and it is irrelevant which level is set as the reference.

Effect-coding can be accomplished by setting the reference level to -1 , instead of 0 as in dummy coding. Table 4 shows an example of effect-coding two attributes with $M = 3$ and $M = 4$ levels. Figure 7 shows an excerpt of the first two choice sets from the ebook reader dataset. In this dataset, each alternative (indicated by *Alt_id*) is represented by one row such that four rows represent one choice set (indicated by *Set_id*). The none option is included as one of the alternatives, which is represented by the *None* variable. The columns in dark grey show the numeric values for screen size, storage, and price, and text information for color. Effect-coding (columns in light grey) needs two parameters each for the attributes storage, screen size, and color, and three parameters for the effect-coded prices. This means that a partworth model requires ten parameters in total, i.e., nine parameters for the effect-coded variables and one variable for the none option (here, dummy coded). The column *Selected* is a dummy coded variable that shows which alternative was chosen in each choice set. It serves as the dependent variable in the estimation model.

Table 4 Effect-coding of attribute levels

Level	Effect-coded variables for $M = 3$		Effect-coded variables for $M = 4$		
	X_1	X_2	X_1	X_2	X_3
1	1	0	1	0	0
2	0	1	0	1	0
3	-1	-1	0	0	1
4			-1	-1	-1

Resp_ id	Set_ id	Alt_ id	Selected	None	Storage	Screen. size	Color	Price	Storage_ 4GB	Storage_ 8GB	Screen. size_ 5inch	Screen. size_ 6inch	Color_ black	Color_ white	Price_ 79	Price_ 99	Price_ 119
1	1	1	0	0	4	7	Silver	119	1	0	0	-1	-1	-1	0	0	1
1	1	2	1	0	16	5	White	79	-1	-1	1	0	0	1	1	0	0
1	1	3	0	0	8	6	Black	99	0	1	0	1	1	0	0	1	0
1	1	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	2	1	0	0	8	5	Silver	139	0	1	1	0	-1	-1	-1	-1	-1
1	2	2	0	0	16	6	Black	79	-1	-1	0	1	0	1	1	0	0
1	2	3	0	0	4	7	0	119	1	0	-1	-1	1	0	0	0	1
1	2	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 7 Excerpt from the ebook reader dataset

Maximum Likelihood Estimation

Applying OLS procedures for the estimation is not appropriate because CBC analyses provide nominal data. The estimation of the MNL model therefore relies on maximum likelihood procedures. In aggregate-level analyses, all respondents are pooled to estimate one set of partworth utilities for the entire sample (Louviere and Woodworth 1983; Sawtooth 1999).

The maximum likelihood procedure aims at finding the set of partworth utilities that best represents the observed choices. The likelihood function L results from multiplying the MNL probabilities as shown in Eq. (8) across all choice sets $t = 1, 2, \dots, T$ and – in the aggregate-level estimation – across all respondents $c = 1, 2, \dots, C$ (Louviere et al. 2000):

$$L = \prod_c^C \prod_t^T p(i_{tc} | S_{tc}) \tag{9}$$

with,

i_{tc} = chosen alternative in choice set t by respondent c

S_{tc} = alternatives in choice set t presented to respondent c

The parameters can be found by maximizing the function subject to the partworth utilities, i.e., $\frac{\partial L}{\partial \beta} = 0$.

The likelihood function lies in the interval $[0, 1]$ and expresses the aggregate probability to observe the choice data given the set of estimated partworth utilities. However, the minimum of zero is only a theoretical value as choosing randomly between the choice options, i.e., assuming that all betas are zero, would yield a probability of $1/S$, with S being the number of alternatives in the choice set. For example, choosing randomly between three ebook readers and the no-choice option would give a probability of $1/4$ that the choice matches the respondents preferred option. The lowest logical value of the likelihood function is therefore $(1/S)^{(T \cdot C)}$. Since this value is very close to zero, the optimization of the function is typically based on the logarithm, i.e., log-likelihood function (Louviere et al. 2000). The lowest value, and the benchmark to assess the model fit, then is $T \cdot C \cdot \log(1/S)$, e.g., for the ebook reader case with 10 choice sets with four alternatives and 200 respondents: $10 \cdot 200 \cdot \log(1/4) = -2772.6$. The estimation model should exceed this value significantly, i.e., have a log-likelihood value that is less negative (closer to zero), because otherwise the partworth utilities would not predict choices better than a random, NULL model.

Estimating the partworth utilities based on the ebook reader example yields a log-likelihood value of -2277.8 . To test if the difference in log-likelihood between the NULL model and the estimated model is significant, a likelihood ratio test can be applied. The test statistic is $\chi^2 = 2 \cdot (LL_1 - LL_0)$, with LL_1 representing the log-likelihood of the estimated model and LL_0 the log-likelihood value of the NULL model. This test statistic is distributed χ -squared with degrees of freedom (df) equal to the difference in the number of parameters between both models. In this case, χ^2 is

$2 * (-2277.8 - (-2772.6)) = 989.6$, with $df = 10$. This test is highly significant ($p < 0.001$), i.e., the estimated model predicts significantly better than the NULL model.

Another measure to assess the goodness of fit is the Pseudo- R^2 or McFadden's $R^2 = 1 - (LL_1/LL_0)$. For the ebook reader example, it is: $R^2 = 1 - (-2277.8/-2772.6) = 0.178$. McFadden's R^2 can be adjusted according to the number of parameters, i.e., $1 - ((LL_1 - npar)/LL_0)$, with $npar$ being the number of parameters. This R^2 value has a different interpretation than in linear regression models. Typically, values exceeding 0.2–0.4 are considered acceptable. Although the ebook reader model is significantly different from the NULL model, its fit relative to this benchmark is not exceeding the threshold of 0.2. A potential explanation for this low fit is that consumers likely have heterogeneous preferences, e.g., towards screen size or color, which are not acknowledged in the aggregate model and therefore increase the error term.

The estimated partworth utilities are depicted in Table 5 (see “Appendix” for the corresponding R code). The partworth utilities for the attribute levels are effect-coded, which can be seen by checking that the sum across the betas is zero. The betas for storage and price show face validity as increasing the storage (price) yields higher (lower) utilities. There is no such trend regarding screen size as 6-in. models have the highest utility, followed by 5-in. models and 7-in. screens. White ebook readers are more preferred than black and silver models.

The no-choice option was dummy coded in this case, with “no-choice” equal to one and “not the no-choice” equal to zero. As can be seen, not choosing one of the

Table 5 Estimated partworth utilities based on the aggregate-level model

Attributes	Beta	Standard error	t-value	Attribute importance
Storage				21.6%
4 GB	-0.389	0.042	-9.323	
8 GB	-0.051	0.039	-1.322	
16 GB	0.440	0.036	12.143	
Screen size				22.0%
5 in.	-0.049	0.039	-1.274	
6 in.	0.446	0.036	12.352	
7 in.	-0.397	0.042	-9.528	
Color				12.5%
Black	-0.002	0.038	-0.059	
White	0.240	0.037	6.547	
Silver	-0.238	0.040	-5.952	
Price				43.9%
€79	0.840	0.045	18.502	
€99	0.286	0.047	6.103	
€119	-0.284	0.053	-5.416	
€139	-0.842	0.063	-13.447	
No-choice				
	-0.532	0.069	-7.749	

ebook readers shows a negative partworth utility so that on average (i.e., with all attributes at their mean utility of zero), choosing one of the ebook readers provides a higher utility and is therefore more likely than choosing none.

The partworth utilities can be transformed to be more accessible for managerial use compared to the rather abstract units of utility. Three transformations shall be elaborated subsequently: relative attribute importances, willingness-to-pay measures, and calculation of purchase probabilities within market simulations.

Relative Attribute Importance

The attribute importance w_n of an attribute n can be calculated based on the relative range of the partworth utilities, i.e., the difference between the most and least preferred attribute levels related to the sum of ranges across all attributes:

$$w_n = \frac{\max(\beta_n) - \min(\beta_n)}{\sum_{i=1}^N (\max(\beta_i) - \min(\beta_i))} \quad (10)$$

For example, storage exhibits a range of 0.829 ($=0.440 - (-0.389)$). The sum of all attribute ranges is 3.832. The relative importance of storage is therefore $0.829/3.832 = 21.6\%$. The attribute importance serves as a first indicator which attribute is most influential in affecting respondents' choices. However, these attribute importances only consider the extremes of the partworth utilities and not the intermediate levels. Moreover, the importances can only be interpreted in the context of the selected attributes and levels. Additionally, the attribute importance has to be evaluated in the context of the ability to discriminate between market offerings (Bauer et al. 1996). For example, most ebook readers on the market are 6-in. models. Although the attribute is the second most important based on the range of partworth utilities, it is less managerially relevant since most manufacturers are already offering the most preferred size so that using this attribute level does not help to differentiate from the competitors.

Willingness-to-Pay

The willingness-to-pay (WTP) transformation is based on the idea to analyze how much utility is lost (gained) when the price increases (decreases) and to relate this utility difference to the partworth utility of an attribute level. As a result, the partworth utilities for nonprice attributes can be expressed in monetary terms (Orme 2001).

The WTP calculation requires a vector model for the price attribute, which means that these analyses are only meaningful if the price function is indeed linear. The WTP for level m of attribute n can then be derived by dividing the partworth utility for the specific attribute level by the value of the price vector:

$$WTP_{nm} = \frac{\beta_{nm}}{\beta_p} \quad (11)$$

with,

β_{nm} : partworth utility for level m of attribute n

β_p : utility vector for the price attribute

The estimate for the price vector in the ebook reader example is -0.028 , i.e., if price increases by one Euro utility drops by 0.028 units (see Table 6 below). The WTP values for the color attribute can then be calculated as $0.240 / -0.028 = \text{€} -8.57$

Table 6 Estimation results of alternative modeling approaches

Attributes	Partworth model	Vector model for storage and price	Ideal point model for screen size	Interaction effect between screen size and color
Log-likelihood	-2277.8	-2278.3	-2278.3	-2273.3
Storage				
4 GB	-0.389			
8 GB	-0.051			
16 GB	0.440			
(linear)		0.067	0.067	0.067
Screen size				
5 in.	-0.049	-0.050		-0.044
6 in.	0.446	0.446		0.454
7 in.	-0.397	-0.396		-0.410
(linear)			7.854	
(squared)			-0.669	
Color				
Black	-0.002	-0.003	-0.003	-0.015
White	0.240	0.240	0.240	0.255
Silver	-0.238	-0.237	-0.237	-0.240
Price				
€79	0.840			
€99	0.286			
€119	-0.284			
€139	-0.842			
(linear)		-0.028	-0.028	-0.028
No-choice				
	-0.532	-2.965	19.632	-2.984
Screen size × color				
5 in. × black				-0.123
6 in. × black				0.173
7 in. × black				-0.051
5 in. × white				0.018
6 in. × white				-0.164
7 in. × white				0.146

for the color white, €0.07 for black, and €8.50 for silver. The interpretation of these values is that if an ebook reader is not available in, e.g., the preferred color white consumers would accept this drawback only if the price of the reader was, on average, at least €8.57 cheaper. In this case, the negative utility difference of a nonwhite reader is balanced with the positive utility difference of a cheaper price. Vice versa, a consumer would accept paying €8.57 more for a white ebook reader, on average. The least preferred color is silver and consumers would be willing to spend €17.07 for upgrading from a silver ebook reader to a white product. The WTP values can therefore be interpreted directly in terms of consumers' *incremental* willingness to pay for differences in attribute levels. Note, however, that the interpretation needs to consider the differences in signs, i.e., attribute levels with positive utilities have a negative WTP and vice versa.

Market Simulations

The most common ebook readers on the market, e.g., the Amazon Kindle, currently feature 4 GB storage, a 6-in. screen, in the color black for €139. To see how likely it is that consumers buy this product or no ebook reader at all, purchase probabilities can be calculated by applying the MNL function (Eq. 8). These calculations require the specification of a market scenario. A scenario consists of assumptions about the products that are available on the market, i.e., about S , which could include multiple products. In this example, we assume that there are two options, the above-mentioned ebook reader and the no-choice option. On the basis of the aggregate-level estimates, the overall utility of the ebook reader is $V_i = -0.389$ (4 GB) + 0.446 (6 in.) - 0.002 (black) - 0.842 (€139) = -0.787. The utility of the no-choice option is $V_j = -0.532$, i.e., consumers are more likely to buy no ebook reader compared to the one available. The purchase probability for the reader can be calculated by applying Eq. (8):

$$p(i|S) = \frac{\exp(-0.787)}{(\exp(-0.787) + \exp(-0.532))} = 0.437$$

That is, the probability that the sample buys the ebook reader is 43.7%. Market simulations then offer the possibility to see how the market will react if the product configuration is changed. If, e.g., the storage is increased to 8 GB, the overall utility increases to $V_i = -0.449$ and the purchase probability to 0.521. Thus, this modification would be sufficient to make consumers more likely to buy an ebook reader compared to not buying one. Purchase probabilities can be increased further by changing the color to white or reducing the price. These simulations therefore allow detecting promising product modifications. Moreover, a company that wants to enter the market can identify attractive product concepts and assess their effect on purchase probabilities given a specific market scenario that could also consider competitor products. Sophisticated simulation procedures also consider optimal competitive reactions and resulting Nash equilibria (Allenby et al. 2014).

Changing the price in a market simulation, *ceteris paribus*, allows creating a demand function. In the example above, the purchase probability for the ebook

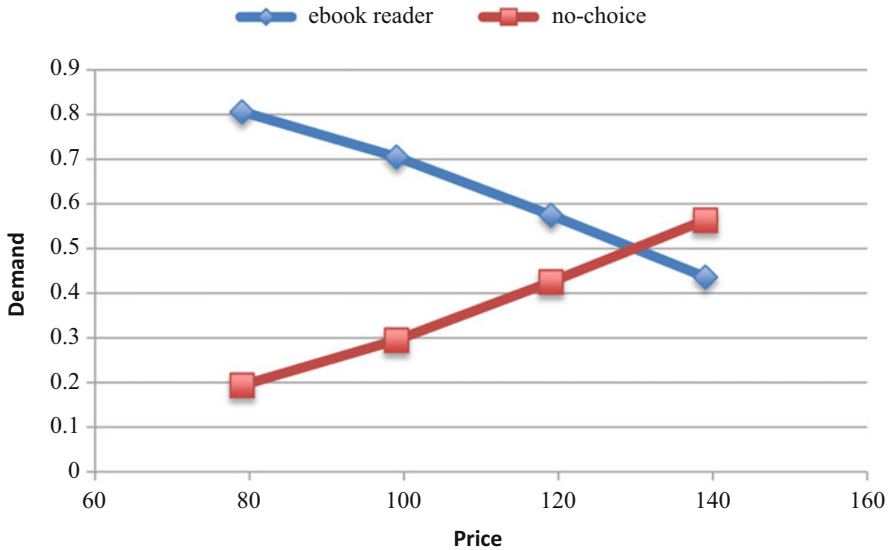


Fig. 8 Demand function for the ebook reader example

reader for €139 is 0.437. Systematically reducing the price increases the probabilities to 0.575 for €119, 0.705 for €99, and 0.806 for €79. The resulting demand function is depicted in Fig. 8. This function can be used to analyze the price elasticity or constitutes an alternative way to calculate WTP. In this example, the price that makes consumers indifferent between choosing the ebook reader and the no-choice option can be taken as the consumers' *absolute* WTP, in this case €130.

The purchase probabilities are frequently interpreted in terms of market shares. Interpreting the predicted probabilities as market shares is ambitious because they have to meet several assumptions (Orme and Johnson 2006). Specifically, probabilities are closer to market shares:

- (a) The more the experiment resembles reality, i.e., all attributes and levels that affect buyers need to be accounted for and all competitors are included in the market scenario (assumptions that are not met in this example).
- (b) The more the real market environment matches the experiment, i.e., all offers are available, e.g., the products are equally distributed, consumers are aware of the available offers, and there are no switching costs between the offers.

Furthermore, predictions are closer the less consumers' choices are influenced by errors that are introduced by the CBC experiment. It has been shown that incentive alignment is a suitable procedure to accomplish more valid answers so that predictions are closer to market shares (Wlömert and Eggers 2016). Moreover, it is often beneficial to consider heterogeneity among consumers via advanced estimation techniques.

Modeling Alternatives

Table 6 depicts the estimation results of alternative modeling approaches. Next to the partworth model interpreted above, it also shows an approach that uses a vector model for storage and price, i.e., that uses their numeric values instead of its effect codes. It can be seen that the model fit changes only marginally as the difference in log-likelihood is only -0.5 , despite using three parameters less. According to a likelihood ratio test this difference is not significant ($p = 0.793$), i.e., this vector model achieves the same fit, while being more parsimonious. The vector model shows that utility increases linearly by 0.067 with every additional GB storage and decreases by -0.028 with every Euro more in purchase price. As the attributes are orthogonal, the other estimates remain largely unaffected. Only the estimate of the no-choice option changes substantially because the numeric values of storage and price are not zero-centered, unlike using effect-coding. This shift does not affect the implications, however.

The third model shown in Table 6 demonstrates the estimation of an ideal-point model for screen size. It requires two parameters, one for the linear effect and one for the squared term. Its model fit is therefore identical to the model in which screen size is represented by a partworth model, which also uses two parameters. The utility for screen size is given by the function $v_{\text{screen size}} = 7.854 * \text{screen size} - 0.669 * \text{screen size}^2$. Accordingly, the ideal point can be calculated as $\partial v / \partial \text{screen size} = 7.854 - 2 * 0.669 * \text{screen size} = 0$, which shows a maximum at 5.87 in.

Finally, the last column of Table 6 adds an interaction effect between the attributes screen size and color. Screen size and color are both represented by two parameters so that $2 * 2$ additional parameters are required. Adding these four parameters significantly increases the model fit ($p = 0.039$), i.e., there is an interaction between these two attributes. Accordingly, consumers prefer a black ebook reader in 6 in. and a white version in 7 in.

Advanced Estimation Techniques

The assumption of aggregate-level analyses that consumers are all identical is usually too restrictive. Considering consumer heterogeneity with advanced estimation techniques is therefore beneficial in reducing the error term. Finite mixture (latent class) procedures assume that the sample consists of distinct segments and estimates different utilities for these segments. Continuous mixture (hierarchical Bayes) models are able to estimate individual-level partworth utilities by assuming that the utilities are drawn from a common distribution, e.g., normal distribution. As a result, partworth utilities are generated for each segment or each individual. These values can subsequently be interpreted analogously to the procedures described in sections “[Relative Attribute Importance](#),” “[Willingness-to-Pay](#),” and “[Market Simulations](#).”

Segment-Level Estimation

Segment-level estimation procedures, i.e., latent class estimation, are assuming that a finite number of (homogeneous) segments can represent the heterogeneity of the respondents in the sample. A segment-level perspective is also in line with discovering market segments with distinct preferences that are an attractive target group for a company's market offerings (i.e., following the segmentation, targeting, and positioning approach).

There are two general approaches for segmentation. The first approach determines segments based on socio-demographic data, e.g., separating males and females and estimating aggregate-level preferences for each of these segments. This a priori segmentation, however, is usually not able to detect segments that reflect systematically different preferences towards the attribute levels. The second approach, i.e., the latent class procedure, aims at finding segments that differ in their choice behavior and estimates segment-specific partworth utilities. These segments are latent, i.e., each respondent belongs to the segments with a certain probability (DeSarbo et al. 1995). If a consumer differs in his/her choice behavior from the partworth utilities of the respective segment, this is reflected by a lower probability to belong to this segment (Teichert 2001b).

Before the estimation starts, the researcher needs to define a specific number of segments. In a first step of an iterative-recursive procedure, the segment-specific partworth utilities for the given number of segments are estimated via maximizing the likelihood function. Afterwards, the utility functions are evaluated given the individual respondent's choices in order to allocate the respondents probabilistically to the segments. This results in posterior probabilities of segment membership based on conditional probabilities according to Bayes' rule (DeSarbo et al. 1995). These calculated probabilities form the basis for the iterative process of re-estimating segment-specific utilities. This loop is repeated until only minor changes in the probabilistic allocation of respondents to segments are observed (Sawtooth 2004).

The iterative-recursive process should be repeated for several numbers of segments. The "optimal" number of segments is not determined by the algorithm and has to be based on information criteria, e.g., AIC, BIC, or CAIC (Wedel and Kamakura 2000; Sawtooth 2004). Moreover, a measure of entropy should be inspected, which reflects the accuracy of the segmentation. It is based on the posterior membership probabilities of the respondents. The entropy can exhibit values in the interval $[0, 1]$ and values close to "1" indicate that the segments are well separated, i.e., respondents can be allocated to one of the segments with almost certainty (DeSarbo et al. 1995).

By weighing the segment-level estimates with the membership probability, individual level estimation can be calculated. However, these values lie in the convex hull of the segment-specific utilities so that it is questionable if they can represent individual-level data well (Wedel et al. 1999). Applying the hierarchical Bayes procedures is more appropriate to estimate individual-level preferences (see next chapter).

Table 7 Segment-level estimates

Attributes	Segment 1	Segment 2	Segment 3
Relative segment size	0.592	0.249	0.158
Storage			
4 GB	-0.323	-0.544	-1.195
8 GB	-0.102	0.122	-0.091
16 GB	0.425	0.422	1.286
Screen size			
5 in.	-0.243	0.815	-0.945
6 in.	0.859	0.011	0.108
7 in.	-0.616	-0.826	0.837
Color			
Black	0.302	-0.593	-0.067
White	-0.240	1.246	0.259
Silver	-0.062	-0.653	-0.192
Price			
€79	1.009	0.920	1.423
€99	0.425	0.411	-0.195
€119	-0.318	-0.250	-0.542
€139	-1.116	-1.081	-0.686
No-choice			
	0.118	-2.383	-0.876

Applying the latent class estimation procedure with three segments to the ebook reader case results in a log-likelihood value of -2056.6 , i.e., an acceptable McFadden's R^2 of 0.258. The entropy value of 0.948 shows a good separation between the segments. The segment-specific partworth utilities are depicted in Table 7 (not showing standard errors and t-values for better readability).

Based on the membership probabilities, segment 1 is the largest segment with about 60% of the respondents. Segment 2 includes a quarter of the sample and segment 3 follows in size with about 15%. As in the aggregate-level case, the estimates for storage and price show face validity for each segment. Moreover, the segmentation is able to discover segments that prefer smaller screens (segment 2) and larger screens (segment 3). The color white is preferred by segments 2 and 3, however, not by segment 1 that prefers black ebook readers. Finally, segment 1 shows a positive value for the no-choice option, which reflects that this segment is more likely to choose no ebook reader compared to the other segments.

Note that in the aggregate-level analysis 6-in. screens and the color white are preferred by the sample. The conclusion to launch this kind of ebook reader would have been suboptimal as none of the segments prefer this product, i.e., segment 1 prefers 6-in. screens but not the color white, and segment 2 and 3 prefer white but smaller or larger screens.

Individual-Level Estimation

An estimation of individual-level partworth utilities with the MNL model is possible with the hierarchical Bayes (HB) procedure. The idea of the procedure is that the aggregate sample is used to determine the distribution of partworth utilities. The distribution then serves as a basis to draw conditional estimates for each individual given the respondent’s choice data. The HB model therefore consists of two coupled layers (Lindley and Smith 1972). The first model layer describes the choice probabilities given the individual partworth utilities, i.e., the MNL model (Eq. 8). The second layer relates the respondents’ partworth utilities to each other by assuming a multivariate (normal) distribution of the utilities with unknown mean (Arora et al. 1998).

The model parameters can then be estimated in an iterative process, e.g., with the Metropolis-Hasting algorithm (Chen et al. 2000). Figure 9 depicts the sequence of the HB procedure.

The researcher first needs to specify the type and parameters of the distribution of the utilities. Based on the distribution and the observed choice data, estimates for the individual partworth utilities are drawn in an iterative recursive process. These utilities, in turn, affect the parameters of the distribution, which then serves as a basis to draw a new set of individual-level partworth utilities in a next iteration. This process runs for a large number of iterations, e.g., 20,000, until the parameters converge. Typically, the first set of individual-level utilities draws is discarded as “burn-in” (Sawtooth 2000). The second set of individual-level draws can be used to make inferences about consumer preferences (Allenby et al. 1995).

Figure 10 shows the distribution of individual-level partworth utilities of the ebook reader dataset as boxplots. The mean and median values are plausible and in

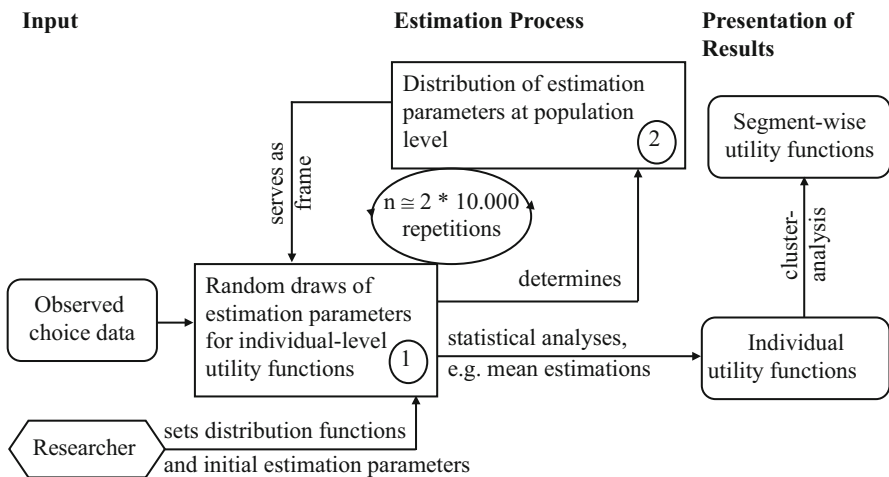
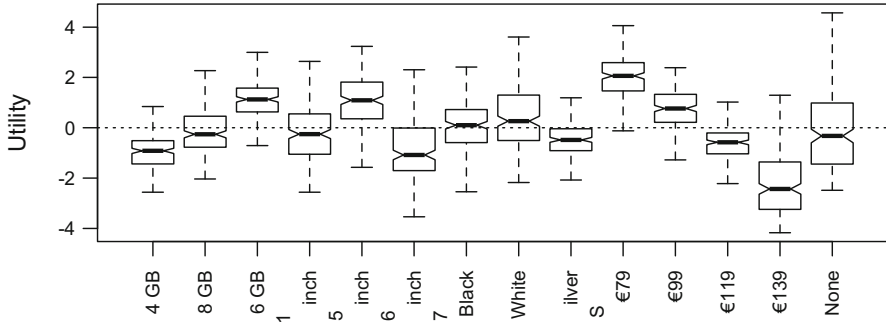


Fig. 9 HB estimation procedures (Teichert 2001b)



	Mean	Standard deviation
Storage		
4 GB	-0.960	0.749
8 GB	-0.173	0.820
16 GB	1.134	0.731
Screen size		
5 inch	-0.187	1.210
6 inch	1.067	0.921
7 inch	-0.880	1.151
Color		
Black	0.026	0.990
White	0.445	1.260
Silver	-0.472	0.668
Price		
€79	2.030	0.804
€99	0.739	0.736
€119	-0.565	0.663
€139	-2.204	1.275
No-choice		
	-0.088	1.593

Fig. 10 Boxplots of partworth utilities and summary statistics for individual-level preferences

line with the aggregate-level model. The distribution and the standard deviation across the respondents' utilities indicate those attributes and attribute levels that exhibit a larger amount of heterogeneous preferences, e.g., screen size, the color white, the highest price, or the no-choice option.

Outlook

Conjoint analysis has emerged from the traditional rating- or ranking-based method in marketing to a general experimental method to study individual's discrete choice behavior with the choice-based conjoint variant. It is therefore not limited to classical applications in marketing, such as new product development, but can be applied to study research questions from related disciplines, e.g., how marketing managers choose their ad campaign, how managers select internationalization options, why consumers engage in or react to social media, etc.

This chapter aims at providing the necessary terminology of conjoint analysis and the requirements to conduct and interpret discrete choice experiments. It also lays the foundation to understand more sophisticated methods and models.

Given the large scope of discrete choice experiments, this attempt is also limited. CBC taps into general theories of how individuals (or groups) choose. These are vast theoretical and empirical grounds, which we cannot cover in detail in this chapter. Understanding CBC models requires not only knowledge of the statistical properties but also understanding behavioral aspects and biases, such as context effects (e.g., compromise, attraction, similarity effects) or trade-off aversion. While knowledge about these aspects is important when running discrete choice experiments, CBC can likewise be used to identify these effects, e.g., by incorporating context effects (Roederkerk et al. 2011) or by measuring price-quality heuristics (Rao and Sattler 2003).

Although CBC is well developed and documented, many areas are still under research, ranging from, e.g., optimal experimental designs, incentive alignment procedures, to estimation techniques. It will therefore remain an active research area with numerous managerial applications in marketing in the future.

Appendix: R Code

The R code and dataset that correspond to the ebook reader example and estimated models can be found at: <http://www.preferencelab.com/data/CBC.R>. The estimation uses the `mlogit` package (Croissant 2012), which needs to be installed first. A less documented version of the R code can be found below (# indicates a comment):

```
# load the library to estimate multinomial choice models.
library(mlogit)

# load (simulated) data about ebook readers
cbc <- read.csv(url("http://www.preferencelab.com/data/
Ebook_Reader.csv"))

# convert data for mlogit
cbc <- mlogit.data(cbc, choice="Selected", shape="long", alt.
var="Alt_id", id.var = "Resp_id")
```



```
### calculate models ###

### partworth model ###
m11 <- mlogit(Selected ~ Storage_4GB + Storage_8GB +
  Screen.size_5inch + Screen.size_6inch +
  Color_black + Color_white +
  Price_79 + Price_99 + Price_119 +
  None | 0, cbc)
summary(m11)

# recover reference level estimates (effect-coding)

# Storage_16GB
-(coef(m11) ["Storage_4GB"] + coef(m11) ["Storage_8GB"])

# Screen.size_7inch
-(coef(m11) ["Screen.size_5inch"] + coef(m11) ["Screen.size_6inch"])

# Color_silver
-(coef(m11) ["Color_black"] + coef(m11) ["Color_white"])

# Price_139
-(coef(m11) ["Price_79"] + coef(m11) ["Price_99"] + coef(m11)
["Price_119"])

# standard errors of the effects are given by the
# square root of the diagonal elements of the
# variance-covariance matrix
covMatrix <- vcov(m11)
sqrt(diag(covMatrix))

# with effect-coding, the standard error of the reference
# level needs to consider the off-diagonal elements of the
# corresponding attribute levels

# Std. Error Storage_16GB
sqrt(sum(covMatrix[1:2, 1:2]))

# Std. Error Screen.size_7inch
sqrt(sum(covMatrix[3:4, 3:4]))

# Std. Error Color_silver
sqrt(sum(covMatrix[5:6, 5:6]))

# Std. Error Price_139
sqrt(sum(covMatrix[7:9, 7:9]))
```

```
### Vector model ###
# Storage and Price follow a linear trend. Replacing
# parameters leads to a more parsimonious model.

m12 <- mlogit(Selected ~ Storage +
  Screen.size_5inch + Screen.size_6inch +
  Color_black + Color_white +
  Price +
  None | 0, cbc)
summary(m12)

# likelihood ratio test
lrtest(m12, m11)

# incremental willingness-to-pay for storage
coef(m12)["Storage"]/coef(m12)["Price"]

# WTP to upgrade from a black to a white ebook reader
(coef(m12)["Color_white"] - coef(m12)["Color_black"])/coef(m12)
["Price"]

### Vector model for screen size has sig. worse fit ###
m13 <- mlogit(Selected ~ Storage + Screen.size + Color_black +
  Color_white + Price + None | 0, cbc)
summary(m13)

lrtest(m13, m12)

### Testing an ideal point model for screen size ###
m14 <- mlogit(Selected ~ Storage +
  Screen.size + I(Screen.size**2) +
  Color_black + Color_white +
  Price +
  None | 0, cbc)
summary(m14)

# same model fit because no differences in df
lrtest(m14, m12)

### Adding interactions between screen size and color ###
m15 <- mlogit(Selected ~ Storage +
  Screen.size_5inch + Screen.size_6inch +
  Color_black + Color_white +
  Price +
  Screen.size_5inch * Color_black +
  Screen.size_6inch * Color_black +
```

```

Screen.size_5inch * Color_white +
Screen.size_6inch * Color_white +
None| 0, cbc)
summary(ml5)

# likelihood ratio test
lrtest(ml2, ml5)

```

References

- Addelman, S. (1962). Orthogonal main-effect plans for asymmetrical factorial experiments. *Technometrics*, 4(1), 21–46.
- Allenby, G. M., Arora, N., & Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32(2), 152–162.
- Allenby, G. M., Brazell, J. D., Howell, J. R., & Rossi, P. E. (2014). Economic valuation of product features. *Quantitative Marketing and Economics*, 12(4), 421–456.
- American Marketing Association. (2015). American Marketing Association AMA. <https://www.ama.org/resources/Pages/Dictionary.aspx>. Accessed 15 Nov 2015.
- Arora, N., Allenby, G. M., & Ginter, J. L. (1998). A hierarchical Bayes model of primary and secondary demand. *Marketing Science*, 17(1), 29–44.
- Batsell, R. R., & Louviere, J. J. (1991). Experimental analysis of choice. *Marketing Letters*, 2(3), 199–214.
- Bauer, H., Herrmann, A., & Homberg, F. (1996). *Analyse der Kundenwünsche zur Gestaltung eines Gebrauchsgutes mit Hilfe der Conjoint Analyse*. Universität Mannheim, Lehrstuhl für ABWL und Marketing II, Working Paper Nr. 110.
- Becker, G. M., Degroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3), 226–232.
- Brazell, J. D., Diener, C. G., Karniouchina, E., Moore, W. L., Séverin, V., & Uldry, P.-F. (2006). The no-choice option and dual response choice designs. *Marketing Letters*, 17(4), 255–268.
- Burmester, A., Eggers, F., Clement, M., & Prostka, T. (2016). Accepting or fighting unlicensed usage – Can firms reduce unlicensed usage by optimizing their timing and pricing strategies? *International Journal of Research in Marketing*, 33(2), 434–356.
- Chakraborty, G., Ball, D., Gaeth, G. J., & Jun, S. (2002). The ability of ratings and choice conjoint to predict market shares – A Monte Carlo simulation. *Journal of Business Research*, 55(3), 237–249.
- Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. New York: Springer Series in Statistics.
- Croissant, Y. (2012). Estimation of multinomial logit models in R: The mlogit packages. *R package version 0.2-2*. <http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf>.
- De Bekker-Grob, E. W., Ryan, M., & Gerard, K. (2012). Discrete choice experiments in the health economics: A review of the literature. *Health Economics*, 21(2), 145–172.
- DeSarbo, W. S., Ramaswamy, V., & Cohen, S. (1995). Market segmentation with choice-based conjoint analysis. *Marketing Letters*, 6(2), 137–147.
- Ding, M. (2007). An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research*, 44(2), 214–223.
- Ding, M., Grewal, R., & Liechty, J. (2005). Incentive-aligned conjoint analysis. *Journal of Marketing Research*, 42(2), 67–82.
- Ding, M., Park, Y.-H., & Bradlow, E. T. (2009). Barter markets for conjoint analysis. *Management Science*, 55(6), 1003–1017.

- Dong, S., Ding, M., & Huber, J. (2010). A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing*, 27(1), 25–32.
- Eggers, F., & Sattler, H. (2009). Hybrid individualized two-level choice-based conjoint (HIT-CBC): A new method for measuring preference structures with many attribute levels. *International Journal of Research in Marketing*, 26(2), 108–118.
- Eggers, F., Hauser, J. R., & Selove, M. (2016). The effects of incentive alignment, realistic images, video instructions, and ceteris paribus instructions on willingness to pay and price equilibria. *Proceedings of the Sawtooth Software conference*, 1–18 September.
- Elrod, T., Louviere, J. J., & Davey, K. S. (1992). An empirical comparison of ratings-based and choice-based conjoint models. *Journal of Marketing Research*, 29(3), 368–377.
- Frischknecht, B., Eckert, C., Geweke, J., & Louviere, J. J. (2014). A simple method for estimating preference parameters for individuals. *International Journal of Research in Marketing*, 31(1), 35–48.
- Gensler, S., Hinz, O., Skiera, B., & Theysohn, S. (2012). Willingness-to-pay estimation with choice-based conjoint analysis: Addressing extreme response behavior with individually adapted designs. *European Journal of Operational Research*, 219(2), 368–378.
- Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5, 103–123.
- Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54, 3–19.
- Haaijer, R., & Wedel, M. (2003). Conjoint experiments. general characteristics and alternative model specifications. In A. Gustafsson, A. Herrmann, & F. Huber (Eds.), *Conjoint measurement: Methods and applications* (3rd ed., pp. 371–412). Berlin: Springer.
- Haaijer, R., Wedel, M., Vriens, M., & Wansbek, T. (1998). Utility covariances and context effects in conjoint MNP models. *Marketing Science*, 17(3), 236–252.
- Haaijer, R., Kamakura, W. A., & Wedel, M. (2001). The “no-choice” alternative to conjoint choice experiments. *International Journal of Market Research*, 43(1), 93–106.
- Hartmann, A. (2004). *Kaufentscheidungsprognose auf Basis von Befragungen. Modelle, Verfahren und Beurteilungskriterien*. Wiesbaden: Gabler.
- Hensher, D. A. (1994). Stated preference analysis of travel choices: The state of practice. *Transportation*, 21(2), 107–133.
- Hensher, D. A., & Johnson, L. W. (1981). *Applied discrete choice modelling*. New York: Wiley.
- Huber, J., & Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33(3), 307–317.
- Johnson, R. M. (1987). Adaptive conjoint analysis. In *Sawtooth software conference proceedings*. Ketchum: Sawtooth Software.
- Johnson, R. M., & Orme, B. K. (1996). *How many questions should you ask in choice-based conjoint studies?* (Sawtooth software research paper series). Sequim: Sawtooth Software.
- Kraus, S., Ambos, T. C., Eggers, F., & Cesinger, B. (2015). Distance and perceptions of risk in internationalization decisions. *Journal of Business Research*, 68(7), 1501–1505.
- Lindley, D. V., & Smith, A. F. (1972). Bayes estimates for the linear models. *Journal of the Royal Statistical Society, Series B*, 34(1), 1–41.
- Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments. An approach based on aggregated data. *Journal of Marketing Research*, 20(4), 350–367.
- Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods. Analysis and application*. Cambridge: Cambridge University Press.
- Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods, and applications*. Cambridge: Cambridge University Press.
- Lusk, J. L., & Schroeder, T. C. (2004). Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics*, 86(2), 467–482.
- McFadden, D. (1981). Econometric models of probabilistic choice. In C. Manski & D. McFadden (Eds.), *Structural analysis of discrete data* (pp. 198–272). Cambridge: MIT-Press.

- Meissner, M., Oppewal, H., & Huber, J. (2016). How many options? Behavioral responses to two versus five alternatives per choice. *Proceedings of the Sawtooth Software conference*, 1–18 September.
- Miller, K. M., Hofstetter, R., Krohmer, H., & Zhang, Z. J. (2011). How should Consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches. *Journal of Marketing Research*, 48(1), 172–184.
- Moore, W. L. (2004). A cross-validity comparison of rating-based and choice-based conjoint analysis models. *International Journal of Research in Marketing*, 21(3), 299–312.
- Moore, W. L., Gray-Lee, J., & Louviere, J. J. (1998). A cross-validity comparison of conjoint analysis and choice models at different levels of aggregation. *Marketing Letters*, 9(2), 195–207.
- Orme, B. (2001). *Assessing the monetary value of attribute levels with conjoint analysis: Warnings and suggestions* (Sawtooth software research paper series). Sequim: Sawtooth Software.
- Orme, B. (2002). *Formulating attributes and levels in conjoint analysis* (Sawtooth software research paper series). Sequim: Sawtooth Software.
- Orme, B. K. (2016). *Results of the 2017 Sawtooth Software User Survey*. <https://www.sawtoothsoftware.com/about-us/news-and-events/news/1693-results-of-2016-sawtooth-software-user-survey>.
- Orme, B., & Johnson, R.M. (2006). *External effect adjustments in conjoint analysis* (Sawtooth software research paper series). Sequim: Sawtooth Software.
- Page, A. L., & Rosenbaum, H. F. (1992). Developing an effective concept testing program for consumer durables. *Journal of Product Innovation Management*, 9, 267–277.
- Park, Y.-H., Ding, M., & Rao, V. R. (2008). Eliciting preference for complex products: A web-based upgrading method. *Journal of Marketing Research*, 45(5), 562–574.
- Rao, V. R., & Sattler, H. (2003). Measurement of price effects with conjoint analysis: Separating informational and allocative effects of price. In *Conjoint Measurement* (pp. 47–66). Berlin/Heidelberg: Springer.
- Roederkerk, R. P., Van Heerde, H. J., & Bijmolt, T. H. (2011). Incorporating context effects into a choice model. *Journal of Marketing Research*, 48(4), 767–780.
- Sattler, H. (2005). Markenbewertung: State-of-the-Art. *Zeitschrift für Betriebswirtschaft*, 2, 33–57.
- Sattler, H. (2006). Methoden zur Messung von Präferenzen für Innovationen. *Zeitschrift für Betriebswirtschaftliche Forschung*, 54(6), 154–176.
- Sattler, H., Hartmann, A., & Kröger, S. (2004). Number of tasks in choice-based conjoint analysis. *Conference proceedings of the 33rd EMAC conference*. Murcia.
- Sawtooth (1999). *The choice-based conjoint (CBC) technical paper* (Sawtooth software technical paper series). Sequim: Sawtooth Software.
- Sawtooth. (2000). *The CBC/HB system for hierarchical Bayes estimation version 4.0* (Sawtooth software technical paper series). Sequim: Sawtooth Software.
- Sawtooth. (2004). *The CBC latent class technical paper (version 3)* (Sawtooth software technical paper series). Sequim: Sawtooth Software.
- Sawtooth. (2013). *The MaxDiff system – Technical paper* (Sawtooth software technical paper series). Orem: Sawtooth Software.
- Sawtooth. (2014). *ACBC – Technical paper* (Sawtooth software technical paper series). Orem: Sawtooth Software.
- Schlereth, C., & Skiera, B. (2016). Two new features in discrete choice experiments to improve willingness-to-pay estimation that result in SDR and SADR: Separated (adaptive) dual response. *Management Science*, 63(3), 829–842.
- Shocker, A. D., & Srinivasan, V. (1973). Linear programming techniques for multidimensional analysis of preference. *Psychometrika*, 337–369.
- Sloan, N. J. A. (2015). A library of orthogonal arrays. <http://neilsloane.com/oadir/>. Accessed 15 Nov 2015.
- Srinivasan, V., & Park, C. S. (1997). Surprising robustness of the self-explicated approach to customer preference structure measurement. *Journal of Marketing Research*, 34(2), 286–291.

- Teichert, T. (2001a). *Nutzenschätzung in Conjoint-Analysen: Theoretische Fundierung und empirische Aussagekraft*. Wiesbaden: Springer.
- Teichert, T. (2001b). Nutzenermittlung in wahlbasierten Conjoint-Analysen. Ein Vergleich zwischen Latent-Class- und hierarchischem Bayes-Verfahren. *Zeitschrift für Betriebswirtschaftliche Forschung*, 53(8), 798–822.
- Toubia, O., Simester, D. I., Hauser, J. R., & Dahan, E. (2003). Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3), 273–303.
- Toubia, O., Hauser, J. R., & Simester, D. I. (2004). Polyhedral methods for adaptive choice-based conjoint analysis. *Journal of Marketing Research*, 41, 116–131.
- Toubia, O., Hauser, J., & Garcia, R. (2007). Probabilistic polyhedral methods for adaptive choice-based conjoint analysis: Theory and application. *Marketing Science*, 26(5), 596–610.
- Toubia, O., de Jong, M. G., Stieger, D., & Füller, J. (2012). Measuring consumer preferences using conjoint poker. *Marketing Science*, 31(1), 138–156.
- Train, K. (2009). *Discrete choice models with simulation* (2nd ed.). Cambridge: Cambridge University Press.
- Urban, G. L., & Hauser, J. R. (1993). *Design and marketing of new products* (2nd ed.). Englewood Cliffs: Prentice Hall.
- Urban, G. L., Weinberg, B. D., & Hauser, J. R. (1996). Premarket forecasting of really-new products. *Journal of Marketing*, 60(1), 47–60.
- Verlegh, P. W. J., Schifferstein, H. N. J., & Wittink, D. R. (2002). Range and number-of-levels in derived and stated measures of attribute importance. *Marketing Letters*, 13(1), 41–52.
- Voeth, M. (1999). 25 Jahre conjointanalytische Forschung in Deutschland. *Zeitschrift für Betriebswirtschaft*, Ergänzungsheft 2, 153–176.
- Vriens, M., Oppewal, H., & Wedel, M. (1998). Rating-based versus choice-based latent class conjoint models – An empirical comparison. *Journal of the Market Research Society*, 40(3), 237–248.
- Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3), 303–343.
- Wedel, M., & Kamakura, W. A. (2000). *Market segmentation. conceptual and methodological foundations* (2nd ed.). Boston: Springer.
- Wedel, M., Kamakura, W. A., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., Johnson, R. M., Lenk, P., Neslin, S., & Poulsen, C. S. (1999). Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters*, 10(3), 219–232.
- Wertenbroch, K., & Skiera, B. (2002). Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research*, 39(2), 228–241.
- Wittink, D. R., Vriens, M., & Burhenne, W. (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, 11, 41–52.
- Wlömert, N., & Eggers, F. (2016). Predicting new service adoption with conjoint analysis: External validity of BDM-based incentive-aligned and dual-response choice designs. *Marketing Letters*, 27(1), 195–210.
- Zeithammer, R., & Lenk, P. (2009). Statistical benefits of choices from subsets. *Journal of Marketing Research*, 46(6), 816–831.



Exploiting Data from Field Experiments

Martin Artz and Hannes Doering

Contents

Introduction	822
Motivation	822
Field Experiments	825
Difference-in-Differences Method	828
Introduction	828
Core Area of Application	830
Critical Assumptions	832
Application in Goldfarb and Tucker (2011)	834
Regression Discontinuity Designs	835
Introduction	835
Core Area of Application	838
Critical Assumptions	839
Application in Flammer (2015)	840
Instrumental Variables	841
Introduction	841
Core Area of Application	845
Critical Assumptions	846
Application in Bennedsen et al. (2007)	848
Application of Methods in Standard Software	849
Conclusions	853
Cross-References	854
References	854

Abstract

This chapter gives an introduction on how to exploit data from field experiments and aims to provide an intuitive understanding for managers and researchers alike. We outline the relevance and hurdles in identifying causal effects compared to observing purely correlational associations in studies which take place in the

M. Artz (✉) · H. Doering
School of Business and Economics, University of Münster, Münster, Germany
e-mail: martin.artz@wiwi.uni-muenster.de; hannes.doering@wiwi.uni-muenster.de

real world. We further provide a framework to classify different kinds of field experiments, such as quasi field experiments and natural field experiments. The core of this chapter focuses on giving an understanding of three standard econometric methods to exploit data from field experiments: difference-in-differences, regression discontinuity, and instrumental variables. For each method, we provide an intuitive understanding of the core features and its critical assumptions. We complement those explanations with an in-depth look at one practical application of each method in a field experiment setting and with a variety of practical examples from recently published research. Lastly, we provide a brief overview on how to implement each method in standard software packages such as STATA, R, and SPSS.

Keywords

Field experiment · Quasi experiment · Natural experiment · Causality · Causal inference · Difference-in-differences · Regression discontinuity · Instrumental variable

Introduction

“There is two ways to get fired from Harrah’s: Stealing from the company or failing to include a proper control group in your business experiment.” (Gary Loveman, economist and former CEO of Harrah’s Entertainment)

Motivation

Suppose a fashion retailer decides to place a specific jeans brand right at the stores’ entrance in order to temporarily increase in-store sales. For the weeks following the rearrangement, the retailer observes an increase in sales for this jeans brand. Clearly, this entrance placement co-moves with an increase in sales. Nevertheless, the question remains whether the rearrangement has caused the sales to increase or whether the increase is due to a long list of alternative reasons. For instance, in the weeks after the rearrangement, customer preferences for this particular brand could have changed or competitors offering the same brand could have increased prices. A co-movement between the two factors – change in placement and increase in sales – is easy to observe, but claiming causality between those two factors is difficult.

Suppose now, the fashion retailer offers free shipping for all orders. Subsequently, the amount of orders and the value per order in the fashion retailer’s online store increase. Again, some action (i.e., the free shipping offer) is linked to some outcome. And again, one has to be cautious to credibly claim that this action indeed caused the outcome. Maybe the retailer simultaneously launched an advertising campaign, a competitor filed for bankruptcy or some of the brick-and-mortar stores were temporarily closed for renovations. All those events may interfere with the free shipping

offer and also potentially co-move with sales. So, top management of this retail chain will once again raise the question: Which of those events caused the increase in sales? Has it been particularly one of these actions, the interplay of various actions, or some (unobserved) events which ultimately caused the increase in online sales?

As these rather simple examples demonstrate, managers and researchers are typically interested in causal relationships to improve the quality of business decisions. “*However, data reveal only associations, which are a combination of causal and non-causal (i.e., spurious) components*” (Keele 2015, p. 102). In order to separate those two components and establish a clear causal path between an action and a particular observable outcome, three general conditions have to be fulfilled (Kenny 1979):

- The cause has to precede the outcome, that is, the cause must occur before the effect temporarily. In the fashion retailer examples, the change in placement of jeans in the store or the free shipping offer occurred before any increase in sales had been observed.
- Cause and effect have to co-move, that is, changes in the cause must be accompanied by changes in the effect. In the fashion retailer examples, this has been the case. After each initiative (i.e., the change in placement or the free shipping initiative), the retailer observed a co-movement between the initiative and sales. Standard measures often used in business practice and in academic research are contingency tables or various forms of correlation coefficients such as Pearson or Spearman (Spearman 1904). These statistical performance measures suggest a weaker or stronger existence of a (linear) co-movement between two variables as well as whether the co-movement is negative or positive.
- The relation between presumed cause and effect cannot be explained by alternative reasons. Stated differently, the new placement or the free shipping initiative is supposed to be the sole driver of any observed sales outcome. However, excluding alternative explanations with certainty has demonstrated to be difficult in both fashion retailer examples and is indeed often the most difficult hurdle to establish causality. A potential way out of this dilemma is drawing conclusions from data generated via a controlled experimental design. Since such data is generated under controlled conditions, external influences can be fully eliminated via the design of the experiment.

In management practice, *field* experiments – experiments that take place in real-world environments instead of laboratory settings – have gained much importance in the last years. As an illustration, the solid line in Fig. 1 depicts the number of search results for the term “field experiment” for the abstract of academic, peer-reviewed business publications for the past 20 years. In practice, online business models are able to vary treatments (i.e., design of marketing campaigns, product offers, or price discounts) between randomized customer groups (i.e., A/B testing) and offline business models try new initiatives in some business units before rolling them out through the whole corporation (i.e., pilot studies). In any case, an experiment requires two stages: a design stage before the experiment is implemented, including,

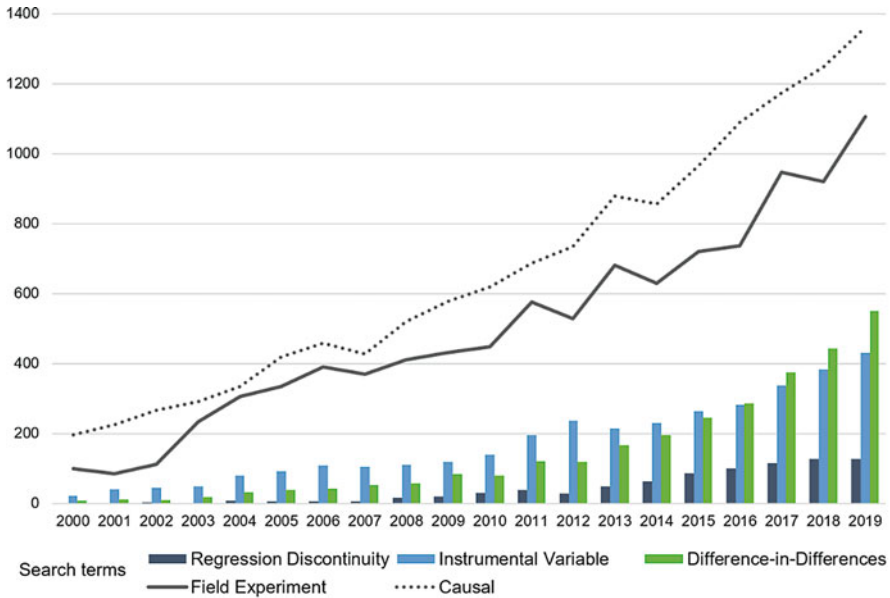


Fig. 1 Yearly number of academic, peer-reviewed publications in business and economics (via *EBSCOhost Business Source Premier*) for which the respective search term is mentioned in the article’s abstract (own illustration)

for instance, the selection of the initiative as well as the number and randomization of groups, and an examination stage after the experiment took place to analyze the data which had been generated.

This chapter focuses on the second part. Its objective is to present, describe, and explain different methods to analyze data generated by field experiments, thus focusing on the examination stage after the experiment took place. Further, it follows the approach of giving the interested reader an intuitive understanding of the standard methods to analyze data from field experiments, often at the cost of a detailed statistical discussion. We refer the reader to Angrist and Pischke (2009), Antonakis et al. (2010), Verbeek (2008), and Wooldridge (2012) for a more detailed discussion of statistical and econometric methods. Further, each section will provide literature relevant to each method for the interested reader. For a detailed introduction on how to design and execute a (field) experiment, we refer to Bornemann and Hattula (chapter ► “[Experiments in Market Research](#)”) as well as Valli et al. (chapter ► “[Field Experiments](#)”) in this handbook.

The chapter is structured as follows: Section two clarifies the term field experiment. Sections three, four, and five refer to three core methodologies to analyze data generated by a field experiment such as the difference-in-differences methodology (section “[Difference-in-Differences Method](#)”), a regression discontinuity design (section “[Regression Discontinuity Designs](#)”), and an instrumental variable approach (section “[Instrumental Variables](#)”). Section “[Application of Methods in Standard Software](#)” provides an overview of how these three methods are

implemented in standard software packages such as STATA, R, and SPSS and provides a selection of data sets that are suitable for one own's replication efforts for each method. The last section concludes this chapter.

Field Experiments

The understanding and interpretation of what constitutes a field experiment and where to draw the line to similar and more distinct research designs widely varies among researchers and practitioners. Among all groups, it seems universally accepted that an experiment represents a study design which requires at minimum two different groups that are equal along each characteristic except for the fact that one gets access to a treatment (i.e., the treatment group), while the other does not (i.e., the control group). Often, this treatment is also called an intervention or manipulation in an experiment. Any observed outcome differences are only due to this intervention, guaranteed via a rigorous design that fulfills the conditions of causality.

Turning to the more specific case of *field* experiments, definitions and interpretations of the term vary. Harrison and List (2004) interpret the term *field* as any experimental intervention which includes a treatment that is related to the real world, independent of whether the experiment takes place inside or outside of the laboratory. In their taxonomy, for example, inviting managers to perform job-related tasks in a laboratory would constitute a field experiment. Other authors refer to field experiments as experiments that take place outside the laboratory environment. For instance, Lourenço (2019) defines field experiments as experiments which take place in the natural environment of the subjects, where researchers are in control of the random assignment of treatment and where subjects “*are not aware that they are part of an experiment*” (Lourenço 2019, p. 2). Contrarily, Harrison and List (2004) would refer to those experiments as “*natural field experiments*” (Harrison and List 2004, p. 1014). Notably, these definitions also include field experiments that are not designed and implemented by firms, but by other institutions (e.g., regulatory or governmental institutions), or occur as natural events (e.g., extreme weather events).

Beyond these differences in academic contributions, practitioners use additional terms to describe certain types of field experiments in different settings and business models. The term *A/B testing* is often used when referring to field experiments in online environments, meaning that, for example, one randomly selected group of customers is presented with packaging design A and another group is presented with packaging design B for an otherwise identical product (Goldfarb and Tucker 2014). In traditional market research, the term *pilot test* describes a setting where a product is only offered to a selected group of customers, often in one store whose customer base reflects a representative set of all customers.

Given these differences in terminology and definitions, we suggest a broader classification including all possible types of field experiments conducted in business practice and research. Figure 2 characterizes four different types of field experiments in a two-by-two matrix along the dimensions of *Control over*

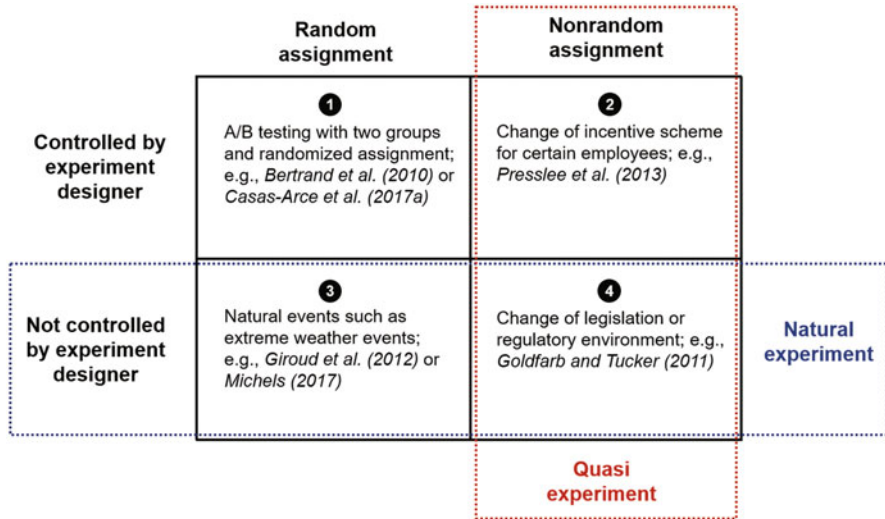


Fig. 2 Classification of field experiments (own illustration)

Treatment and Randomness of Assignment borrowing the terminology of Shadish et al. (2002):

- *Control over Treatment* describes the researcher’s ability to design and have full control over the type of treatment and the selection of subjects and circumstances. Here, we differentiate between controlled and not controlled by experiment designer. In the fashion retailer example, management is able to decide where and when to place the jeans brand and whether other promotions are conducted simultaneously. However, sometimes the treatment is outside of the designer’s control and is either determined by exogenous conditions (i.e., natural weather events) or by executives at higher decision-making levels (i.e., regulatory or governmental institutions). These exogenous interventions are often simply called “shocks” (Atanasov and Black 2016). We classify such designs as natural field experiments.
- *Randomness of Assignment* describes whether the experimental design classifies subjects randomly (or not) to the treatment and the control group. Here, we differentiate between random and nonrandom assignment. In the fashion retailer example, the firm could randomly offer the free shipping option to some online shop visitors (i.e., the treatment group) but not to others (i.e., the control group). Although a nonrandom allocation may not represent a rigorous experimental design, this type of experiment plays a significant role in research and practice (Shadish et al. 2002). We classify such designs as quasi field experiments.

Along those two dimensions, we differentiate four types of field experiments:

- Quadrant 1 (controlled by designer, random assignment) characterizes field experiments which satisfy the definition of Lourenço (2019). Here, the designer

is able to select the treatment and can randomize this treatment across subjects. An example for such a field experiment is Casas-Arce et al. (2017a). For this field experiment, the authors collaborate with an insurance company and randomize the frequency and level of detail of customer satisfaction reports provided to independent contractors. Consequently, they can observe and measure performance improvements for the different contractor groups and report a causal link between the frequency and detail of reports and performance. Another example is Bertrand et al. (2010) who partner with a consumer lender in South Africa to randomly vary different aspects of advertising content to existing customers. The authors can clearly identify that particular aspects of the advertising content affect lending demand of various customer groups.

- Quadrant 2 (controlled by designer, nonrandom assignment) constitutes experiments in which the designer can select the treatment but the assignment of the treatment to subjects itself is not random. Often, the collaborating partner (i.e., a firm or regulatory body) does not allow for true random assignment as it would be favored to establish causality. For example, Presslee et al. (2013) are able to select different cash and intangible rewards given to employees of a call center for a limited time span. The authors observe the employees' behavior with regard to chosen goal difficulty, goal commitment, and, ultimately, performance. Nevertheless, the call center firm decided which group of employees received which treatment and did so on the basis of the geographic location of its call centers. Due to the lack of complete randomization, Presslee et al. (2013) point out that their setup represents a quasi field experiment.
- Quadrant 3 (not controlled by designer, random assignment) consists of those experiments for which the treatment is assigned randomly but not at the designer's control, that is, the designer does not administer the random assignment by herself. Rather, the random treatment is occurring due to exogenous causes such as extreme weather events. Those events may be floods, wild fires, hurricanes, or heavy snowfall affecting one group of subjects differently than others. For example, Giroud et al. (2012) exploit unexpected snowfall in the Austrian Alps for their natural experiment in the domain of finance research, and Michels (2017) exploits events of floods and hurricanes in the domain of financial accounting research. In a marketing context, Shriver et al. (2013) use the plausibly random variation in wind speeds on Swiss surf spots to explain the generation of content in an online social network for wind surfers. In general, the pure randomness of weather events may be discussed for some instances, either with regard to self-selection into areas of particular weather conditions or with regard to systematic climate changes; this discussion is yet beyond the scope of this more general introduction.
- Finally, quadrant 4 (not controlled by designer, nonrandom assignment) consists of those experiments in which the designer neither is in full control of the treatment nor is the treatment randomly assigned to subjects. Both limitations with regard to a rigorous experimental design may be present because the collaborating partner is either not willing or not able to give up full control. Especially regulatory bodies such as the European Central Bank or national tax authorities have their own, distinct statutes and agendas when deciding for

specific policy changes and are hesitant to implement policies “at random.” This strongly limits the designer’s ability to conduct an ideal field experiment and limits the opportunity to study certain questions such as the change in business tax rates across European firms or spikes in prime lending rates for financial institutions at various points in time. Nevertheless, some interesting research has been conducted around the introduction of or changes in legislation or regulation. For example, Goldfarb and Tucker (2011) exploit the change in privacy regulation in the European Union as a regulatory shock to advertisers and their respective possibility to target customers with advertisement. In this quasi-natural field experiment, the authors are strongly concerned with the limits of their research setting, consequently adding sophisticated statistical analyses to derive causal statements from their results. Thinking in terms of distance from the ideal field experiment, studies in this quadrant are certainly those where approaching causality requires additional (and often complex) statistical analyses and further, partially untestable assumptions to mitigate concerns that other factors might be responsible for the observed effects.

As already indicated, some researchers argue that natural experiments (quadrants three and four) are not rigorous experiments since the designer does not have full control which constitutes a violation of one of the key components of an experiment (Lourenço 2019). Often, this kind of shock-based research relies on an event to happen and requires the event to affect only a certain group of individuals without the primary purpose of exploiting this manipulation for research (Atanasov and Black 2016). Typically, violations of randomization are addressed via supplementary statistical analyses and by carefully considering any interfering factor in order to establish causality. However, despite these concerns, the term experiment is usually not disputed. Conclusively, what distinguishes a field experiment from any other study (potentially also conducted “in the field”) is its clear variation through a (quasi) exogenous treatment allowing to derive a causal treatment effect.

The following sections deal with methods of how to analyze data generated by all types of field experiments classified in Fig. 2. Since these methods do not distinguish between the data generating process, we discuss methods that are applicable to all kinds of field experiments. However, it will become apparent that some methods are more suitable for some particular designs. In particular, we now refer to the difference-in-differences method (section “[Difference-in-Differences Method](#)”), regression discontinuity designs (section “[Regression Discontinuity Designs](#)”), and instrumental variable approach (section “[Instrumental Variables](#)”).

Difference-in-Differences Method

Introduction

In any study, no matter whether it is conducted in the laboratory or in the field, the researcher compares the observable outcome of the treatment group against the outcome of a control group, that is, the group which is (ideally) identical to the

treatment group but does not receive the treatment. As long as treatment and control groups are equal along all characteristics (i.e., when the researcher was able to fully randomize the treatment on a homogeneous group as in quadrant 1), a simple comparison of average outcomes between the two groups is sufficient to derive the average treatment effect (ATE). Unfortunately, for many field experiments, an ideal control group does not exist, is impossible or too costly to observe, or its design would violate ethical or legal boundaries.

To exploit data from field experiments which lack the perfect control group, the difference-in-differences method relaxes the requirements for the control group from being “practically identical” to “showing the same trend,” called “parallel” or “common” trends assumption (Antonakis et al. 2010, pp. 1108–1109). In this setup, the control group is not equal to the treatment group along all possible variables but exhibits the same trend over time along the relevant dimensions prior to the treatment (Angrist and Pischke 2009, pp. 169–172). Thus, levels may be different but the distance between those levels stays constant over time. For example, two customer groups exhibit different levels of product purchases (e.g., in terms of order value) but over time both groups’ spending increases at the same rate so that the difference in levels remains equal.

Figure 3 illustrates how exploiting this requirement allows to draw causal inferences from nonequal groups within the difference-in-differences design. Before the intervention, the treatment group and the control group are not on the same level (e.g., in terms of number of products purchased or order value) but follow the same trend (i.e., order value increases over time at the same rate for both groups). After the treatment, the level of order value continues to increase at the same rate for the control group. Yet, for the treatment group, the trend has changed and the rate at which the order value increases is greater. In this setup, the desired difference

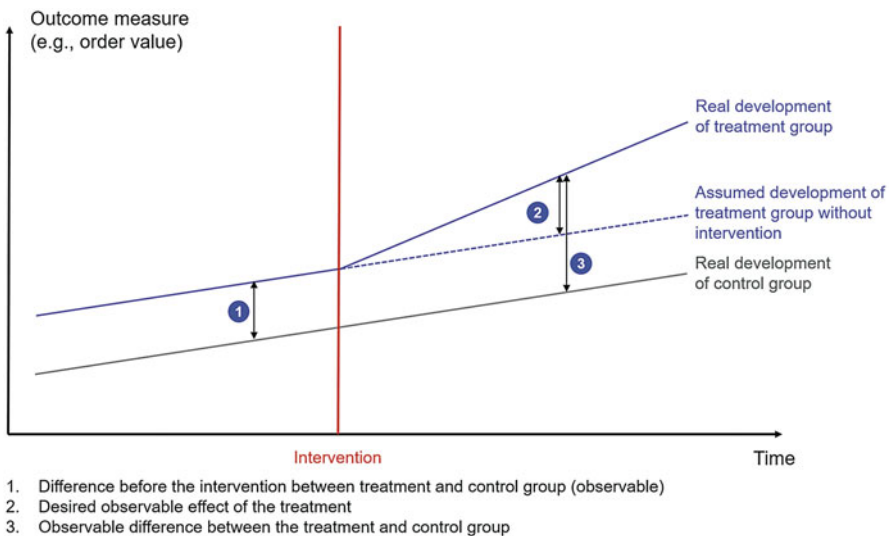


Fig. 3 Graphical illustration of difference-in-differences approach (own illustration)

Stylized difference-in-differences regression
 Outcome = $\beta_0 + \beta_1 * \text{Treated} + \beta_2 * \text{Post} + \beta_3 * \text{Treated} * \text{Post} + \epsilon$

Isolation of the treatment effect			
	Post = 0	Post = 1	Intertemporal Difference
Treated = 0	β_0	$\beta_0 + \beta_2$	β_2
Treated = 1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Cross-Sectional Difference	β_1	$\beta_1 + \beta_3$	β_3

Fig. 4 Isolation of the treatment effect in a stylized difference-in-differences regression (adapted from Wooldridge 2012)

(2) is derived by observing the difference in order value between the two groups after the change (3) and deducting the observed difference in order value between the two groups before the change (1). Thus, the difference-in-differences methodology exploits an intertemporal comparison and a cross-sectional comparison in one design with a “treatment” dummy variable and a “temporal” dummy variable (Wooldridge 2012, pp. 452–455). This results in a two-by-two matrix in which one axis represents the “pre” and “post” period with regard to the timing of the intervention and the other axis represents the “treatment” and “control” units of observation (Goldfarb and Tucker 2014, p. 9). By combining both dimensions, the desired treatment effect can be isolated in β_3 corresponding to the interaction of both dimensions, “treated” and “post,” as presented in the stylized regression in Fig. 4.

Core Area of Application

The difference-in-differences methodology exploits two different comparisons within one design: an intertemporal comparison and a cross-sectional comparison. The first comparison is the intertemporal difference within the treatment group; thus, the difference in observation before and after the treatment only for the group which has been affected by the manipulation (Wooldridge 2012, pp. 453–455). While this would satisfy causality condition one (cause precedes the outcome) and allows to observe a co-movement between the observed outcome and the treatment (causality condition two), it would not satisfy causality condition three (absence of other interfering factors) because general time trends or other global shocks are not accounted for. Thus, the observed outcome cannot be considered causal to the treatment. For example, the fashion retailer switches from a fixed-pay scheme (“pre” period) to a commission-based pay scheme (“post” period) for its shop floor personnel and subsequently the performance of the sales people (e.g., revenue per sales person) increases, satisfying causality condition one and two. Yet, condition three is not satisfied because general time trends such as an economic boom or

shocks such as a competitor filing for bankruptcy are not incorporated, both of which would affect the sales people's potential to sell products and in turn affects the retailer's revenue.

The second difference constitutes the variation between the treatment and the control group. Here, observable outcomes of treatment and control group are compared for one or more points in time only after the intervention, that is, a cross-sectional difference is exploited. For the fashion retailer, this would mean changing the compensation scheme for its personnel in one store but leaving it as it is at another store and comparing the subsequent performance between the two groups after the change. While it is now possible to incorporate the difference between the two groups (and thus a general time trend) and a co-movement between "compensation scheme" and "sales made" can be observed, this approach neglects the difference between the two groups of sales people which was preexistent before the intervention, that is, working in two different stores (Angrist and Krueger 1999, pp. 1298–1299). Most likely, the two stores had different levels of sales before the change in compensation scheme due to their specific locations, product offerings, or manager's discretion with regard to promotions. Thus, a purely cross-sectional comparison would again fail to meet condition three (exclusion of other factors) and could not establish a causal relationship. In combining both differences, intertemporal and cross-sectional, the difference-in-differences methodology allows to draw causal inferences as long as it fulfills specific conditions (Goldfarb and Tucker 2014, p. 9). We discuss these conditions in the subsequent section in detail when we deal with critical assumptions of the difference-in-differences method.

A specific but important application area for the difference-in-differences methodology is a so-called staggered design. Here, researchers either design their field experiment in a way, or exploit the fact, that the treatment does not occur to all treated individuals or groups at the same time but is introduced step-by-step. Staggered designs are often observed when regulations or legal directives are adopted or firms subsequently roll out a new policy over subjects such as business units, facilities, employee groups, or customers. For example, contrary to an EU Regulation (e.g., the General Data Protection Regulation), an EU Directive has to be adopted into national law in each member state resulting in different adoption dates for a law throughout the EU which allows to observe the effects of the adoption at several points in time. In this setting, a member state belongs to the control group as long as the directive has not yet come into effect and belongs to the treatment group from that moment on when it comes into effect. As more countries adopt the new law, the size of the treatment group grows while the size of the control group shrinks. Such a staggered design is strengthening the case that it is unlikely that another, unobserved factor systematically drives the outcome because the treatment effect is observable for various treatment and control group compositions at various points in time. In this way, concerns regarding the randomness of assignment and control over treatment are mitigated. In this fashion, Heese and Pérez-Calvazos (2020) exploit the introduction of new flight routes between major cities throughout the USA over several years. Similarly, Aghamolla and Li (2018) exploit the staggered introduction of new (and enhanced) debt contract enforcement in India as an opportunity to apply a difference-in-differences design in a natural field experiment.

Some authors also consider event studies as a special application of difference-in-differences designs (Atanasov and Black 2016, p. 219). Event studies generally exploit very short time frames such as a few days before and after the treatment and are common for research in capital markets and stock exchanges where market reaction times are short and prices evolve quickly. In these cases, the intertemporal difference stems from a period before and after an event happens (e.g., the announcement of a CEO leave). The cross-sectional difference stems from the comparison of the stock's performance against the broader market, that is, a market performance index such as the S&P500 or the Dow Jones Industrial Index. For example, in a field experiment in cooperation with Yahoo! Finance, Lawrence et al. (2018) investigate the influence of news coverage on the stock prices of the firms which are covered by the news. Yahoo! Finance randomly gave earnings announcements more or less prominent coverage on its website. Subsequent to this treatment, the authors observe the stock price reaction in a one-day period, comparing the difference in stock price movement and the index movement of the "post" period to the same difference of the "pre" period. In this vein, an event study can also be regarded as an application of a difference-in-differences design.

Critical Assumptions

The most relevant assumption for the difference-in-differences approach is the parallel trends assumption which postulates that the treatment and the control group do not necessarily have to be identical but have to be "very similar" or "comparable" (Antonakis et al. 2010, pp. 1108–1109), showing "similar trends" over time (Angrist and Krueger 1999, p. 1297). The difference between groups with regard to the outcome variable is expected to be stable over time so it can be assumed that they would have continued to stay in parallel had there not been the intervention. This idea follows, yet relaxes, the idea of a counterfactual in any experimental design according to which the control group should be identical to the treatment group to represent the unobservable counterfactual (Atanasov and Black 2016, pp. 218–219). Similar to the inherently unobservable true counterfactual, the assumption of a parallel trend continuing after the treatment ("post" period) cannot be tested empirically. Nevertheless, for any pair of treatment and control group, it can be shown that both groups were moving in parallel prior to the intervention ("pre" period) in order to plausibly state the assumption that this trend would have continued in absence of the intervention (Goldfarb and Tucker 2014, p. 15). In order to assure parallel trends, relevant variables and covariates need to be measured repeatedly at different points in time during the "pre" period. Additional plausible reasoning and supporting data can strengthen that the intervention was an exogenous shock in a way that assignment to treatment was as good as random (Atanasov and Black 2016, pp. 238–241). Gill et al. (2017) use an alternative way to ensure comparability of groups in a setting where buyers in a market could voluntarily self-select into the treatment (i.e., using a business-to-business app). They explicitly model the decision to become part of the treatment

via a first-stage selection model, thus correcting statistically for systematic differences between treatment and control group before intervention.

The absence of a single suitable control group that fulfills the parallel trend assumption lead to the development of synthetic controls (Abadie et al. 2010). This method has become popular in economics, marketing, and other disciplines in recent years since it is particularly useful in studying interventions that are implemented at an aggregate level, affecting a small number of large units (such as regions or business units). The synthetic control method is based on the idea that a weighted combination of candidates for a control group provides a more appropriate comparison than a single control group alone. Therefore, the control group is a weighted combination of several control group candidates where the respective weights are the outcome of a prediction model. Based on observable variables, the approach fits the weighted control group to the treatment group before the treatment takes place, thus constructing a hypothetical control, i.e., a synthetic group. In one application, Pattabhiramaiah et al. (2019) require a control group for the readership of the *New York Times* newspaper and are able to compose a weighted average of other newspapers such as the *Washington Post*, the *LA Times*, and the *Chicago Tribune* to study the effects of a paywall introduction on the newspaper's revenue. For further details and application areas of this approach, we refer to Abadie et al. (2010), Abadie (2020), and Acemoglu et al. (2016).

While the parallel trends assumption is basically always addressed in research that employs a difference-in-differences design, many other assumptions are often left implicit (we refer to Atanasov and Black 2016, pp. 237–249 for a thorough review). One of those assumptions addresses the issue that there must not be a second event which has diametrically opposed effects on the treatment and the control group, that is, events to which subjects in the two groups react systematically different based on their group status. Armstrong et al. (2019) explicitly point out three more assumptions with relevance to the difference-in-differences method and provide reasoning to which extent their specific research design meets those assumptions:

- The first additional assumption beyond parallel trends postulates that the treatment status of one unit should not interfere with another unit's outcome. For example, customer A's status (qualified for free shipping) should not affect the amount of money spent by another customer B. This assumption is not testable but has to be analyzed and plausibly argued in each research setting.
- The second additional assumption demands that neither subject in the treatment or in the control group anticipates the intervention or is affected by the treatment prior to the intervention. This assumption might be violated in cases where subjects can voluntarily choose to become part of a group, that is, a customer may decide to purchase an additional item to cross the threshold of the free shipping minimum order amount. Other examples of such violations include changes in accounting rules, voluntary compliance with stricter environmental standards, or customer data protection.
- Finally, the difference-in-differences method implicitly assumes perfect compliance. This means all individuals in the treatment group are in fact treated. In case

of a violation, the method estimates (just) an “intention to treat” effect. This intention to treat effect is more conservative than the treatment on the treated estimation because it is based on a smaller variance within the data set making it harder to detect a statistically significant effect between treatment and control group (see also Gassen and Muhn 2018, pp. 21–22). For the free shipping threshold, it may occur that certain customers may either simply not recognize that free shipping is available or at which threshold they would qualify for the free shipping.

Despite being relevant to difference-in-differences designs in particular, the ideas expressed in these additional assumptions are not exclusive to difference-in-differences models but can be seen as some broader requirements for conducting (field) experiments in general (Lourenço 2019). Most field experiments implicitly consider these assumptions by prohibiting units of observations to interact with each other, so that the treatment is not “diluted.” For example, Casas-Arce et al. (2017a) explicitly state that “*information sharing among professionals [treated units] was not common*” and Casas-Arce et al. (2017b) state that the treatment (here: introduction of a simulation software) “*was installed overnight and a memo was sent to branch managers with instructions for its use,*” assuring that the treatment could not have been anticipated.

Application in Goldfarb and Tucker (2011)

Goldfarb and Tucker (2011) use the difference-in-differences method in their study investigating the (potential) impact of a regulatory reform on advertisement effectiveness based on the introduction of an online privacy law in 2004 by the European Union. Before the implementation of the Privacy Regulation, firms had the ability to broadly monitor users’ online activities and behaviors and thus were able to specifically target users with advertisement based on their activities. For example, knowing that a user searched the terms “holiday” and “Spain” would lend itself to promoting summer fashion or flights to Madrid. With the introduction of the Privacy Regulation, marketers’ possibilities to collect and use such information were severely limited, making it plausible to observe a decline in advertising effectiveness subsequent to the introduction. Goldfarb and Tucker (2011) exploit the change in legislation in a difference-in-differences approach in which they compare advertising effectiveness before and after the change in regulation.

The authors are able to exploit unique data of advertising effectiveness from a repeated survey response study conducted by a research agency between 2000 and 2008 in the EU and the USA. In this survey, the research agency measured ad effectiveness for marketers by randomly showing ads and placebo ads to real online users and subsequently surveying users from both groups with regard to purchase intent of the promoted product. While in itself this (repeated) field experiment already provides the possibility to draw causal inferences on the ad effectiveness, the authors add another dimension, namely time, to measure the change in

effectiveness. Hence, the treatment in this design is the implementation of the new privacy regulation. All users who were shown personalized ads belong to the treatment group, while all users who were shown placebo ads belong to the control group (i.e., a cross-sectional comparison). The addition of the time dimension accounts for the difference-in-differences design of this quasi-natural experiment. Effectively, Goldfarb and Tucker (2011) are even able to exploit a so-called triple difference-in-differences design with changes in three different dimensions: the variation over time, the variation between treatment and control group within the EU, and the differences between the EU and the USA. Using the rigorous difference-in-differences design, their results not only suggest that the advertisement effectiveness declined after the Privacy Regulation was put into effect but also that this outcome is a causal result of the introduction of the Privacy Regulation (Goldfarb and Tucker 2011, p. 70).

Table 1 provides an overview of further applications of the difference-in-differences method in business research literature exploiting various settings and exogenous shocks.

Regression Discontinuity Designs

Introduction

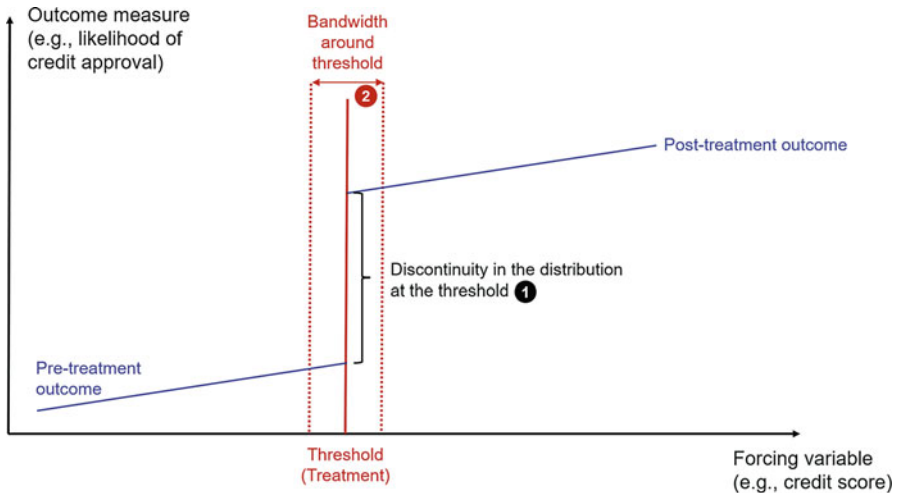
As discussed in prior sections, field experiments are building on the critical aspect of random assignment to the treatment. Quasi field experiments lack this requirement due to a nonrandom assignment of the treatment. For example, in the US banking industry, loans were granted without requesting further proof of income or collateral as long as the applicant had a FICO score (US credit rating) of 620 points or more (Keys et al. 2010). Thus, 620 points serve as the threshold for (not) getting a loan without further documentation, that is, for receiving the treatment. Using the cut-off to differentiate between treatment and control group is not per se helpful to establish causality. The problem is that those two groups are not comparable because a credit score is specifically designed to reflect a wide range of indicators, to differentiate credit worthiness levels, and to facilitate credit approval decisions. However, it may be more likely that an applicant scoring 619 is at least very similar to an applicant scoring 621 despite having a (marginally) lower score. Regression discontinuity designs make use of a comparison between subjects in close proximity to a threshold that are likely to be (almost) equal in terms of characteristics else than the treatment.

Figure 5 depicts the regression discontinuity setting. The horizontal axis may depict the credit score of applicants and the vertical axis may depict the likelihood of receiving a loan without further documentation. Instead of showing a continuous, functional relationship between credit score and the need to provide further documentation, the plotted distribution manifests a jump (or step) at the threshold value of 620 points (1). This arbitrary threshold divides the sample into a treatment and control group. The bandwidth (2) describes the range in close proximity to the

Table 1 Literature table for difference-in-differences applications in business research

Author(s), year, journal	Title	Type of field experiment	Dependent variable(s)	Independent variable(s), exogenous treatment	Findings
Amiram, D., Bauer, A. M., Frank, M.M. (2019) <i>The Accounting Review</i>	Tax Avoidance at Public Corporations Driven by Shareholder Taxes: Evidence from Changes in Dividend Tax Policy	Quasi-natural field experiment	Corporate tax avoidance measured in spread between statutory and effective tax rate	Change in countries' dividend tax policy triggered by ECI decisions, dividend payouts, share repurchases, array of financial controls	Average firm affected by an elimination of imputation systems reduces its cash effective tax rate by 5.5%
Ascarza, E., Ebbes, P., Netzer, O., Danielson, M. (2017) <i>Journal of Marketing Research</i>	Beyond the Target Customer: Social Effects of Customer Relationship Management Campaigns	Field experiment	Telephone usage (in minutes) and churn rates of customers who are socially connected to the targeted customers	CRM campaign targeted at specific, random telecom providers' customers	CRM campaign targeted at changing the behavior of specific customers transmits through social network, socially connected customers also increase consumption and decrease churn rates
Chen, T., Lin, C. (2018) <i>Journal of Financial and Quantitative Analysis</i>	Does Information Asymmetry Affect Corporate Tax Aggressiveness?	Natural field experiment	Five standard measures of tax avoidance (among others: total book-tax difference)	Analyst coverage of firms, closure and mergers of brokers covering firms as exogenous treatment	Firms avoid tax more aggressively after a reduction in analyst coverage
Gill, M., Sridhar, S., Grewal, R. (2017) <i>Journal of Marketing</i>	Return on Engagement Initiatives: A Study of a Business-to-Business Mobile App	Field experiment	Return on Engagement Initiatives (RoEI) and Sales Revenues	Customer engagement via a mobile app; adoption of the app as treatment (combined with mitigation of self-selection problems)	Adoption of the app increased revenues by 19.1 to 22.8% and resulted in positive RoEI
Grullon, G., Michenaud, S., Weston, J.P. (2015) <i>The Review of Financial Studies</i>	The Real Effects of Short-Selling Constraints	Natural field experiment	Corporate investment, external financing activities, stock price sensitivity	Capital market frictions operationalized as the SEC announcement of removal of short sale restrictions	Increase in short-selling activity causes stock prices to decrease; small firms react to these lower prices by reducing equity issues and investments

Houston, J.F., Lin, C., Liu, S., Wie, L. (2019) <i>The Accounting Review</i>	Litigation Risk and Voluntary Disclosure: Evidence from Legal Changes	Quasi-natural field experiment	(Frequency of) Disclosures with regard to financially relevant events	Good/Bad News, three legal decisions with regard to disclosure requirements as plausibly exogenous shocks	Treated firms tend to make fewer (more) management earnings forecasts relative to the control firms when they expect litigation risk to be lower (higher) following the legal event
Huse, C., Kopyug, N. (2017) <i>Journal of Economics and Management Strategy</i>	Bailing on the Car that was not Bailed Out: Bounding Consumer Reactions to Financial Distress	Quasi-natural field experiment	Brand's sales	Consumer confidence, financial distress of durable consumer goods manufacturers as exogenous treatment	Significant decrease in the sales of Saab following its filing for insolvency/restructuring
Gong, S., Zhang, J., Zhao, P., Jiang, X. (2017) <i>Journal of Marketing Research</i>	Tweeting as a Marketing Tool: A Field Experiment in the TV Industry	Field experiment	Percentage of channel's audience viewing particular show	Number of firms' tweets and influencers' retweets, number of noncommercial tweets	Company tweets directly increase viewing; influential retweets increase viewing only if the show tweet is informative
Jiang, J., Wang, I., Wang, K. (2019) <i>The Accounting Review</i>	Big N Auditors and Audit Quality: New Evidence from Quasi-Experiments	Quasi field experiment	Audit quality measured in financial statement divergence score	Big N auditors' acquisition of Non Big N auditors (and forced switch in auditor for the client)	Treated firms' audit quality improves after switch to Big N auditor
Mochon, D., Johnson, K., Schwartz, J., Ariely, D. (2017) <i>Journal of Marketing Research</i>	What Are Likes Worth? A Facebook Page Field Experiment	Field experiment	Customers' online and offline responses to invitation of liking firm's Facebook page	Invitation to like firm's Facebook pages as treatment (randomized subset of existing customers); promotional activities on the page	Acquired (invited) Facebook page likes have a positive effect on offline customer behavior



1. Discontinuity at the threshold (e.g., credit score 620), as soon as the threshold is reached, the outcome changes
2. Bandwidth around the threshold which forms the sample under investigation (e.g., credit score from 610 to 630)

Fig. 5 Graphical illustration of regression discontinuity design (own illustration)

threshold of 620 points for which it is assumed that applicants are reasonably comparable along each dimension despite being above/below the threshold value.

Core Area of Application

Regression discontinuity designs provide an opportunity for causal inference when the assignment to the treatment is unlikely to be random for the whole sample but may be as good as random around a certain threshold. In contrast to a difference-in-differences approach, the regression discontinuity only refers to two different groups and one type of variation: The difference in outcome for the treated group (right of the discontinuity) and the outcome for the control group (left of the discontinuity). The cut-off may be either a cross-sectional difference or an intertemporal difference but cannot be both (as we have seen for the difference-in-differences approach in the previous section).

Antonakis et al. (2010) exemplify this idea in a business context with leadership training for managers, for which the effect on team performance is unclear. In an ideal field experiment, leadership training would be offered to a randomly selected group of managers independent of prior performance. Yet, the firm wants to spend funds as efficiently as possible and wants to prioritize those managers for leadership training who showed lower performance in prior periods. Thus, the firm favors managers for the training who scored below average in their past annual evaluation. While comparing the team performance of all managers below and above the average performance subsequent to the training would not yield satisfactory results, the regression discontinuity approach would compare the teams' performance whose

managers were just above and just below the average. The comparison thus is limited to those managers who were just not granted training against those who just received the training. In this fashion, regression discontinuity designs also demonstrate a potential pathway to accommodating a firm's requirement for efficient spending and the requirements for drawing causal inferences.

Critical Assumptions

As it becomes evident, a regression discontinuity design requires a discontinuous function, that is, the treatment should show a plausible, discrete jump in its distribution. Therefore, it becomes critical to assess whether a jump in a distribution originates from an actual discontinuity or instead represents a nonlinear relationship around the threshold. We discuss three potential ways of verifying the plausibility of a suspected discontinuity.

First, it can be assessed whether the jump is plausible via an understanding of the data generation process. In the example of US credit scores and lending decisions, it is obvious that the discontinuity is an arbitrary threshold and that individuals most likely are unable to specifically manipulate their scores to belong to the treatment or control group. A different case is the example of the fashion retailer's free shipping option. Here, the retailer may decide to offer free shipping above an order value of 40 Euro. While this may lead to a discontinuity in the distribution of order values, this design in itself does not qualify for a regression discontinuity approach. Here, shoppers being close to the threshold have the discretion to become part of the treatment group (the group of people receiving free shipping). Due to this self-selection, the treatment neither is exogenously assigned to shoppers nor is it random, and a regression discontinuity analysis will not yield causal inferences (Goldfarb and Tucker 2014, p. 23).

Second, managers and researchers can statistically test whether the comparability assumption holds and whether subjects within a certain bandwidth around the discontinuity are comparable in observable characteristics. For the online shopper example, it could be useful to compare prior order values of the customers left and right of the discontinuity or to compare characteristics such as age or method of payment as proxies for shoppers' socio-economic background. For a more sophisticated test of self-selection and sorting behavior of subjects in regression discontinuity designs, we refer to McCrary (2008).

Third, it is possible to statistically test for different data generating processes and assess whether those are likely to be a valid origin of the observed data. While it is often advised to test whether a higher order polynomial may have generated the perceived discontinuity in the data, Gelman and Imbens (2019) specifically stress the idea to use local higher order polynomials (i.e., a quadratic function) or local linear functions instead of global higher order polynomials (e.g., sixths order) to control for the possibility of continuous data generating processes and to avoid poor or highly sensitive estimations.

Another crucial point in all regression discontinuity designs is the power of statistical tests. On the one hand, the number of observations around the threshold

is often small, thus, allowing for the possibility that single extreme values drive findings. From this perspective, a wider bandwidth would be favorable. On the other hand, regression discontinuity designs critically build on the assumption that subjects left and right of the cut-off are comparable along observable and unobservable characteristics. This perspective speaks in favor of a narrow bandwidth around the threshold. Hence, widening the bandwidth around the threshold will decrease the risk of influential outliers due to more observations, yet it will simultaneously corrode the assumption of comparableness between treatment and control group. Therefore, the bandwidth around the discontinuity becomes a critical aspect. A simple and effective way of dealing with this issue is to work with different bandwidths around the threshold value such as $\pm 1\%$, $\pm 2.5\%$, $\pm 5\%$, and $\pm 10\%$ to determine which bandwidth yields the optimal trade-off. Another solution is proposed by Imbens and Kalyanaraman (2012) who suggest to select a bandwidth by minimizing the mean squared error which is an efficient, yet economically less intuitive solution.

Application in Flammer (2015)

Flammer (2015) addresses the question of whether the approval of a Corporate Social Responsibility (CSR) related shareholder proposal is firm value enhancing. She motivates her study by presenting two competing theories. On the one hand, a resource-based view predicts that firms only engage in activities which are value enhancing. On the other hand, agency theory argues that private managerial concerns for reputation may motivate executives in engaging in CSR activities in order to increase personal reputation at the cost of firm value. It should be noted, however, that those two arguments are not mutually exclusive. The design of the field experiment in this paper can more easily explain whether the resource-based view holds because it insufficiently addresses the question of whether the agency explanation holds in this research setting due to the shareholders voting instead of management taking decisions.

To address her research question, Flammer (2015) exploits a discontinuity in the data – a majority vote threshold. She collects data from publicly listed firms in the USA concerning their relative stock market performance (cumulative abnormal returns) as outcome variable and data on all CSR-related shareholder proposals which were put to a vote in the annual shareholder meeting. In case a shareholder proposal met the threshold of “50% plus one vote,” it was approved and, thus, regarded as treated. All other proposals that did not pass the majority vote threshold are assigned to the control group. In a second step, she isolates all those proposals which were approved and rejected at a very thin margin (± 5 percentage points around the threshold) and investigates the share price development of those firms subsequent to the shareholder meeting.

As outlined before, it is crucial to provide evidence that the jump in the distribution can be attributed to a discontinuity in the data. The threshold of 50% approval provides plausible grounds. Visual inspection of the distribution of abnormal returns

for the vote share bandwidth of 45% to 55% also reveals a jump in the distribution around the approval threshold of 50%. Flammer (2015) further provides statistical evidence that the discontinuity unlikely stems from a higher order polynomial data generating process and therefore provides sufficient indication that she is in fact exploiting a valid discontinuity in the data.

After having provided evidence for the validity of the discontinuity, she also provides evidence regarding the second relevant requirement, the absence of self-selection effects. To ensure that the distribution is as good as random, she first investigates potential preexisting differences in various variables before and after the shareholder meeting. For the full sample, that is, all firms that held a vote on CSR-related proposals, she finds that firms *“that pass a CSR proposal differ significantly from companies that reject it”* (Flammer 2015, p. 2557). This finding supports the idea that passing a CSR proposal is not independent of other firm characteristics. Nevertheless, when only comparing firms which narrowly pass or reject a CSR proposal (i.e., firms within the $\pm 5\%$ bandwidth), these differences disappear, providing plausible evidence that the passage of a CSR proposal is uncorrelated to firm characteristics for this subsample. Unfortunately, the risk remains that the latter result (statistically insignificant differences in the subsample) is partially driven by a smaller sample size when considering only firms at the threshold and therefore has to be considered with caution. Therefore, this evidence is further underscored by a formal approach which tests for the continuity of vote shares in the data set (McCrary 2008).

Employing a regression discontinuity design, Flammer (2015) is able to provide empirical evidence that a firm’s CSR engagement is not only positively correlated to its financial performance but that it is likely that engaging in CSR initiatives in fact drives firm value. Her study demonstrates how valid thresholds can be exploited to establish causality in quasi field experiments but also demonstrates the method’s limits: Due to the small sample size around the threshold (61 observations), the author cautions against generalizing her findings and suggests additional studies on this issue.

Table 2 provides an overview of further applications of the regression discontinuity method in business research literature exploiting various settings and quasi-random distributions through arbitrary thresholds.

Instrumental Variables

Introduction

While the difference-in-differences method and the regression discontinuity design broadly lend themselves to satisfy causality condition three, the absence of interfering factors, instrumental variables may be used to solve more specific and complex issues. In general, instrumental variables take a special role with regard to field experiments because the method allows to embed an experiment into settings which did not allow for causal inference because the setting did or could not provide a (quasi) exogenous treatment, that is, when the setting was not a field experiment in

Table 2 Literature table for regression discontinuity applications in business research

Author(s), year, journal	Title	Type of field experiment	Dependent variable	Independent variable(s), exogenous treatment and threshold	Findings
Bird, A., Karolyi, S.A. (2019) <i>The Accounting Review</i>	Governance and Taxes: Evidence from Regression Discontinuity	Natural field experiment	Tax avoidance measured in effective tax rates (book-tax expense and cash taxes paid)	Institutional ownership variation over time, threshold of (not) being included in Russel 1000 and Russel 2000 index (entering or leaving index)	Increased institutional ownership around Russell index reconstitutions lead to significant decreases in effective tax rates and to increased use of international tax planning
Chang, Y., Hong, H., Liskovich, I. (2015) <i>The Review of Financial Studies</i>	Regression Discontinuity and the Price Effects of Stock Market Indexing	Natural field experiment	Stock Price	Inclusion and exclusion in Russel 1000 and Russel 2000 index as treatment/threshold	Additions to the Russell 2000 result in price increases and deletions result in price decreases
Chava, S., Roberts, M.R. (2008) <i>The Journal of Finance</i>	How Does Financing Impact Investment? The Role of Debt Covenants	Natural field experiment	Firms' investments (measured in ratio of CAPEX)	Debt covenant violations and subsequent threat of transfer of ownership, violation of the covenant as the arbitrary threshold	Capital investment declines sharply following a financial covenant violation
Chemmanur, T.J., Tian, X. (2018) <i>Journal of Financial and Quantitative Analysis</i>	Do Antitakeover Provisions Spur Corporate Innovation? A Regression Discontinuity Analysis	Natural field experiment	Innovation measured in number of patents and citations of patents	Proposals of antitakeover provisions which did (not) pass the shareholders' vote, threshold at 50% of shareholder votes	Positive effect of passage of antitakeover provisions on innovation

Kajüter, P., Klassmann, F., Nienhaus, M. (2019) <i>The Accounting Review</i>	The Effect of Mandatory Quarterly Reporting on Firm Value	Quasi-natural field experiment	Firm valuation, cumulative abnormal returns	Exemption from quarterly financial reporting; stock market capitalization threshold at fiscal year end	Five percent decrease in firm value, consistent with theory that mandatory quarterly reporting is a net burden for smaller firms
Narayanan, S., Kalyanam, K. (2015) <i>Marketing Science</i>	Position Effects in Search Advertising and their Moderators: A Regression Discontinuity Approach	Quasi-natural field experiment	Click through rates and sales orders	Higher position in Google Search measured as the difference in AdRank	Position effects are stronger when advertiser is smaller; position effects are weaker when keyword phrase has specific brand/product information
Vashishtha, R. (2014) <i>Journal of Accounting and Economics</i>	The Role of Bank Monitoring in Borrowers' Discretionary Disclosure: Evidence from Covenant Violations	Natural field experiment	Firms' disclosure measured as likelihood of issuing earnings forecast	Increased bank monitoring of firms (change in governance structure) subsequent to debt covenant violations	Firms reduce disclosure subsequent to change in governance structure (violation of debt covenant)

the first place (Angrist and Krueger 1999, p. 1300). Second, instrumental variables are used whenever simultaneity poses a challenge to causal inference (thus, violation of condition one, cause has to precede the outcome). Third, instrumental variables are employed specifically when unobservable variables not only drive the observed outcome variable but also influence the presumed treatment variable. For example, this could be the case if subjects self-select into the suspected treatment or are noncompliant with the treatment. This influence of other factors on both treatment and outcome variable is often referred to as “endogeneity of the treatment.” We refer to Ebbes et al. (chapter ► [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers”](#)) in this handbook, Roberts and Whited (2013), and Angrist et al. (1996) for a more detailed and specific discussion about endogeneity.

Suppose a fashion retailer is aiming to increase its online store revenue by increasing its spending on paid search ads. As it is common, the search platform, such as Google or Yahoo!, charges the fashion retailer for every click on the paid search result that directs a customer to the fashion retailer’s online store (pay-per-click model). If the fashion retailer is now interested in the effectiveness of the paid search ads to generate revenues, the retailer would observe that the ad expenses and the online store revenues increase by some factor, for example, for every 1% increase in ad expense, revenues increase by 0.1%. Yet, this simple observation of comovement would ignore that the consumer’s behavior (a click on the ad) both drives the revenues and the expenses alike because the ad expense is not a lump-sum payment (Blake et al. 2015). In this case, instrumental variables are addressing the concern that the treatment cannot be clearly distinguished from other factors which affect the outcome and the treatment itself by “replacing” the original treatment variable with an instrumented variable, that is, an estimated variable based off an instrument. We will present and discuss examples for valid instruments and instrumental variables in the context of field experiments in the following segments.

Before turning to instrumental variables in field experimental settings, it is crucial to understand the mechanics of the instrumental variables approach. An instrument offers exogenous variation which the actual treatment cannot provide (Angrist and Krueger 1999, p. 30). In brief, an instrument is a “third” variable which explains as much variation in the treatment variable as possible but is exogenous to the outcome variable of interest, except for its influence through the treatment. The instrument thus replaces the treatment variable with estimated “cleaned, exogenous” values of the treatment instead of actual values by estimating the treatment variable in the first stage.

This idea is depicted in Fig. 6. The first box shows the original relationship of interest, for example, the effect of education (X) on income after graduation (Y). As the treatment education is endogenous (i.e., future income expectations are likely to influence education choices), an appropriate instrument (Z) is required. This instrument should be valid (i.e., it explains the number of years an individual spends on education) but exogenous (i.e., it is only related to income after graduation via education). Such an instrument could be mandatory school attendance (Angrist and Krueger 1991). In a first step, this instrument is used to estimate (mandatory) years of education (\hat{X}) which in turn is used to estimate income after graduation. The terms ϵ and u represent the respective residuals of the performed regression whose potential interrelation may be one reason for unobserved effects in the model.

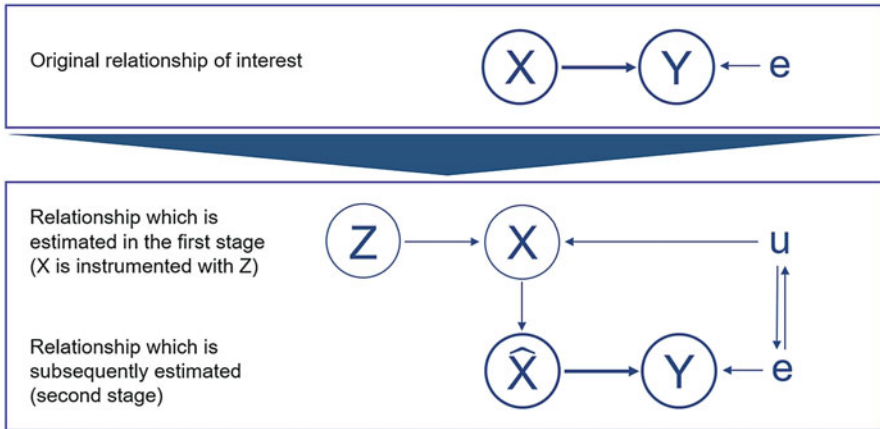


Fig. 6 Graphical illustration of instrumental variables method (own illustration)

Core Area of Application

Instrumental variables are useful in cases in which the main research question does not constitute a field experiment due to the lack of a (quasi) exogenous intervention as described for difference-in-differences settings (Angrist and Krueger 1999, p. 1300; Goldfarb and Tucker 2014, p. 26). Putting it in a more pointed way, the second stage of the instrumental variables design is not an experiment with (quasi) exogenous treatment because if it were, the research question would not require an instrumental variables design in the first place but could be addressed with a comparison of means or possibly a sophisticated difference-in-differences design. In that sense, an instrumental variables approach implements a (quasi) natural field experiment before the original relationship of interest is addressed. Turning to the causality conditions, instrumental variables provide an opportunity to establish causality in cases where either condition one (cause precedes effect) or condition three (absence of interfering factors) is difficult to meet. In other words, in these cases, it is difficult to isolate the effect of the treatment variable from an array of other unobservable factors. This also includes cases in which subjects deliberately decide to become part of the treatment group, that is, when (quasi) random assignment cannot be ensured. Here, instrumental variables have become the standard method to address this problem, developed and driven mostly by research in economics (e.g., Angrist and Krueger 1991). Nevertheless, it has found its way into business research as a result of stronger emphasis on causal inferences.

For any field experiment which can be considered being part of the first quadrant, that is, the designer has full control over the treatment and the treatment is randomized, an instrumental variable approach is generally not necessary because the treatment (i.e., the cause) precedes the outcome by design and ideally there are no unobservable characteristics which influence the treatment and the outcome alike. For natural field experiments or quasi field experiments, the issue of other (un)observable influences becomes of increasing concern as managers or researchers are

lacking control over those circumstances. In most cases, potentially interfering factors are nevertheless observable and only affect the outcome but not the treatment and can thus be integrated into the estimation as control variables. For example, Goldfarb and Tucker (2011) employ their difference-in-differences design on the introduction of a new privacy law in the EU but include a range of control variables in their estimation such as age or income of the survey participants. Furthermore, the intervention is quasi exogenously determined by the EU regulatory body. Contrarily, in the case of Giroud et al. (2012), the authors are not able to isolate the effect of an increase in financial leverage from unobserved, opportunistic behavior of the hotels' managers, which they suspect to also influence the outcome variable, the likelihood of bankruptcy. Therefore, the authors resort to an instrument to estimate their treatment variable. Disaggregating their study, the core question of interest does not include an experiment because it lacks the randomized intervention from an external body (experiment's designer or regulatory body). Including an instrumental variable allows the authors to exploit an experimental setting (excessive amount of snowfall) in a first step before addressing the actual research question, making their research design qualify as a quasi-natural field experiment.

Another case can be made for situations in which the assignment to treatment is not random but individuals choose to become part of the treatment group, for example, if firms voluntarily disclose nonfinancial information. In an example of overcoming the problem of voluntary decisions by firms, Ladika and Sautner (2020) investigate the question whether CEOs are likely to scale back investments when they are presented with more short-term incentives. While so-called "accelerated option vesting" by firms is generally regarded as a valid indicator for shortening top management's incentive horizon, doing so is a voluntary decision by the firm. Thus, it is most likely correlated with factors which also drive investment decisions. In that sense, one may also argue that this setting per se does not qualify as a field experiment as the treatment is not exogenously determined. In order to mitigate this concern, Ladika and Sautner (2020) exploit a regulation by the Financial Accounting Standard Board (FASB) which creates a strong incentive for firms to accelerate option vesting. Importantly, this incentive becomes effective quasi-randomly for each firm at different points in time as it is tied to a firm's fiscal year end. Therefore, it can be used as an instrument for actual accelerated option vesting, again "attaching" an experiment to a research question which in itself would not qualify as a field experiment.

Critical Assumptions

Any instrument has to satisfy two fundamental criteria, the relevance condition, and the exclusion restriction (Wooldridge 2012, p. 508):

- According to the relevance condition, the instrument must be informative about the independent (exogenous) variable. For example, in their field experiment, Shriver et al. (2013) investigate the relationship between social ties and the

creation of content in an online social network. As those two concepts are intertwined, it poses a challenge to causality condition one (cause precedes effect). The authors choose wind speeds at the observed surfing locations as their instrument and establish that wind speeds are informative about the blogging behavior of surfers in their online social network. Here, wind speeds higher than the median enable surfing at a specific location and in turn enable surfers to write about their experience on an online platform.

- According to the exclusion restriction, the instrument must be exogenous to the dependent variable. This means that the instrument should only affect the observed outcome variable via the treatment. Again, Shriver et al. (2013) claim that wind speeds in itself should have no direct effect on the likelihood of surfers developing social ties in an online network. Stated differently, above median winds are per se not increasing the likelihood of surfers getting and sending friend requests in their online social network. Wind speeds should affect social ties in the online social network only through the creation of content of past surfing experiences.

Providing evidence that the instrument meets the relevance and exclusion restriction is challenging and statistical tests can at best only strengthen the credibility of the instrument. The relevance of instruments can be tested in the first-stage regression via the respective F-statistic of whether the instruments are jointly significantly different from zero. As a rule of thumb, an F-statistic smaller than 10 for a single treatment variable indicates the presence of a weak instrument, that is, that the chosen instrument is not sufficiently relevant for the relationship of interest. Yet, for certain cases, this rule of thumb may not be sufficient to determine the relevance of the instrument, for example, Lee et al. (2020) demonstrate the insufficiencies of this rule of thumb in many research applications and propose that a more sensible F-statistic value would lie north of 100. For further details on extended testing for an instrument's relevance, we refer to Stock and Yogo (2005) and more recent development for testing an instrument's validity put forth by Sanderson and Windmeijer (2016).

The exclusion restriction is even more difficult to assess because no single, universal statistical test exists to date. Thus, arguing in favor of a valid instrument which satisfies the exclusion restriction requires in-depth knowledge on the domain and a very sound understanding of the data generating process. Therefore, authors often caution their readers against taking results at face value. For instance, Shriver et al. (2013) state that they "*present an IV approach based on wind speeds, which delivers estimates of causal effects under the assumption of exogeneity of these instruments. If these assumptions are violated, the causal interpretation no longer holds.*" (Shriver et al. 2013, p. 1430).

Two further assumptions are required to sensibly employ an instrumental variable approach: the independence condition and the monotonicity assumption. The independence condition states that the instrument itself must not suffer from the same property as the variable that is instrumented, namely it must not suffer from endogeneity itself. This condition aligns with the idea of exploiting a field

experimental setting in the first stage. Secondly, the monotonicity condition states that the instrument must only have a one directional effect on the treatment variable. Specifically, the instrument may partially not affect certain individuals (e.g., because they effectively withdraw from being treated) but it must not adversely affect the likelihood of being treated (Angrist and Pischke 2009, p. 114). Circling back to Shriver et al. (2013), while wind speed may not affect those surfers who may have had a “rest day” in their schedule or were sick on one of the days in the observation period, above median wind speeds should not decrease the likelihood of surfing for some of the surfers but it should only increase the likelihood for all surfers. This assumption may be questioned, potentially arguing that for some extreme wind speeds, especially beginners may be discouraged from surfing. On the other hand, extreme winds may also encourage content creation for those beginners without an imminent surfing experience, for example, by creating content on the extreme weather conditions. This brief discussion again illustrates the problem of those assumptions being mostly untestable and heavily relying on convincing argumentation and discussions. Nevertheless, meeting all the aforementioned conditions and assumptions paves the way to deriving the local average treatment effect (LATE) in the desired research setting.

Application in Bannedsen et al. (2007)

Bannedsen et al. (2007) pursue the research question whether choosing a family CEO or an external CEO positively affects the financial performance of a family firm. Choosing a CEO from the family has the advantage of insider knowledge and greater alignment between the interests of the owning family and the CEO, which mitigates agency conflicts. In contrast, choosing an outside CEO enables the firm to choose a skilled manager from a larger talent pool. Thus, from a theoretical point of view, both scenarios, a family CEO being more and being less valuable to the firm compared to an outsider, are plausible.

Ideally, researchers have full control over the treatment and could randomly assign a family CEO or an outside CEO to identical family firms and measure their subsequent financial performance. Unsurprisingly, this is not feasible. Similarly, it is not helpful to simply investigate the correlation between “family vs. outsider CEO” on subsequent financial performance of all firms. This analysis would satisfy causality condition one and two but certainly not condition three, the absence of other influencing factors. The problem is that the choice of appointing a CEO from the ranks of the family is endogenous to its performance, that is, some factors influence both the appointment decision and subsequent financial performance. Thus, the research question itself does not constitute a field experiment due to the lack of an exogenous intervention. Therefore, Bannedsen et al. (2007) instrument the choice of appointing a family CEO with a plausibly exogenous variable: the gender of the CEO’s first-born child.

The authors gather financial data from privately held Danish firms and match information on whether the CEO’s first-born child is male or female. The gender of

the first-born is plausibly random and not influenced by their parents, especially as techniques to identify the gender before birth were not widely employed prior to 1980 (Bennedsen et al. 2007, p. 650). Further, the gender of the first-born is plausibly independent of firm's financial performance. While the chance of being born a male is 50%, the rate of family succession within Danish firms is 39% when the first-born is male while it is only 29.4% when the first-born is female (Bennedsen et al. 2007, p. 650). Stated differently, a male first-born has a 32.7% higher likelihood of becoming a CEO in his parents' company than a female first-born. The chosen instrument satisfies both the exclusion restriction (gender being uncorrelated to firm's financial performance) and the relevance condition (gender being informative about CEO appointment choice). As outlined before, an instrumental variable setup critically builds on plausible reasoning and extended empirical tests to demonstrate that the relevance condition and exclusion condition are fulfilled.

Exploiting the instrumental variable in this field experiment, the authors are able to identify a causal relationship between appointing a family CEO and subsequent financial performance. They find that family successions hurt firm performance: in particular, firms with the parting CEO's first child being male demonstrate an average decline in operative return on assets of 0.8 to 1.2 percentage points, statistically significant at the 5% level. Correlational analyses suggest that the negative effect persists, meaning that firms whose performance is relatively weaker subsequent to appointing a family CEO do not catch up to their peers in the 3 years following the CEO succession (Bennedsen et al. 2007, p. 678).

Table 3 provides further applications of the instrumental variables method in business research literature exploiting various settings and treatment variations. As discussed above, applications of instrumental variables in field experimental settings are rather rare as the method is only useful in specific situations which (by construction) are less prevalent in field experiments.

Application of Methods in Standard Software

As the presented methods to exploit data from field experiments have become increasingly more popular and wide-spread as demonstrated in Fig. 1, this has also materialized in their implementation in various standard software solutions. While a manual application of each method is generally possible in STATA, R, and SPSS, especially STATA and R often provide dedicated functions or packages to execute the methods in a more user-friendly, easy-to-handle way. This provides the user with a wide variety of possibilities to apply the methods tailored to the specific question at hand. Table 4 provides an overview of the application of the methods difference-in-differences, regression discontinuity, and instrumental variables in STATA, R, and SPSS.

In order to familiarize oneself with the presented methods, we suggest replicating research findings that build on publicly available data. As this data is often cumbersome to collect and manipulate to the point where an analysis becomes meaningful, the

Table 3 Literature table for instrumental variables applications in business research

Author(s), year, journal	Title	Type of field experiment	Dependent variable	Independent variable(s), instrument	Findings
Bernstein, S. (2015) <i>The Journal of Finance</i>	Does Going Public Affect Innovation?	Quasi-natural field experiment	Number of patents, citations of patents, employee turnover	IPO completion, instrumented by likelihood of completion measured in NASDAQ fluctuation subsequent to IPO filing (book building phase)	Quality of internal innovation declines after IP; key innovators leave firm and often found competing companies
Cannon, J.N. (2014) <i>The Accounting Review</i>	Determinants of "Sticky Costs": An Analysis of Cost Behavior using United States Air Transportation Industry Data	Quasi-natural field experiment	(Log) Change in available transportation capacity	Change in demand and capacity unit cost change, instrumented by industry-level unit costs and selling price changes	Managers retain idle capacity and lower selling prices to utilize existing capacity when demand falls, but add capacity (rather than raise selling prices) when demand grows
Hui, S.K., Inman, J.J., Huang, Y., Suher, J. (2013) <i>Journal of Marketing</i>	The Effect of In-Store Travel Distance on Unplanned Spending: Applications to Mobile Promotion Strategies	Field experiment	Unplanned customer spending	Mobile promotion, in-store path length, instrumented by reference path determined by store layout and shoppers' planned purchases	Increasing path length by 10% increases unplanned spending by about 16.1% (\$2.54); relocating physical products may increase unplanned spending by approximately 7%
Luong, H., Moshirian, F., Nguyen, L., Tian, X., Zhang, B. (2017) <i>Journal of Financial and Quantitative Analysis</i>	How Do Foreign Institutional Investors Enhance Firm Innovation?	Quasi-natural field experiment	Number of patents, citations of patents	Foreign institutional ownership, instrumented by varying membership in the MSCI All Country World Index foreign institutional ownership has a positive, causal effect on firm innovation	Foreign institutional ownership has a positive, causal effect on firm innovation

Table 4 Application of methods in standard software packages

Method	STATA	R	SPSS statistics
Difference-in-differences	<p>Manual application Generate a “post” and a “treated” dummy variable Generate an interacted variable from “post” and “treated” dummies Regression with interacted dummy variable Package “diff” <i>diff</i>: Estimate treatment effect based on a difference-in-difference method</p>	<p>Manual application Generate a “post” and a “treated” dummy variable Generate an interacted variable from “post” and “treated” dummies Regression with interacted dummy variable No package available</p>	<p>Manual application Generate a “post” and a “treated” dummy variable Generate an interacted variable from “post” and “treated” dummies Regression with interacted dummy variable No package available</p>
Regression discontinuity	<p>Manual application Create dummy variable for treatment status Restrict sample to bandwidth around threshold Regression with independent variable and the treatment status dummy Package “rdrobust” <i>rdrobust</i>: Estimate a treatment effect based on regression discontinuity method <i>rdbwselect</i>: Manual selection of the bandwidth around the threshold <i>rdplot</i>: Plot data to depict the discontinuity</p>	<p>Manual application Create dummy variable for treatment status Restrict sample to bandwidth around threshold Regression with independent variable and the treatment status dummy Package “rdd” <i>RDEstimate</i>: Estimate a treatment effect based on regression discontinuity method <i>IKbandwidth</i>: Optimal bandwidth according to Imbens-Kalyanaram <i>Plot.RD</i>: Plot data to depict the discontinuity</p>	<p>Manual application Create dummy variable for treatment status Restrict sample to bandwidth around threshold Regression with independent variable and the treatment status dummy No package available</p>

(continued)

Table 4 (continued)

Method	STATA	R	SPSS statistics
<p>Instrumental variables</p>	<p>Manual application Estimate instrumented independent variable using the chosen instruments as explanatory variables (first stage) Estimate dependent variable using the instrumented independent variable as explanatory variable (second stage) Built-in solution “ivregress” <i>ivregress</i>: Estimate a treatment effect based on instrumental variables method Package “ivreg2” <i>ivreg2</i>: Estimate a treatment effect based on instrumental variables method; more sophisticated than “ivregress” Built-in solution “ivprobit” <i>ivprobit</i>: Estimate a treatment effect based on instrumental variables method with a binary dependent variable</p>	<p>Manual application Estimate instrumented independent variable using the chosen instruments as explanatory variables (first stage) Estimate dependent variable using the instrumented independent variable as explanatory variable (second stage) Package “AER” <i>ivreg</i>: Estimate a treatment effect based on instrumental variables method</p>	<p>Manual application Estimate instrumented independent variable using the chosen instruments as explanatory variables (first stage) Estimate dependent variable using the instrumented independent variable as explanatory variable (second stage) Built-in solution “Two-stage Least Squares” Select Two-Stage Least Squares from Regression Menu</p>

following papers provide their applied data set ready for any researcher to use and replicate the findings. Furthermore, the authors of those studies also provide their code which provides a useful check for one own's replication efforts. Employing difference-in-differences designs in natural experiments, Calzada and Gil (2020) investigate the role of online news aggregators on news providers, De Silva et al. (2010) research the causal relationship of migration on wages, and Seiler et al. (2017) address the question of whether and how online word of mouth increases demand. Bronzini and Iachini (2014) and Shapiro (2018) are using regression discontinuity approaches; the former investigates whether incentives for R&D are effective and the latter derives causal claims with regard to advertising in the health insurance market. Lastly, instrumental variables approaches are used by Barron et al. (2020) who address the effect of home-sharing on house prices using Airbnb data and by Draca et al. (2011) who aim to disentangle the effect of police presence on crime using the London 2005 terror attacks as an exogenous event.

Conclusions

Drawing causal inferences instead of purely relying on associations has gained importance in business practice and research in the past 20 years. Traditionally, conducting experiments in the controlled environment of a laboratory with participants who were randomly selected from a certain pool of candidates was the primary method to gain insights into causal relationships. More recently, field experiments have gained relevance for researchers and practitioners alike due to new (online) possibilities for conducting self-designed experiments outside the laboratory. Yet, as field experiments do not provide the same degree of randomization and controllability, exploiting such data in a way which still provides the means to draw causal conclusions requires a set of selected methods. Moreover, it becomes apparent that the farther one parts from the ideal field experiment, namely towards quasi, natural, or quasi-natural field experiments, the more effort and sophistication is likely to be required in analyzing the data. In order to tackle short-comings in a field experiment's design, three methods have become a standard set in business research: difference-in-differences, regression discontinuity, and instrumental variables. The difference-in-differences method lends itself often when treatment and control group (s) are not necessarily equal but are sufficiently similar so that they are affected by the same environmental conditions and, thus, move in parallel with regard to a set of selected variables. The regression discontinuity method becomes useful when one is able to exploit an arbitrary cut-off, a threshold value, which quasi-randomly divides observations into a treatment and control group so that a prerequisite of causal inferences (random assignment to treatment) is restored for a sufficiently large subsample of the data. Finally, the instrumental variables method helps to exploit data in which one may be confronted with self-selection into treatment and control group, if critical influential variables cannot be observed or if simultaneity is likely present. Those issues may be solved by identifying a "third," exogenous, variable

which is able to estimate the treatment variable sufficiently well in order to address the present endogeneity problem.

The three methods presented in this chapter can be employed irrespective of the setting, that is, they may serve practitioners when evaluating the introduction of a new product in one or more markets, when analyzing the effect of a new legislation regarding advertising to children or when testing different purchase processes in the firm's online sales channels. This chapter aimed at providing an introduction to the three methods and their respective applications by providing intuitive, nontechnical explanations for each approach on how to exploit data from field experiments. Understanding, generating, and integrating insights from data created by field experiments will become more relevant even despite the growing availability of big data in business research and practice. Big data often only provides insights into correlational relationships potentially providing misleading guidance for decision making, creating the potential of misusing insights from big data, and leading to worse business decisions. Similar to Keele (2015), in a recent Harvard Business Review publication, Zoumpoulis et al. (2015) subsume that carefully conducted and analyzed field experiments can serve as a remedy and as a complement of increasing importance to make sense of purely correlational evidence from big data.

Cross-References

- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Experiments in Market Research](#)
- ▶ [Field Experiments](#)

References

- Abadie, A. (2020). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*. (forthcoming).
- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505.
- Acemoglu, D., Johnson, S., Kermani, A., Kwak, J., & Mitton, T. (2016). The value of connections in turbulent times: Evidence from the United States. *Journal of Financial Economics*, 121(2), 368–391.
- Aghamolla, C., & Li, N. (2018). Debt contract enforcement and conservatism: Evidence from a natural experiment. *Journal of Accounting Research*, 56(5), 1383–1416.
- Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4), 989–1014.
- Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1277–1366). Amsterdam: North-Holland.
- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics*. Princeton: Princeton University Press.
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455.

- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21(1), 1086–1120.
- Armstrong, C. S., Glaeser, S., & Huang, S. (2019). *Controllability of risk and the design of incentive-compensation contracts*. Working paper. Available at SSRN 2896147.
- Atanasov, V., & Black, B. (2016). Shock-based causal inference in corporate finance and accounting research. *Critical Finance Review*, 5(1), 207–304.
- Barron, K., Kung, E., & Prosperpio, D. (2020). The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Science*, 40(1), 23–47. Data available at: <https://services.informs.org/dataset/mksc/download.php?doi=mksc.2020.1227>
- Bennedsen, M., Nielsen, K. M., Perez-Gonzalez, F., & Wolfenzon, D. (2007). Inside the family firm: The role of families in succession decisions and performance. *The Quarterly Journal of Economics*, 122(2), 647–691.
- Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., & Zinman, J. (2010). What's advertising content worth? Evidence from a consumer credit marketing field experiment. *The Quarterly Journal of Economics*, 125(1), 263–306.
- Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *Econometrica*, 83(1), 155–174.
- Bronzini, R., & Iachini, E. (2014). Are incentives for R&D effective? Evidence from a regression discontinuity approach. *American Economic Journal: Economic Policy*, 6(4), 100–134. Data available at: <https://www.aeaweb.org/articles?id=10.1257/pol.6.4.100>
- Calzada, J., & Gil, R. (2020). What do news aggregators do? Evidence from Google News in Spain and Germany. *Marketing Science*, 39(1), 134–167. Data available at: <https://services.informs.org/dataset/mksc/download.php?doi=mksc.2019.1150>
- Casas-Arce, P., Lourenço, S. M., & Martínez-Jerez, F. (2017a). The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction. *Journal of Accounting Research*, 55(5), 1051–1088.
- Casas-Arce, P., Martínez-Jerez, F., & Narayanan, V. G. (2017b). The impact of forward-looking metrics on employee decision-making: The case of customer lifetime value. *The Accounting Review*, 92(3), 31–56.
- De Silva, D. G., McComb, R. P., Moh, Y.-K., Schiller, A. R., & Vargas, A. J. (2010). The effect of migration on wages: Evidence from a natural experiment. *American Economic Review*, 100(2), 321–326. Data available at: <https://www.aeaweb.org/articles?id=10.1257/aer.100.2.321>
- Draca, M., Machin, S., & Witt, R. (2011). Panic on the streets of London: Police, crime and the July 2005 terror attacks. *American Economic Review*, 101(5), 2157–2181. Data available at: <https://www.aeaweb.org/articles?id=10.1257/aer.101.5.2157>
- Flammer, C. (2015). Does corporate social responsibility lead to superior financial performance? A regression discontinuity approach. *Management Science*, 61(11), 2549–2568.
- Gassen, J., & Muhn, M. (2018). *Financial transparency of private firms: Evidence from a randomized field experiment*. Working paper. Available at SSRN 3290710.
- Gelman, A., & Imbens, G. W. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, 37(3), 447–456.
- Gill, M., Sridhar, S., & Grewal, R. (2017). Return on engagement initiatives: A study of a business-to-business mobile app. *Journal of Marketing*, 81(4), 45–66.
- Giroud, X., Mueller, H. M., Stomper, A., & Westerkamp, A. (2012). Snow and leverage. *The Review of Financial Studies*, 25(3), 680–710.
- Goldfarb, A., & Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, 57(1), 57–71.
- Goldfarb, A., & Tucker, C. E. (2014). *Conducting research with quasi-experiments: A guide for marketers*. Working paper. Available at SSRN 2420920.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Heese, J., & Pérez-Calvazos, G. (2020). When the boss comes to town: The effect of headquarter's visits on facility-level misconduct. *The Accounting Review*, 95(6), 235–261.
- Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3), 933–959.

- Keele, L. (2015). The discipline of identification. *PS: Political Science & Politics*, 48(1), 102–106.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience.
- Keys, B. J., Mukherjee, T., Seru, A., & Vig, V. (2010). Did securitization lead to lax screening? Evidence from subprime loans. *The Quarterly Journal of Economics*, 125(1), 307–362.
- Ladika, T., & Sautner, Z. (2020). Managerial short-termism and investment: Evidence from accelerated option vesting. *Review of Finance*, 24(2), 305–344.
- Lawrence, A., Ryans, J., Sun, E., & Laptev, N. (2018). Earnings announcement promotions: A Yahoo Finance field experiment. *Journal of Accounting and Economics*, 66(2–3), 399–414.
- Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2020). Valid t-ratio inference for IV. *ArXiv*, 2010.05058. Available from: <http://arxiv.org/abs/2010.05058>
- Lourenço, S. M. (2019). Field experiments in managerial accounting research. *Foundations and Trends in Accounting*, 14(1), 1–72.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698–714.
- Michels, J. (2017). Disclosure versus recognition: Inferences from subsequent events. *Journal of Accounting Research*, 55(1), 3–34.
- Pattabhiramaiah, A., Sriram, S., & Manchanda, P. (2019). Paywalls: Monetizing online content. *Journal of Marketing*, 83(2), 19–36.
- Presslee, A., Vance, T. W., & Webb, R. A. (2013). The effects of reward type on employee goal setting, goal commitment, and performance. *The Accounting Review*, 88(5), 1805–1831.
- Roberts, M. R., & Whited, T. M. (2013). Endogeneity in empirical corporate finance. In G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.), *Handbook of the economics of finance* (Vol. 2 (A), pp. 493–572). Oxford: North Holland.
- Sanderson, E., & Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, 190(2), 212–221.
- Seiler, S., Yao, S., & Wang, W. (2017). Does online word of mouth increase demand? (and how?) Evidence from a natural experiment. *Marketing Science*, 36(6), 838–861. Data available at: <https://services.informs.org/dataset/mksc/download.php?doi=mksc.2017.1045>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shapiro, B. T. (2018). Advertising in health insurance markets. *Marketing Science*, 39(3), 587–611. Data available at: <https://services.informs.org/dataset/mksc/download.php?doi=mksc.2018.1086>
- Shriver, S. K., Nair, H. S., & Hofstetter, R. (2013). Social ties and user-generated content: Evidence from an online social network. *Management Science*, 59(6), 1435–1443.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101.
- Stock, J. H., & Yogo, M. (2005). Asymptotic distributions of instrumental variables statistics with many instruments. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models*. New York: Cambridge University Press.
- Verbeek, M. (2008). *A guide to modern econometrics* (3rd ed.). West Sussex: John Wiley & Sons Ltd.
- Wooldridge, J. M. (2012). *Introductory econometrics: A modern approach* (4th ed.). Mason: South-Western Cengage.
- Zoumpoulis, S., Simester, D., & Evgeniou, T. (2015 November 12). Run field experiments to make sense of your big data. *Harvard Business Review*. Available from: <https://hbr.org/2015/11/run-field-experiments-to-make-sense-of-your-big-data>. Accessed 15 Sept 2020.



Mediation Analysis in Experimental Research

Nicole Koschate-Fischer and Elisabeth Schwillie

Contents

Introduction	858
Conceptual and Statistical Basics of Mediation Analysis	859
The Single Mediator Model	860
Mediation Models Including More Than One Mediator: The Parallel and Serial Multiple Mediator Model	871
Mediation Models Including a Moderator: Conditional Process Models	879
Further Mediation Models	887
Strengthening Causal Inference in Mediation Analysis	891
Strengthening Causal Inference Through Design	891
Strengthening Causal Inference Through the Collection of Further Evidence	893
Strengthening Causal Inference Through Statistical Methods	894
Questions Arising When Implementing Mediation Analysis	894
Sample Size and Power in Mediation Analysis	895
Mean Centering in Conditional Process Analysis	896
Coding of Categorical Independent Variables	897
Regression Analysis Versus Structural Equation Modeling	898
Software Tools for Mediation Analysis	900
Summary	900
Cross-References	901
References	901

Abstract

This chapter introduces the conceptual and statistical basics of mediation analysis in the context of experimental research. Adopting the respective terminology, mediation analysis can be referred to as an array of quantitative methods developed to investigate the causal mechanism(s) through which an independent variable influences a dependent variable. The chapter takes a regression-based

N. Koschate-Fischer (✉) · E. Schwillie
University of Erlangen-Nuremberg, Nuremberg, Germany
e-mail: nicole.koschate-fischer@fau.de; elisabeth.schwillie@fau.de

approach to mediation analysis and focuses on mediation models likely to be tested in experiments (i.e., the single mediator model, parallel and serial multiple mediator models, and conditional process models). Yet, the scope of mediation analysis beyond an experimental setting will also be touched upon. Furthermore, the chapter addresses the question how to strengthen causal inference in mediation analysis through design, the collection of additional evidence, and statistical methods. It closes with a discussion of common topics of relevance when implementing mediation analysis such as sample size and power, mean centering in conditional process analysis, coding of categorical independent variables, advantages and disadvantages of a regression-based approach to mediation analysis, and software options to perform mediation analysis.

Keywords

Mediation analysis · Conditional process analysis · Regression analysis · Bootstrapping · Experiments

Introduction

One focal goal of market research is to gain insight into whether and why marketing stimuli, such as price or advertising, affect consumer behavior. That is, it is not only important to demonstrate the causal effect of a marketing measure on consumer behavior (e.g., through conducting experiments, Koschate-Fischer and Schandelmeyer 2014; chapter ► “[Field Experiments](#)” by Valli et al., this volume), but it is also crucial to understand the causal mechanism(s) through which an effect occurs. A deeper understanding of the “why” or “how” of an effect is often gained through qualitative methods (e.g., focus groups or interviews). This chapter provides an introduction to regression-based mediation analysis, an array of quantitative methods developed to investigate the causal mechanism(s) through which an independent variable influences a dependent variable, which has gained increasing popularity in experimental research in marketing and market research over the last decade (e.g., Cavanaugh 2014; Koschate-Fischer et al. 2012, 2016; Savary et al. 2014; Touré-Tillery and McGill 2015).

The remainder of this chapter is structured as follows: Starting with the most simple mediation model, the single mediator model, the conceptual and statistical principles of mediation analysis are explained. These will then be applied to mediation models including multiple mediating variables (multiple mediator models) as well as a moderating variable (conditional process models). Further mediation models that take into account additional variables, time (longitudinal mediation models), and nested data (multilevel mediation models) are also briefly addressed. Subsequently, an overview is given of how to strengthen causal inference through design, the collection of further evidence, and statistical methods. The chapter closes with a discussion of selected questions arising when implementing mediation analysis.

Conceptual and Statistical Basics of Mediation Analysis

Adopting the terminology from an experimental context, *mediation* refers to a situation in which the effect of an independent variable on a dependent variable is transmitted through an intervening variable, the *mediator* (e.g., MacKinnon et al. 2007b; Preacher 2015; see also Mathieu and Taylor 2006). A mediator is a third variable included in the conceptual framework describing the simple effect of an independent variable on a dependent variable and can conceptually, yet not necessarily statistically (MacKinnon et al. 2000), be distinguished from other third variables: confounding variables, covariates, and moderators (MacKinnon et al. 2007b). *Confounding variables* influence both the independent variable and the dependent variable and, if unaccounted for (i.e., omitted from the model), bias the estimate of the relationship between the independent variable and the dependent variable. In an experimental context, *covariates* (also called concomitant variables) are variables that share variance with the dependent variable and controlling for them improves the estimation of the relationship between the independent variable and the dependent variable (Miller and Chapman 2001). *Moderators* influence the effect of the independent variable on the dependent variable, such that the magnitude or sign of the relationship changes depending on the values of the moderator.

In marketing and market research, mediators are likely to be psychological processes evoked by marketing stimuli affecting consumer judgment and behavior such as brand-, other-, or self-related cognitions and emotions. We agree with MacKinnon (2008) that mediators should be selected a priori based on theoretical considerations and the careful review of existing literature. If such a basis is not available, possible mediators could also be identified through, for example, qualitative interviews, and tested in subsequent quantitative studies. In either case, it is of paramount importance to carefully consider whether the mediator is a variable that can be causally affected by the independent variable and can, in turn, causally affect the dependent variable. Thus, stable consumer traits, such as cultural norms or values, cannot be mediators (unless the mediating variable denotes a change in such relatively stable traits, see Koschate-Fischer et al. 2017) and the mediator and the dependent variable have to be clearly distinguishable from each other (Pieters 2017).

The concepts and analyses described in this chapter will be illustrated with the help of a hypothetical experiment exploring the effect of sales promotions on consumers' positive word-of-mouth (WOM) intentions. Specifically, the experiment investigates whether a "free gift with purchase" promotion (e.g., "Receive a free summer gift with any \$10 purchase") increases positive WOM intentions through providing hedonic benefits to consumers (e.g., making the shopping experience more interesting and fun), a research question derived from the benefit congruency framework of sales promotion effectiveness (Chandon et al. 2000), and literature on WOM generation (Berger 2014). The independent variable in this example is manipulated on two levels (small vs. large free gift) and a between-subject design is employed. Hence, there are two experimental groups to which participants are randomly assigned. In both groups, participants read a scenario in which they

imagine they are browsing through an online store to buy a T-shirt. The store offers a wide variety of different T-shirts and brands. They encounter a banner on which it says that for each T-shirt of a specific brand bought today, consumers receive a free gift card for a future purchase in the online shop. In the small free gift condition, the banner states that participants will receive a “\$2 gift card with every T-shirt,” and in the large free gift condition, it states they will receive a “\$10 gift card with every T-shirt.” Participants are then asked about the hedonic benefits the sales promotion provides them and their willingness to positively talk about and recommend the promoted products by answering questions such as how inclined they are to like social media content referring to the promoted products. Hence, while the independent variable is manipulated, the mediator as well as the dependent variable are measured, which makes the experiment a measurement-of-mediation design (Spencer et al. 2005). Note that, as the experiment itself, the data the following analyses are based upon are hypothetical, i.e., simulated. Hence, the results reported do not allow to draw conclusions as to the ability of free gifts to stipulate positive WOM intentions through increasing hedonic benefits.

The Single Mediator Model

To illustrate the conceptual and statistical idea of mediation analysis, the most simple mediation model, the single mediator model, is described in the following section. It will be extended in later sections of the chapter by including additional mediating (multiple mediator models) and a moderating variable (conditional process models).

Conceptual Description of the Single Mediator Model

To explain the single mediator model, we start with a conceptual diagram (Hayes 2018) showing the simple causal effect of an independent variable X on a dependent variable Y (Fig. 1). In mediation analysis, this relationship is referred to as the *total effect* of X on Y .

In the single mediator model, this basic causal relationship is extended by a mediator M , which is causally located between X and Y (see Fig. 2). By including M , the total effect is split into direct and indirect components: X affects Y indirectly through M (path a and path b). In addition, X also affects Y directly (path c'). Together, path a and path b are referred to as the *indirect effect* of X on Y . The indirect effect indicates the effect of X on Y that is transmitted through M . Path c' denotes the *direct effect* of X on Y , which corresponds to the effect of X on Y that is

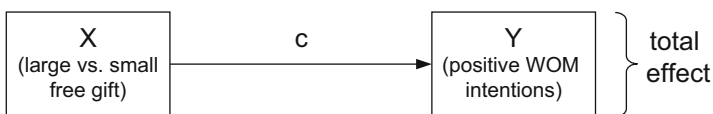


Fig. 1 The conceptual diagram of the total effect of X on Y

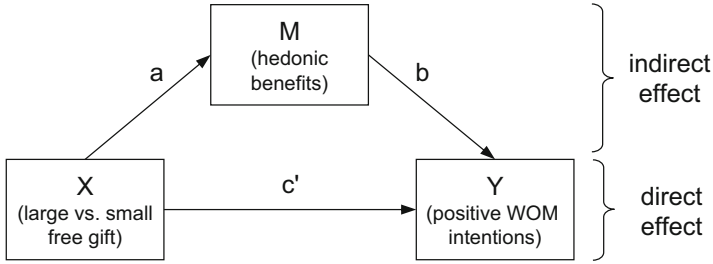


Fig. 2 The conceptual diagram of the single mediator model depicting the direct as well as indirect effect of *X* on *Y*

not transmitted through *M*. The direct effect (path *c'*) differs from the total effect (path *c*) in that it estimates the effect of *X* on *Y* while controlling for the indirect effect of *X* on *Y* through *M*. The total effect, however, denotes the overall effect of *X* on *Y*.

Figures 1 and 2 illustrate these concepts with the help of the free gift example introduced above. In this example, the total effect refers to the overall effect of the large (vs. small) free gift on positive WOM intentions without taking into account a possible mediator. The indirect effect denotes the effect of the large (vs. small) free gift on positive WOM intentions that is transmitted through hedonic benefits. It is hypothesized to be positive, because a large (vs. small) free gift provides hedonic benefits to consumers which, in turn, are associated with increased positive WOM intentions. The direct effect denotes the effect of the large (vs. small) free gift on positive WOM intentions which is not transmitted through hedonic benefits.

Statistical Description of the Single Mediator Model

The statistical diagrams (Hayes 2018) for the total effect and the single mediator model are depicted in Fig. 3. The diagrams can be described by a set of linear equations.

The total effect (see panel A in Fig. 3) is quantified by

$$Y = i_y + cX + e_y \tag{1}$$

where *i_y* denotes the intercept and *c* the effect of *X* on *Y*. The error term of *Y* is denoted as *e_y*.

To describe the single mediator model (see panel B in Fig. 3), two equations are necessary, one predicting *M* and the other predicting *Y*:

$$M = i_m + aX + e_m \tag{2}$$

$$Y = i_y + bM + c'X + e_y \tag{3}$$

The *i* parameters in Eqs. 2 and 3 denote the intercepts, *a* estimates the effect of *X* on *M*, *b* estimates the effect of *M* on *Y* controlling for *X*, and *c'* estimates the effect of *X* on *Y* controlling for *M*. The error terms are denoted by the *e* parameters, respectively. Note that *i_y* as well as *e_y* in Eqs. 1 and 3 are not equivalent.

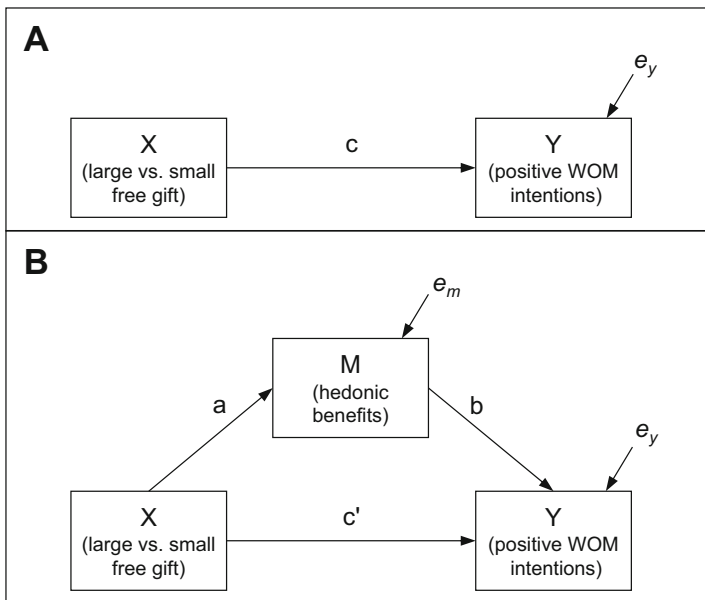


Fig. 3 The statistical diagrams of the total effect of *X* on *Y* (panel A), as well as the single mediator model depicting the direct as well as indirect effect of *X* on *Y* (panel B)

The indirect effect corresponds to the product of path *a* and path *b*, i.e., *ab*. The total effect equals the sum of the indirect effect and the direct effect (Eq. 4). Hence, the indirect effect can be also expressed as the difference between the total effect and the direct effect (Eq. 5):

$$\text{Total effect : } c = ab + c' \tag{4}$$

$$\text{Indirect effect : } ab = c - c' \tag{5}$$

The effects are usually reported as unstandardized regression coefficients, especially when *X* is dichotomous (as in the free gift example), as standardized coefficients are not meaningful in this case (Hayes 2018).

In the previously introduced free gift example, *X* is coded with “0” (small free gift) and “1” (large free gift). Hence, a total effect of $c = 0.218$ denotes that in comparison to a small free gift, a large free gift increases positive WOM intentions by 0.218 units. An indirect effect of $ab = 0.103$ indicates that a large (vs. a small) free gift increases positive WOM intentions by 0.103 units through its effect on hedonic benefits which, in turn, is associated with positive WOM intentions. A direct effect of $c' = 0.115$ shows that controlling for the indirect effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits, a large (vs. small) free gift increases positive WOM intentions by 0.115 units. As can be seen, the total effect $c = 0.218$ is the sum of the indirect effect $ab = 0.103$ and the direct effect $c' = 0.115$, as $0.103 + 0.115 = 0.218$.

Please note that the direct effect c' does not have to be smaller (i.e., closer to zero) than the total effect c even though the total effect c is the sum of both, the indirect effect ab and the direct effect c' . There are situations when the total effect c is equal in size to the direct effect c' or even smaller, for instance, if M acts as a suppressor (MacKinnon et al. 2000).

Statistical Inference for the Single Mediator Model

Various approaches have been suggested to determine whether or not mediation occurs, that is, whether the indirect effect ab is significantly different from zero (for an overview, see Hayes and Scharkow 2013; MacKinnon et al. 2002). MacKinnon et al. (2002) classify these according to the coefficients tested within the respective approaches: While the *causal steps approach* (Baron and Kenny 1986; Judd and Kenny 1981) establishes mediation through testing whether the individual paths in a mediation model are significantly different from zero, the *difference in coefficients approach* and the *product of coefficients approach* test whether the indirect effect (as indicated by the difference $c - c'$ or the product ab) is different from zero. Each approach will be described subsequently.

The causal steps approach consists of a series of regressions run separately and sequentially in order to demonstrate mediation. Mediation is logically inferred if, first, the total effect c is significant, second, path a is significant, third, path b is significant, and, fourth, the direct effect c' is not significant (Judd and Kenny 1981) or considerably reduced in size as compared to the total effect c (Baron and Kenny 1986).

The causal steps approach has been extraordinarily influential. The paper illustrating it – Baron and Kenny (1986) – is one of the most cited papers in the social sciences. However, it has also been criticized heavily (e.g., Hayes 2009; MacKinnon et al. 2000; Rucker et al. 2011; Shrout and Bolger 2002; Zhao et al. 2010). One major criticism is that the method requires a significant total effect to establish mediation (“an effect to be mediated”). It has been argued that there are situations in which significant mediation occurs even though the total effect of X on Y is not significant. Hence, requiring an effect to be mediated impairs the power of the approach (MacKinnon et al. 2002). Other criticism refers to the fact that testing the individual paths in the mediation model separately is statistically as well as conceptually different from directly testing the indirect effect: Testing the constituent paths answers the question whether they, considered individually, are different from zero. Testing the indirect effect, however, answers the question whether the indirect effect as a whole is different from zero. It might be intuitive to assume that, if the paths comprising the indirect effect are significant, the indirect effect must also be significant. However, this is not necessarily the case. Hence, to establish mediation, the focus should be on testing the indirect effect.

As opposed to the causal-steps approach, the difference in coefficients approach and the product of coefficients approach directly test the indirect effect. While the difference in coefficients approach does so using the right side of Eq. 5, $c - c'$, the product of coefficients approach refers to the left side of Eq. 5, ab . The indirect effect can be tested through computing its standard error which can then be used to create a

test statistic or confidence interval. The indirect effect is assumed to be significantly different from zero if the value of the test statistic exceeds some critical value of the normal distribution, or if the confidence interval around the indirect effect excludes zero. The latter approach is based on the following interpretation of a confidence interval: If a study were replicated many times and a confidence interval were computed around the indirect effect in each study, respectively, a large percentage of the confidence intervals obtained (e.g., 95% for a 95% confidence interval) would include the true value of the indirect effect. Hence, if zero is included in a confidence interval around the estimated indirect effect, zero is a fairly plausible value for the true indirect effect. If zero is not included, however, zero is a rather implausible value.

Several different methods have been suggested to estimate the standard error for both expressions, ab as well as $c-c'$ (MacKinnon et al. 2002). The most prominent of those methods is probably the one suggested by Sobel (1982), which is based on the *delta method* and used in the so-called *Sobel test*. With se_a and se_b being the standard errors of the coefficients a and b , respectively, to estimate the standard error of the indirect effect ab , Sobel (1982) suggests the following (first order delta estimator):

$$se_{ab} = \sqrt{a^2 se_b^2 + b^2 se_a^2} \quad (6)$$

The test statistic Z used in the Sobel test is then

$$Z = \frac{ab}{se_{ab}} \quad (7)$$

If Z exceeds the critical value of the z -distribution (for a two-sided test and $\alpha = 0.05$, $z_{\text{crit}} = \pm 1.96$), the indirect effect ab is assumed to be significantly different from zero. For an overview of other methods to compute standard errors and corresponding test statistics of ab and $c-c'$, see MacKinnon et al. (2002).

The values necessary to compute the Sobel test can be obtained through two ordinary least squares (OLS) regressions fitting the linear equations illustrated in Eqs. 2 and 3. In the free gift example, these regressions yield the following estimates: $a = 0.135$, $b = 0.767$, $se_a = 0.049$, and $se_b = 0.079$. Consequently, se_{ab} can be estimated as $\sqrt{0.135^2 \times 0.079^2 + 0.767^2 \times 0.049^2} = 0.0391$ (see Eq. 6). The test statistic Z is computed as $\frac{0.135 \times 0.767}{0.0391} = 2.650$ (see Eq. 7). As $Z > z_{\text{crit}}$ ($z_{\text{crit}} = 1.96$), it can be concluded that there is a significant and positive indirect effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits. That is, as compared to consumers receiving a small free gift, the consumers receiving a large free gift perceive greater hedonic benefits ($a = 0.135$), which, in turn, is associated with increased intentions to positively talk about and recommend the promoted products ($b = 0.767$). Note that one would obtain the same indirect effect if the regression coefficients denoting path a and path b were negative. Hence, although not being the focus when testing ab , the individual path coefficients a and b should by no means be ignored.

One issue with the Sobel test is that inference is based on the assumption that the product of two normally distributed random variables (e.g., regression coefficients) is normally distributed as well. That means, it is assumed that the sampling distribution of the indirect effect ab is normal. However, the sampling distribution of the indirect effect differs from the standard normal distribution with regard to its skewness and kurtosis, especially in small samples (e.g., Bollen and Stine 1990; Kisbu-Sakarya et al. 2014). Hence, relying on the assumption that the sampling distribution of the indirect effect is normal may lead to false conclusions about mediation (MacKinnon et al. 2002, 2004).

Taking this concern into account, MacKinnon and colleagues (e.g., MacKinnon et al. 2007a) developed approaches to compute se_{ab} that are based on the assumption that the indirect effect ab follows the distribution of the product of two normally distributed random variables. Simulation studies show that as compared to the normal theory approach, this *distribution of the product approach* leads to more accurate Type I error rates and higher statistical power to detect a possible indirect effect (MacKinnon et al. 2002), as well as to more precise confidence intervals (MacKinnon et al. 2004). However, the distribution of the product approach is not always easily applicable to more complicated mediation models (Taylor et al. 2008; Preacher and Hayes 2008).

Another set of methods establishes mediation by creating confidence intervals around the indirect effect through resampling (Monte Carlo resampling, Preacher and Selig 2012; jackknife resampling, MacKinnon et al. 2004; bootstrapping, Bollen and Stine 1990). Of these, the presumably most commonly applied method in consumer science is *bootstrapping* (Pieters 2017). The particularly convenient characteristic of bootstrapping is that, unlike the previously presented approaches, it does not require any assumptions to be made about the distribution of the indirect effect, nor does it rely on an estimate of the standard error of the indirect effect. It is a resampling procedure, which means that the distribution of the indirect effect is empirically obtained, that is, obtained from the data itself. To do so, k bootstrap samples ($k_{\min} = 1,000$, Shrout and Bolger 2002; $k_{\max} = 10,000$, Hayes 2018; $k_{\text{recommended}} > 5,000$, Hayes 2018) with N cases each are drawn with replacement from the original dataset (N denotes the original sample size). From each bootstrap sample k , the indirect effect ab_k is estimated. As the bootstrap sample is drawn with replacement (i.e., a participant from the original sample may be selected not at all, once, or multiple times in a bootstrap sample), the individual bootstrap samples will not only differ from the original sample, but also from each other. Consequently, the k estimates of the indirect effect vary. Sorting them from smallest to largest creates a distribution of the indirect effect. This distribution can then be used to compute a confidence interval around the point estimate of ab estimated from the original sample.

Figure 4 depicts such an empirically obtained distribution of the indirect effect from the free gift example. It is based on $k = 5,000$ bootstrap samples. The point estimate for the indirect effect from the original sample is $ab = 0.1035$. Three types of bootstrap confidence intervals are depicted: the percentile bootstrap confidence interval, the bias-corrected bootstrap confidence interval, and the bias-corrected and accelerated bootstrap confidence interval.

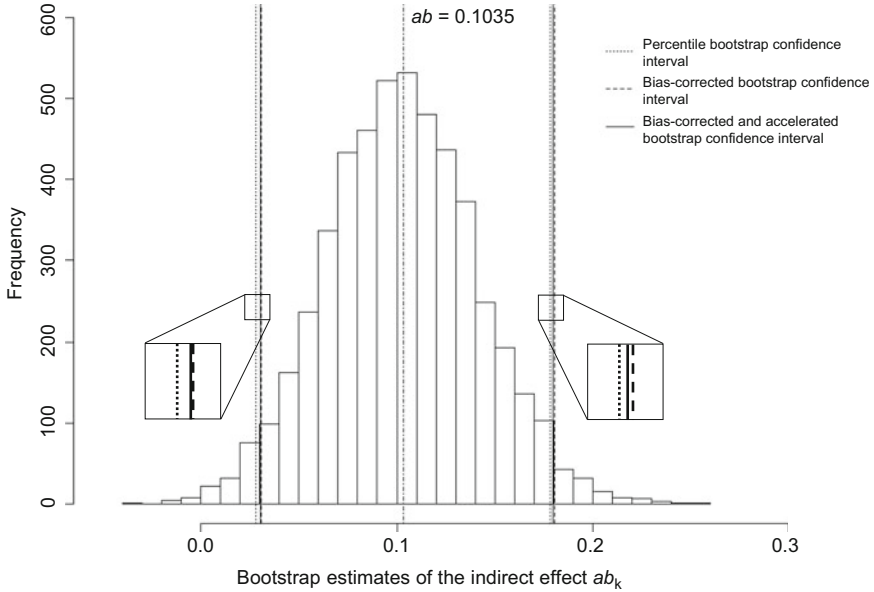


Fig. 4 Distribution of 5000 bootstrap estimates of an indirect effect ($ab = 0.1035$) and the corresponding percentile, bias-corrected, and bias-corrected and accelerated bootstrap confidence intervals (p.bci_{95%} [0.0281; 0.1781], bc.bci_{95%} [0.0310; 0.1805], bca.bci_{95%} [0.304; 0.1797])

Indicated by the two dotted lines are the lower and upper limits of the percentile bootstrap confidence interval (p.bci), which denotes the values at position $k \times (\frac{\alpha}{2})$ (lower limit, ll) and $k \times (1 - \frac{\alpha}{2}) + 1$ (upper limit, ul). Setting $\alpha = 0.05$, the lower limit of the percentile bootstrap confidence interval (p.bci.ll) corresponds to $5,000 \times (\frac{0.05}{2}) = 125$, that is, the 125th value in the sorted distribution (p.bci.ll = 0.0281 in Fig. 4); the upper limit of the percentile bootstrap interval (p.bci.ul) corresponds to $5,000 \times (1 - \frac{0.05}{2}) + 1 = 4,876$, that is, the 4876th value (p.bci.ul = 0.1781 in Fig. 4). Note that due to the skewness of the distribution, bootstrap confidence intervals are usually asymmetric unlike confidence intervals based on the standard normal distribution. As the confidence interval does not include zero (p.bci_{95%} [0.0281; 0.1781]), zero is not a plausible value for the indirect effect. Hence, the indirect effect is assumed to be significantly different from zero.

In this example, testing the indirect effect with a percentile bootstrap confidence interval leads to the same conclusion as the Sobel test: The effect of the large (vs. small) free gift on positive WOM intentions is significantly mediated by hedonic benefits ($ab = 0.1035$, p.bci_{95%} [0.0281; 0.1781]). This reflects the notion that although there may be inconsistent results from different approaches to test mediation, they quite frequently agree (Hayes and Scharkow 2013).

As mentioned, Fig. 4 also depicts the bias-corrected bootstrap confidence interval (bc.bci, broken lines) and the bias-corrected and accelerated bootstrap confidence interval (bca.bci, solid lines, Efron 1987). The bias correction adjusts the confidence

limits for differences between the point estimate of the indirect effect from the original data set ab and the bootstrap estimates of the indirect effect ab_k (bc.bci_{95%} [0.0310; 0.1805]). The bias-corrected and accelerated bootstrap confidence interval additionally accounts for the skew of the bootstrapped distribution (bca.bci_{95%} [0.304; 0.1797]).

The bootstrap approach to mediation analysis has some disadvantages: It requires raw data, which may not always be available. In addition, as the resampling process is random, the limits of the confidence interval may slightly differ when the analysis is repeated. Finally, statistical software has to be set up to perform bootstrapping. However, statistical software generally allows one to save bootstrap estimates for further analysis or to specify a seed to replicate the bootstrap samples. Furthermore, there are an increasing number of software options to perform mediation analysis with bootstrapping.

Most importantly, though, bootstrapping has considerable statistical advantages over normal theory based approaches and the distribution of the product approach: As noted, it makes no assumptions about the sampling distribution of the indirect effect, but empirically determines it through resampling. As a consequence, bootstrap confidence intervals have been shown to be more accurate and perform better with regards to statistical power while maintaining reasonable Type I error rates (Fritz and MacKinnon 2007; Hayes and Scharkow 2013; MacKinnon et al. 2004). The bias-corrected confidence interval has been demonstrated to perform best with regard to power, although somewhat liberally under some conditions (Fritz et al. 2012). Researchers reluctant to take this risk may use the percentile bootstrap confidence interval which is more powerful than the Sobel test, but less liberal than the bias-corrected bootstrap confidence interval. Furthermore, bootstrapping is relatively easy to apply to more complex mediation models (Taylor et al. 2008; Williams and MacKinnon 2008). Hence, in accordance with many others (e.g., Hayes 2018; MacKinnon 2008; Jose 2013), we recommend bootstrapping confidence intervals for testing mediation over normal theory based methods and the distribution of the product approach.

Assumptions of the Single Mediator Model

OLS regression-based mediation analysis relies on assumptions that apply to OLS regression analysis in general (see chapter ► “Regression Analysis” by Skiera et al., this volume). Some of these assumptions are particularly crucial in mediation analysis. For instance, it has been shown that measurement error can heavily bias estimates of the indirect effect, especially if it affects the mediator. While this can lead to either overestimation or underestimation of effects in a mediator model, in experimental settings, measurement error affecting the mediator tends to lead to underestimation of the indirect effect (Fritz et al. 2016). Mediation analysis with structural equation modeling accounts for this issue to some degree as it enables measurement error to be estimated (MacKinnon 2008; see also ► “Crafting Survey Research: A Systematic Process for Conducting Survey Research,” this volume).

Omitting causally relevant variables in mediation analysis can similarly bias the estimate of the indirect effect (see also chapter ► “Dealing with Endogeneity: A

[Nontechnical Guide for Marketing Researchers](#)” by Ebbes et al., this volume). Pieters (2017) differs between bias as a result of pre- and posttreatment confounding variables. Pretreatment confounding variables are independent of the treatment, that is, the independent variable, but affect the mediator and the dependent variable. One example of pretreatment confounding variables pointed out by Pieters (2017) is common method bias, that is, common variance in the mediator and the dependent variable due to being measured in a similar way or in close proximity to each other. Posttreatment confounding variables are consequences of the treatment and affect the dependent variable. Hence, they are omitted mediators. In experimental studies, the omission of a confounding variable affecting the mediator and the dependent variable likely leads to an overestimation of the indirect effect (Fritz et al. 2016). Pieters (2017) concludes that bias resulting from confounding variables can be addressed through employing different methods to measure the mediator and the dependent variable, the inclusion of possible confounding variables as covariates in the mediation model, the use of advanced statistical methods that account for the influence of unobserved confounding variables on mediation (for an overview, see MacKinnon and Pirlott 2015, see also chapter ▶ [“Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)” by Ebbes et al., this volume), and by running studies in which the mediator is manipulated (for an overview, see Pirlott and MacKinnon 2016).

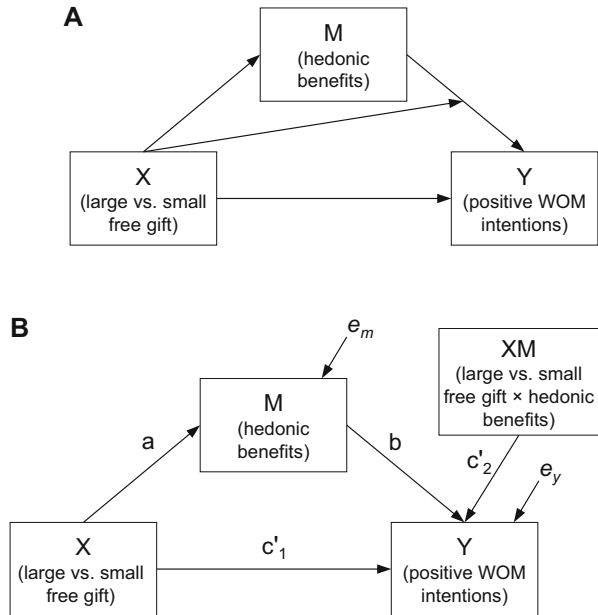
To address violations of the assumption of homoskedasticity, robust regression analysis (e.g., Hayes and Cai 2007) could be applied which corrects for possible bias in the standard errors of the regression coefficients. However, note that bootstrapping does not rely on estimates of the standard errors of the regression coefficients comprising the indirect effect, but solely on estimates of the regression coefficients which are unaffected by heteroskedasticity (Darlington and Hayes 2017). Hence, the bootstrap approach to testing mediation is generally robust against violations of the assumption of homoskedasticity.

Another assumption in mediation analysis requires that the causal order of variables be correctly specified (i.e., the causal order is assumed to be unidirectional). To test this assumption, it would be necessary to demonstrate that X causes M , which in turn causes Y . The section on causal inference in this chapter will address this issue in greater detail.

Furthermore, it is generally assumed that X and M do not interact to predict Y . Hence, the interaction between X and M is not included in the linear equation predicting Y (see Eq. 3, but see also Fig. 5 for a model including this interaction). That means, neither is the direct effect of X on Y assumed to be affected by M , nor is the effect of M on Y assumed to be affected by X . However, it has been argued that this is not justified and that the interaction between X and M should be estimated in mediation analysis (e.g., Kraemer et al. 2002, 2008; Valeri and VanderWeele 2013), for instance, to test whether mediation differs across levels of X . The questions of when to estimate the interaction term XM in mediation analysis and how to estimate the direct and indirect effects in such a case are addressed by VanderWeele (2015).

Additional assumptions in mediation analysis refer to the correct timing and level of the mediated effect (MacKinnon 2008). Specifically, conclusions based on a

Fig. 5 Conceptual (panel A) and statistical diagram (panel B) of a single mediator model including the interaction between the independent variable X and the mediator M



single (as compared to repeated) assessment of X , M , or Y assume that the variables and relationships of interest do not change over time. In addition, inferring mediation without taking into account possible nesting of the data (e.g., consumers nested in geographical locations or repeated measurements nested in one participant) relies on the assumption that mediation is unaffected by such nesting.

Classifying Mediation

Mediation can be classified depending on the significance of the indirect and the direct effect. Zhao et al. (2010) broadly differentiate between *mediation* (if the indirect effect is significant) and *nonmediation* (if the indirect effect is not significant), and further distinguish different types of mediation and nonmediation depending on the significance of the direct effect as well as whether the direct effect and indirect effect have the same sign (see Table 1). Competitive mediation has also been referred to as inconsistent mediation (MacKinnon et al. 2000) or suppression (Shrout and Bolger 2002). Baron and Kenny (1986) refer to complementary mediation as partial mediation and to indirect-only mediation as complete mediation, for which the attributes “perfect” or “full” are also used. According to Zhao et al. (2010), different types of mediations give hints for subsequent theory building. Specifically, complementary and competitive mediation as well as direct-only nonmediation may indicate that a relevant mediator or moderator may have been omitted from the mediation model. However, when drawing theoretical conclusions or classifying mediation based on the significance of the direct effect, note that a nonsignificant direct effect may also be the result of too little power (Rucker et al. 2011). Furthermore, mind that qualitatively classifying mediation as, for example, partial,

Table 1 Classification of mediation according to Zhao et al. (2010)

Classification		1. Is the indirect effect significant?	2. Is the direct effect significant?	3. Do direct and indirect effects have equal signs?
Mediation	Complementary mediation	Yes	Yes	Yes
	Competitive mediation	Yes	Yes	No
	Indirect-only mediation	Yes	No	–
Non-mediation	Direct-only nonmediation	No	Yes	–
	No-effect nonmediation	No	No	–

complete, or competitive, does not allow conclusions to be drawn about the magnitude of an indirect effect.

Effect Size

To quantify mediation, several effect size measures have been proposed with a comprehensive overview provided by Preacher and Kelley (2011). However, none of the measures is without limitations. For instance, effect size measures and their corresponding variance estimates may be inaccurate unless sample size or effect size is large (e.g., $N > 500$, MacKinnon et al. 1995), may not be applicable to specific variable metrics (e.g., exclusively suitable for dichotomous X , Hansen and MacNeal 1996), or may not work in any mediation model more complex than the single mediator model (e.g., Wen and Fan 2015). Hence, we agree with Preacher and Kelley (2011) that, when reported, effect size measures have to be carefully discussed with regard to whether they are bounded (i.e., whether there is an upper and lower limit of possible values of the measure), robust to changes in scales (i.e., standardized) as well as sample size, precise (as indicated by, e.g., a confidence interval), and meaningfully scaled.

Variable Metrics

OLS regression-based mediation analysis can incorporate dichotomous, multicategorical, or continuous independent variables. A multicategorical independent variable can be included as a set of indicator variables, each representing, for example, a pairwise comparison with a reference group (Hayes and Preacher 2014). Different strategies to create indicator variables will briefly be discussed in the section on coding of categorical independent variables later in the chapter. Including a dichotomous or multicategorical mediator or dependent variable goes beyond what OLS regression-based mediation analysis can accommodate. Yet, the equations presented above can be rewritten for logit models and logistic regressions (Iacobucci 2012; MacKinnon 2008). Note, that in this case, Eq. 5 does not hold anymore, as the difference in coefficients $c - c'$ may be biased due to scale boundness (MacKinnon and Dwyer 1993). Applying generalized linear models to

mediation analysis further allows for the analysis of mediation models with mediators and outcomes taking the form of counts (e.g., VanderWeele and Vansteelandt 2014) or survival rates (e.g., VanderWeele 2015). Additionally, several approaches to incorporate nonnormally distributed but continuous variables into mediation analysis have been described (e.g., Yuan and MacKinnon 2014).

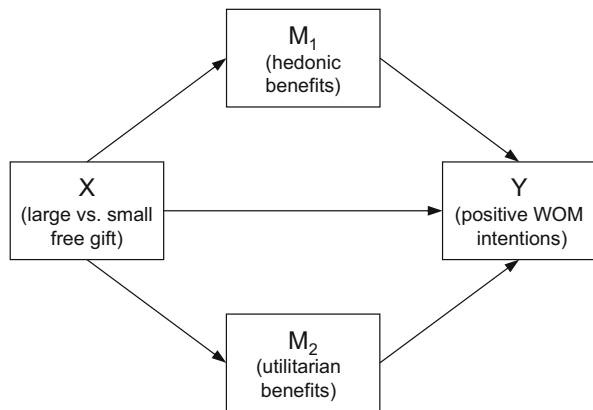
Mediation Models Including More Than One Mediator: The Parallel and Serial Multiple Mediator Model

Some research questions require inclusion of more than one mediator into the conceptual model. It could be of interest, for instance, to pit competing explanations for an effect of X on Y against each other or to test whether an effect of X on Y operates through multiple mechanisms at once. The single mediator model can be extended to include multiple mediators in two different ways: They can be assumed to be causally independent of each other and work in parallel, or form a causal chain from X to Y and operate in serial. The former is referred to as the *parallel multiple mediator model* and will be described next. The latter is referred to as the *serial multiple mediator model* and will be described subsequently. General notes on statistical inference in multiple mediator models are included at the end of this section on mediation models with more than one mediator.

Conceptual Description of the Parallel Multiple Mediator Model

In the parallel multiple mediator model, X is assumed to affect Y through two or more mediators M_i , which are assumed *not* to be causally related (see Fig. 6). Yet, they are also not expected to be completely uncorrelated as they share a common cause. Parallel multiple mediator models are hence particularly suited to disentangle the respective mediating ability of multiple (ideally not too strongly) correlated mediators from each other (Preacher and Hayes 2008).

Fig. 6 The conceptual diagram of a parallel multiple mediator model in which the effect of X on Y is transmitted through two mediators M_1 and M_2 operating in parallel



For instance, in the free gift example, it might be of interest to investigate whether the increase in positive WOM intentions in response to the large (vs. small) free gift is solely explained by hedonic benefits or also by utilitarian benefits (e.g., making a purchase decision more efficient because of reduced search costs). In order to investigate these alternative explanations, a second mediator (utilitarian benefits) is added to the previously tested single mediator model, which included hedonic benefits as the only mediator. Furthermore, since it is not assumed that there is a causal relationship between hedonic and utilitarian benefits, the two mediators are hypothesized to act in parallel.

Just like the single mediator model, the parallel multiple mediator model can be separated into indirect and direct components. However, there are multiple indirect components in parallel multiple mediator models, namely, the *specific indirect effects* $a_i b_i$ associated with each mediator M_i , respectively. The crucial characteristic of the specific indirect effect in a multiple mediator model compared to the indirect effect in a single mediator model is that it estimates the specific mediating ability of a mediator while controlling for the remaining specific indirect effects of all other mediators included in the model. Consequently, specific indirect effects are affected by the degree to which the mediators in a parallel multiple mediator model conceptually overlap (i.e., correlate). The specific indirect effects in a parallel multiple mediator model sum up to form the *total indirect effect*, which denotes the ability of a set of mediators to transmit an effect from X on Y . The direct effect denotes the remaining effect of X on Y , controlling for the total indirect effect. Thus, together, the total indirect effect and the direct effect add up to the total effect of X on Y .

Preacher and Hayes (2008) describe several advantages of testing one parallel multiple mediator model instead of multiple single mediator models. First, by testing the total indirect effect, a parallel multiple mediator model allows conclusions to be drawn regarding a set of multiple mediators. Second, disentangling the mediating ability of each mediator enables researchers to identify the specific indirect effect of each mediator as well as to quantitatively compare the specific indirect effects of the different mediators with each other. Third, the parallel multiple mediator model partially accounts for the limitation of the single mediator model with regard to possible bias due to omitted variables.

Statistical Description of the Parallel Multiple Mediator Model

The parallel multiple mediator model with two mediators M_1 and M_2 can be described by the following linear equations predicting M_1 , M_2 , and Y , respectively (see also Fig. 7):

$$M_1 = i_{m1} + a_1 X + e_{m1} \quad (8)$$

$$M_2 = i_{m2} + a_2 X + e_{m2} \quad (9)$$

$$Y = i_y + b_1 M_1 + b_2 M_2 + c' X + e_y \quad (10)$$

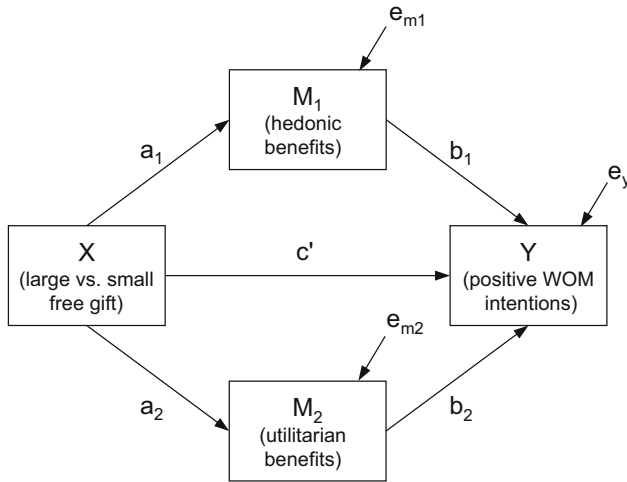


Fig. 7 The statistical diagram of a parallel multiple mediator model in which the effect of X on Y is transmitted through two mediators M_1 and M_2 operating in parallel

The i parameters denote the intercepts and a_i estimates the effect of X on M_i . The coefficient b_1 estimates the effect of M_1 on Y , controlling for X and M_2 . Likewise, b_2 estimates the effect of M_2 on Y , controlling for X and M_1 . Finally, c' estimates the effect of X on Y , controlling for M_1 and M_2 . The error terms are denoted by the respective e parameters. In a parallel multiple mediator model with j mediators, $j + 1$ equations are required to describe the model, one to predict each M_i and one to predict Y .

Analogous to the single mediator model, a specific indirect effect through a mediator M_i is the product of the two unstandardized regression coefficients of path a_i and path b_i , $a_i b_i$. Therefore, in a parallel multiple mediator model with j mediators, there are j specific indirect effects to be estimated. The total indirect effect is the sum of all j specific indirect effects:

$$\text{Total indirect effect} : \sum_{i=1}^j (a_i b_i) \tag{11}$$

The total effect is the sum of the total indirect effect and the direct effect c' :

$$\text{Total effect} : c = \sum_{i=1}^j (a_i b_i) + c' \tag{12}$$

Again, the total indirect effect can also be computed by subtracting the direct effect from the total effect as follows:

$$\text{Total indirect effect : } c - c' = \sum_{i=1}^j (a_i b_i) \quad (13)$$

In the parallel multiple mediator model sketched out above, a specific indirect effect through the first mediator (hedonic benefits) of $a_1 b_1 = 0.103$ indicates that controlling for the specific indirect effect of the large (vs. small) free gift on positive WOM intentions through utilitarian benefits, the large (vs. small) free gift increases positive WOM intentions through hedonic benefits by 0.103 units. Similarly, a specific indirect effect through the second mediator (utilitarian benefits) of $a_2 b_2 = 0.005$ denotes that controlling for the specific indirect effect of the large (vs. small) free gift on positive WOM intentions through hedonic benefits, the large (vs. small) free gift increases positive WOM intentions through utilitarian benefits by 0.005 units. The total indirect effect of $a_1 b_1 + a_2 b_2 = 0.103 + 0.005 = 0.108$ shows that together, changes in hedonic and utilitarian benefits in response to a large (vs. small) free gift account for a 0.108 unit increase in positive WOM intentions. A direct effect of $c' = 0.110$ indicates that independent of the total indirect effect of the large (vs. small) free gift on positive WOM intentions through hedonic as well as utilitarian benefits, the large (vs. small) free gift increases positive WOM intentions by 0.110 units. As can be seen, the total effect $c = 0.218$ is the sum of the specific indirect effect through the first mediator, $a_1 b_1 = 0.103$, the specific indirect effect through the second mediator, $a_2 b_2 = 0.005$, and the direct effect $c' = 0.110$, as $0.103 + 0.005 + 0.110 = 0.218$.

Statistical Inference for the Parallel Multiple Mediator Model

Testing for mediation in the parallel multiple mediator model involves testing the total as well as the specific indirect effects. As with the single mediator model, this can be accomplished with the help of bootstrap confidence intervals. There are also other suitable approaches; however, they do not perform as well as the bootstrap approach (Preacher and Hayes 2008; Williams and MacKinnon 2008). Note that specific indirect effects should be investigated irrespective of whether the total indirect effect is significant or not (Preacher and Hayes 2008).

In the free gift example, the bias-corrected bootstrap confidence interval for the specific indirect effect through hedonic benefits lies completely above zero. Hence, the effect is positive and significant ($a_1 b_1 = 0.103$, bc.bci_{95%} [0.029; 0.181]). However, the bias-corrected bootstrap confidence interval for the specific indirect effect through utilitarian benefits straddles zero and is hence not significant ($a_2 b_2 = 0.005$, bc.bci_{95%} [-0.014; 0.029]). The total indirect effect is positive and significant too ($a_1 b_1 + a_2 b_2 = 0.108$, bc.bci_{95%} [0.031; 0.189]). That is, as compared to consumers receiving a small free gift, consumers receiving a large free gift perceive greater hedonic benefits ($a_1 = 0.135$), which, in turn, is associated with increased intentions to positively talk about and recommend the promoted products ($b_1 = 0.766$, $a_1 b_1 = 0.103$, bc.bci_{95%} [0.029; 0.181]). Furthermore, hedonic and utilitarian benefits jointly mediate the effect of a large (vs. small) free gift on positive WOM intentions ($a_1 b_1 + a_2 b_2 = 0.108$, bc.bci_{95%} [0.031; 0.189]).

In parallel multiple mediator models, it is also possible to address the question of whether specific indirect effects differ from each other. For instance, a researcher

may be interested in finding out whether one specific indirect effect is larger than another one. This, too, can be achieved with bootstrapping. The idea is straightforward: If two specific indirect effects $a_i b_i$ and $a_j b_j$ significantly differ from each other, their difference must be different from zero. To test this, the distribution of $a_i b_i - a_j b_j$ is bootstrapped, and a confidence interval is determined. If it excludes zero, the specific indirect effects in question differ significantly from each other. Note, however, that the conclusion that one specific indirect effect is larger than the other can only be drawn if both specific indirect effects compared have the same sign (i.e., are both positive or both negative, Preacher and Hayes 2008). To contrast specific indirect effects with different signs, the difference in absolute values may be determined and tested analogously with the help of a bootstrap confidence interval (Hayes 2018). Furthermore, note that a (specific) indirect effect is scaled in the metrics of X and Y : A change in X by one unit leads to a change in Y through M of ab units. Hence it does not matter whether the same response scales (e.g., 7-point vs. 5-point) are used to assess the respective mediators when comparing two specific indirect effects (Preacher and Hayes 2008).

Conceptual Description of the Serial Multiple Mediator Model

In a serial multiple mediator model, X is assumed to affect Y through two or more mediators M_i . However, in contrast to the parallel multiple mediator model, the mediators in a serial multiple mediator model are hypothesized to form a causal chain (see Fig. 8). Thus, serial multiple mediator models are suitable when a causal chain of mediators is assumed to account for the effect of X on Y . Research investigating serial multiple mediator models may be less common than research on parallel multiple mediator models. Yet, it is frequently possible to assume that mediators are part of a longer causal chain.

In the free gift example, one could hypothesize, for instance, that the effect of a large (vs. small) free gift on positive WOM intentions is in fact the result of an immediate positive emotional response to encountering a sales promotion which is then used as a basis for judging the hedonic benefits provided by the specific promotion. That is, one could argue that a causal chain consisting of, first, a positive emotional response and, second, perceived hedonic benefits, transmits the effect of the large (vs. small) free gift on positive WOM intentions.

As in the case of the parallel multiple mediator model, the serial multiple mediator model can be divided into total and specific indirect and direct components. The total

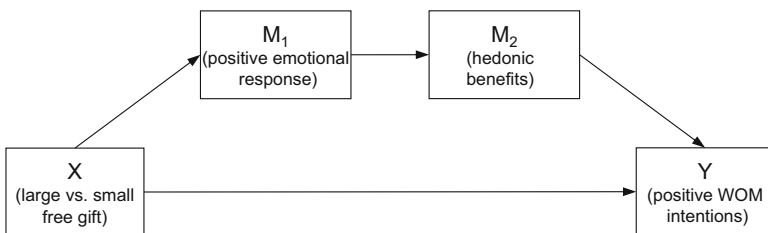


Fig. 8 The conceptual diagram of a serial multiple mediator model in which the effect of X on Y is transmitted through two mediators M_1 and M_2 operating in serial

indirect effect is the sum of the specific indirect effects of X on Y through the respective mediators, considered individually and in sequence (see Fig. 9). The direct effect denotes the remaining effect of X on Y after controlling for the total indirect effect. Together, the direct effect and the total indirect effect add up to the total effect of X on Y .

Statistical Description of the Serial Multiple Mediator Model

Analogous to the previously presented mediation models, the serial multiple mediator model with two mediators M_1 and M_2 (see Fig. 9) can be described by the following linear equations predicting M_1 , M_2 , and Y :

$$M_1 = i_{m1} + a_1X + e_{m1} \tag{14}$$

$$M_2 = i_{m2} + a_2X + d_{21}M_1 + e_{m2} \tag{15}$$

$$Y = i_y + b_1M_1 + b_2M_2 + c'X + e_y \tag{16}$$

The i parameters denote the intercepts, a_1 estimates the effect of X on M_1 , and a_2 estimates the effect of X on M_2 , controlling for the effect of M_1 on M_2 which is captured by d_{21} . The coefficient b_1 denotes the effect of M_1 on Y , controlling for X and M_2 . Likewise, b_2 denotes the effect of M_2 on Y , controlling for X and M_1 . The coefficient c' estimates the effect of X on Y , controlling for M_1 and M_2 . The error terms are denoted by the respective e parameters. Generally, in a serial multiple mediator model with j mediators, there are $j + 1$ equations required to describe the model, one to predict each M_i and one to predict Y .

In a serial multiple mediator model with two mediators, three specific indirect effects are to be estimated, one through each mediator (a_1b_1 and a_2b_2), and one through both mediators ($a_1d_{21}b_2$). The total indirect effect is the sum of the specific indirect effects:

$$\text{Total indirect effect} : a_1b_1 + a_2b_2 + a_1d_{21}b_2 \tag{17}$$

The total effect is the sum of the total indirect effect and the direct effect c' :

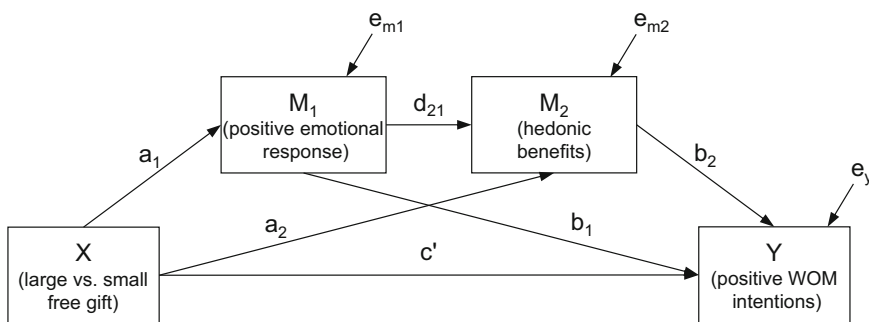


Fig. 9 The statistical diagram of a serial multiple mediator model in which the effect of X on Y is transmitted through two mediators M_1 and M_2 operating in serial

$$\text{Total effect : } c = a_1b_1 + a_2b_2 + a_1d_{21}b_2 + c' \quad (18)$$

Again, the total indirect effect can also be computed by subtracting the direct effect from the total effect:

$$\text{Total indirect effect : } c - c' = a_1b_1 + a_2b_2 + a_1d_{21}b_2 \quad (19)$$

These effects correspond to the following effects from the free gift example: A specific indirect effect through the first mediator (positive emotional response) of $a_1b_1 = 0.001$ indicates that controlling for all other indirect effects in the model, a large (vs. small) free gift increases positive WOM intentions through a positive emotional response by 0.001 units. Similarly, a specific indirect effect through the second mediator (hedonic benefits) of $a_2b_2 = 0.067$ denotes that controlling for all other indirect effects in the model, a large (vs. small) free gift increases positive WOM intentions through hedonic benefits by 0.067 units. Finally, a specific indirect effect through both mediators (first, the positive emotional response and then hedonic benefits) of $a_1d_{21}b_2 = 0.037$ denotes that a large (vs. small) free gift increases positive WOM intentions by 0.037 units sequentially through a positive emotional response which is associated with hedonic benefits which, in turn, is associated with positive WOM intentions. The total indirect effect of $a_1b_1 + a_2b_2 + a_1d_{21}b_2 = 0.105$ shows that together, all three specific indirect effects in the model account for a 0.105 unit increase in positive WOM intentions in response to the large (vs. small) free gift. A direct effect of $c' = 0.114$ indicates that controlling for all specific indirect effects in the model (i.e., the total indirect effect), a large (vs. small) free gift increases positive WOM intentions by 0.114 units. The total effect $c = 0.218$ is the sum of the specific indirect effect through the first mediator (positive emotional response), $a_1b_1 = 0.001$, the specific indirect effect through the second mediator (hedonic benefits), $a_2b_2 = 0.067$, the specific indirect effect through both mediators $a_1d_{21}b_2 = 0.037$, and the direct effect $c' = 0.114$, as $0.001 + 0.067 + 0.037 + 0.114 \approx 0.218$ (the difference is due to rounding).

Hayes (2018) notes that while it is relatively easy to incorporate more than two mediators in a parallel multiple mediator model, the number of specific indirect effects quickly increases in a model with more than two mediators operating in serial. This follows because there is one specific indirect effect through each mediator, one specific indirect effect through each combination of two mediators, one specific indirect effect through each combination of three mediators, and so on. Yet, the same relations apply in such models: The specific indirect effects comprise the total indirect effect and together with the direct effect they sum up to the total effect. Moreover, the total indirect effect can be estimated by subtracting the direct effect from the total effect.

Statistical Inference for the Serial Multiple Mediator Model

As for the parallel multiple mediator model, testing for mediation in the serial multiple mediator model involves testing the total indirect effect as well as the specific indirect effects. Again, there are several suitable methods to do so, yet bootstrap confidence intervals have been demonstrated to perform very well

(Taylor et al. 2008). In addition, as in the case of the parallel multiple mediator model, specific indirect effects should be investigated irrespective of whether or not the total indirect effect is significant (Hayes 2018).

In the free gift example, the results of the analysis are as follows: The bias-corrected bootstrap confidence intervals for the specific indirect effect of the large (vs. small) free gift on positive WOM intentions solely through the positive emotional response as well as the specific indirect effect solely through hedonic benefits include zero, that is, neither of the two effects is significant ($a_1b_1 = 0.001$, bc.bci_{95%} [-0.043; 0.046], $a_2b_2 = 0.067$, bc.bci_{95%} [-0.010; 0.145]). However, the specific indirect effect of the large (vs. small) free gift on positive WOM intentions through both mediators sequentially, the positive emotional response and hedonic benefits, is positive and significant ($a_1d_{21}b_2 = 0.037$, bc.bci_{95%} [0.012; 0.072]). The total indirect effect is positive and significant as well ($a_1b_1 + a_2b_2 + a_1d_{21}b_2 = 0.001 + 0.067 + 0.037 = 0.105$, bc.bci_{95%} [0.026; 0.192]). That is, as compared to consumers receiving a small free gift, consumers receiving a large free gift show a more positive emotional response to the promotion ($a_1 = 0.199$), which, in turn, is associated with greater perceived hedonic benefits ($d_{21} = 0.241$), and, as a consequence, positive WOM intentions are increased ($b_2 = 0.767$, $a_1d_{21}b_2 = 0.037$, bc.bci_{95%} [0.012; 0.072]). Furthermore, the positive and significant total indirect effect suggests that a large (vs. small) free gift increases positive WOM intentions through all three specific indirect effects at once ($a_1b_1 + a_2b_2 + a_1d_{21}b_2 = 0.105$, bc.bci_{95%} [0.026; 0.192]).

Analogous to the parallel multiple mediator model, specific indirect effects can also be contrasted in serial multiple mediator models. The logic is the same and, again, a significant difference between two specific indirect effects can only be interpreted as the one effect being larger than the other one if both specific indirect effects have the same sign. In case they do not, the strength of the two specific indirect effects may be compared by testing the difference between the absolute values of the effects.

How to Interpret Results from Multiple Mediator Models

The crucial difference between a multiple mediator model and a single mediator model lies in the number of mediators included in the model. As a consequence, the interpretation of an “indirect effect” in a multiple mediator model differs from the “indirect effect” in a single mediator model: In a single mediator model, the *indirect effect* denotes the ability of a mediator M to transmit the effect of X on Y . In a multiple mediator model, however, there are multiple indirect components: While a *specific indirect effect* quantifies the unique ability of a mediator M_i to transmit the effect of X on Y taking the specific indirect effects of X on Y through other mediating variables M_j into account, the *total indirect effect* denotes the ability of a set of mediators to transmit the effect of X on Y . These distinctions should be carefully considered when interpreting and comparing results from different single and multiple mediator models (Hayes 2018).

Mediation Models Including a Moderator: Conditional Process Models

The single mediator model can also be extended to include a moderator. Doing so allows to account for the fact that mediation may work differently, for example, for different people or under different circumstances. For instance, it could be of interest in the free gift example to investigate whether the indirect effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is the same across consumers differing in their general responsiveness to sales promotions. One might argue that mediation is especially large for consumers that are very deal-prone, that is, overall highly responsive to sales promotions (Lichtenstein et al. 1995), but smaller for consumers that are overall not particularly deal-prone. Moreover, one could hypothesize that as hedonic benefits are more likely to affect consumer judgment if consumers pursue a hedonic consumption motive as compared to a utilitarian consumption motive (Chandon et al. 2000), a large (vs. small) free gift should increase positive WOM intentions through hedonic benefits if the promoted product is associated with a hedonic consumption motive (e.g., a fashionable T-shirt by a high-end brand), but not if it is generally purchased out of utilitarian motives (e.g., a plain T-shirt to wear under a shirt).

Conceptual Description of Conditional Process Models

Mediation models with added moderators can generally be referred to as conditional process models (Hayes 2018), but sometimes, a conceptual distinction is made between *moderated mediation* and *mediated moderation* (e.g., Muller et al. 2005; Preacher et al. 2007). In the prototypical moderated mediation, the research focus lies on whether or not the mediation of X on Y through M is influenced by a moderating variable W , which may affect different paths in the conditional process model. The conceptual diagrams depicted in panel A and B in Fig. 10 are exemplary cases of moderated mediation. Excluding the broken line, panel A shows a conditional process model in which a moderator W affects path a , the effect of X on M . Including the broken line, the model assumes that W also moderates path c' , the direct effect of X on Y . These models correspond to the first research question mentioned above, namely, whether the effect of a large (vs. small) free gift on positive WOM intentions through perceived hedonic benefits is affected by consumers' deal proneness, as consumers' deal proneness could be argued to influence the effect of a large (vs. small) free gift on hedonic benefits (i.e., path a and possibly also path c').

Panel B shows a conditional process model in which a moderator V affects path b , the effect of M on Y (excluding the broken line), but also a mediation model in which path c' is additionally affected by V (including the broken line). These models correspond to the second question raised above, namely, whether the effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is affected by the consumption motive associated with the promoted product, as this moderator can be hypothesized to influence the effect of hedonic benefits on positive WOM intentions (i.e., path b and possibly also path c').

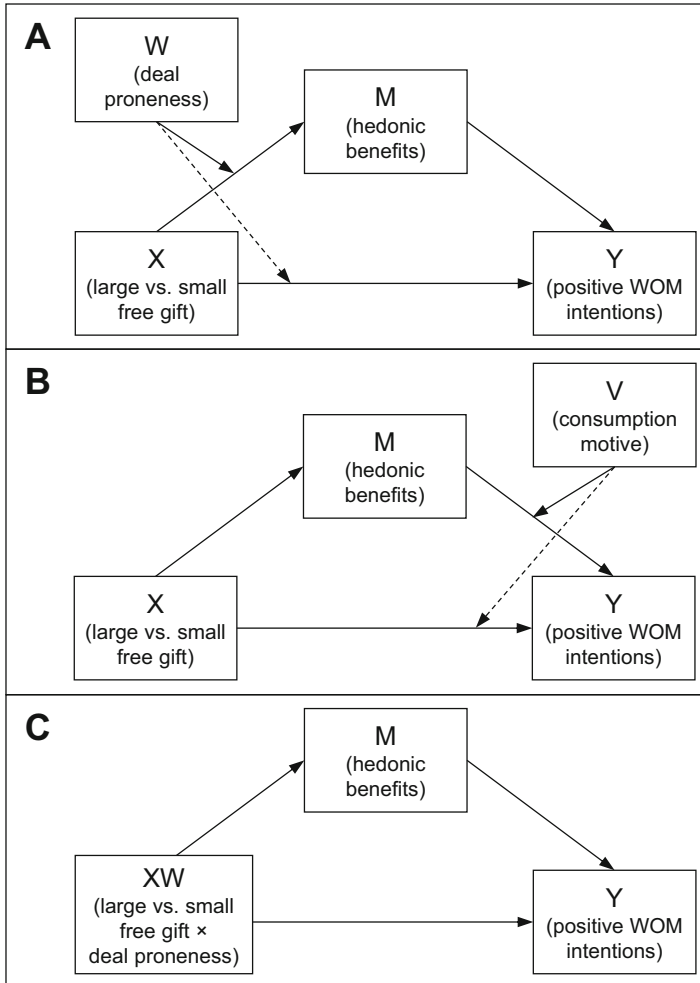


Fig. 10 Conceptual diagrams of exemplary conditional process models. While the model in panel A and panel B are examples of moderated mediation, the model in panel C corresponds to a case of mediated moderation

In the prototypical mediated moderation, the focus lies on whether or not the moderation of an effect of X on Y by W can be explained by a mediator. Thus, in mediated moderation, the effect of an interaction between two variables (i.e., XW) on Y is transferred through M (Hayes 2018). The respective conceptual diagram is depicted in panel C in Fig. 10. This would correspond to the question whether the effect of the interaction between the large (vs. small) free gift and consumers' deal proneness on positive WOM intentions can be explained by hedonic benefits. As it will become clearer in the following section, mediated moderation is a special case of moderated mediation. Hence, mediated moderation can be framed as moderated

mediation. Hayes (2018) even argues that mediated moderation should be framed as moderated mediation, as the product between two variables and thus its indirect effect may be difficult to interpret meaningfully.

Just like the previously presented mediation models, a conditional process model can be divided into direct and indirect components which, if qualified by a moderator, are referred to as *conditional indirect effect* and *conditional direct effect*. Together, the (conditional) direct effect and the conditional indirect effect may be added up to the total effect of X on Y .

Statistical Description of Conditional Process Models

In the following section, the three conditional process models depicted in Fig. 10 will be described statistically, starting with the conditional process model depicted in panel A in which a moderator W affects path a (see panel A in Fig. 10 as well as Fig. 11, excluding the broken lines). This conditional process model is described by the following equations:

$$M = i_m + a_1X + a_2W + a_3XW + e_m \tag{20}$$

$$Y = i_y + bM + c'_1X + e_y \tag{21}$$

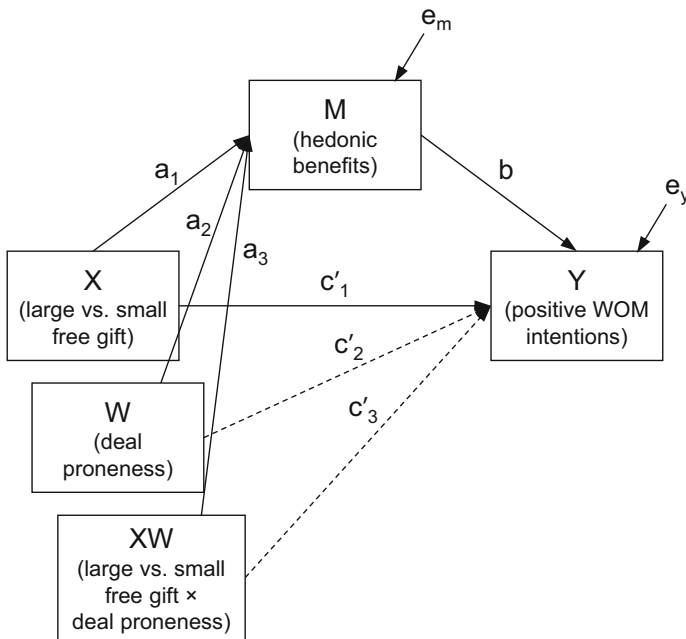


Fig. 11 The statistical diagram of the conditional process model depicted in panel A in Fig. 10 in which a moderator W affects path a (solid lines) as well as path c' (broken lines)

The i parameters in Eqs. 20 and 21 denote the intercepts, a_1 estimates the effect of X on M , a_2 the effect of W on M , and a_3 the effect of their interaction XW on M , respectively, controlling for the remaining variables in the equation. The effect of M on Y controlling for X is indicated by b and the effect of X on Y controlling for M is estimated by c' . The error terms are denoted by the respective e parameters. If the direct effect is conditional on W as well (see panel A in Fig. 10 as well as Fig. 11, including the broken lines), additional c' parameters would be added to Eq. 21 to denote the partial effects of W and XW on Y resulting in the following equation:

$$Y = i_y + bM + c'_1X + c'_2W + c'_3XW + e_y \tag{22}$$

Irrespective of whether or not a moderator W affects path c' , if W moderates path a , the effect of X on M , the conditional indirect effect is denoted as

$$\text{Conditional indirect effect : } \omega_{xw} = (a_1 + a_3W)b \tag{23}$$

Rewritten as

$$\omega_{xw} = a_1b + a_3bW \tag{24}$$

it becomes evident that the conditional indirect effect ω_{xw} is a linear function of the moderator W . That is, there is no single numeric estimate for the conditional indirect effect but many, depending on the value of W .

Hayes (2015) coined the term *index of moderated mediation* for the weight of the moderator a_3b in Eq. 24 which quantifies the linear dependency of the conditional indirect effect on the moderator W . Specifically, the index of moderated mediation denotes the difference between the conditional indirect effects of participants differing by one unit on the moderator W . Hence, if W is a dichotomous variable indicating two experimental groups and coded with “0” and “1,” the index of moderated mediation denotes the difference of the conditional indirect effects in those two groups (Hayes 2015).

The direct effect is denoted by c' if unconditional. If conditional on a moderator W (see broken lines in Fig. 11), it is denoted by

$$\text{Conditional direct effect : } c'_1 + c'_3W \tag{25}$$

which shows that if the direct effect is moderated by W , the remaining effect of X on Y controlling for the conditional indirect effect varies depending on values of W .

The total effect c is again the sum of the (conditional) direct effect and the conditional indirect effect. If W solely moderates the effect of X on M (Fig. 11, excluding the broken lines), the total effect is denoted by

$$\text{Total effect : } c = \omega_{xw} + c' = (a_1 + a_3W)b + c' \tag{26}$$

The linear dependency of a conditional indirect effect on a moderator described above is visualized in Fig. 12 which is based on results from the free gift example:

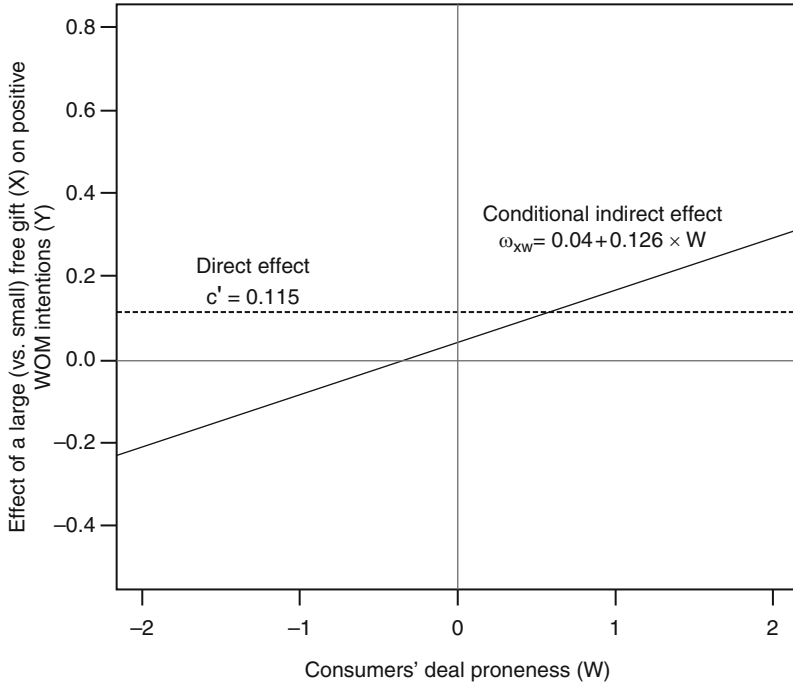


Fig. 12 Plot of the conditional indirect and the direct effect of a large (vs. small) free gift (X) on positive WOM intentions (Y) for different values of consumers' deal proneness (W)

While the y-axis denotes the magnitude of the effect of a large (vs. small) free gift (X) on positive WOM intentions (Y), the x-axis indicates the values of the metric (and mean centered) moderator deal proneness (W). The solid line represents the conditional indirect effect of X on Y through M , which is $\omega_{xw} = a_1b + a_3bW = 0.052 \times 0.767 + 0.165 \times 0.767 \times W = 0.040 + 0.126 \times W$ (see also Eq. 24). As the conditional indirect effect ω_{xw} is dependent on W , the slope of the solid line in Fig. 12 is nonzero: The conditional indirect effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is descriptively positive for highly deal-prone consumers (e.g., for $W = 2$, $\omega_{xw} = 0.04 + 0.126 \times 2 = 0.292$), but negative for low deal-prone consumers (e.g., for $W = -2$, $\omega_{xw} = 0.04 + 0.126 \times (-2) = -0.212$). The slope of the graph depicting the conditional indirect effect corresponds to the index of moderated mediation, which is $a_3b = 0.126$.

The broken line in Fig. 12 denotes the direct effect of X on Y which, in contrast to the conditional indirect effect, is not affected by a moderator and denoted by $c' = 0.115$. That means, the direct effect of a large (vs. small) free gift on positive WOM intentions is $c' = 0.115$ for highly deal-prone consumers, but also for low deal-prone consumers. Hence, the broken line is horizontal.

The equations describing a conditional process model in which path b is moderated by V (see panel B in Fig. 10, excluding the broken lines) follow the same

principles just applied to the case in which a moderator W affects path a . The conditional process model is described by

$$M = i_m + aX + e_m \quad (27)$$

$$Y = i_y + b_1M + b_2V + b_3MV + c'_1X + e_y \quad (28)$$

The i parameters denote the intercepts and a estimates the effect of X on M . Controlling for the remaining variables in the equation, the effect of M on Y is estimated by b_1 , the effect of V on Y by b_2 , the effect of their interaction MV on Y by b_3 , and the effect of X on Y by c'_1 . The error terms are denoted by the respective e parameters. As Eqs. 27 and 28 show, M is a simple function of X . Y , however, is not only affected by X and M but also by V and by the interaction MV . If the direct effect were conditional on V as well (see panel B in Fig. 10, including the broken line), an additional c' parameter would be added to Eq. 28 to denote the partial effect of XV on Y . The resulting equation would be

$$Y = i_y + b_1M + b_2V + b_3MV + c'_1X + c'_2XV + e_y \quad (29)$$

Irrespective of whether or not a moderator V affects path c' , if V moderates path b , the effect of M on Y , the conditional indirect effect is denoted as

$$\omega_{mv} = a(b_1 + b_3V) = ab_1 + ab_3V \quad (30)$$

In this conditional process model, the index of moderated mediation is indicated by ab_3 . The direct effect is again denoted by c'_1 if unconditional, yet, if conditional on the moderator V (see panel C in Fig. 10, including the broken line), it is described by the following equation:

$$\text{Conditional direct effect} : c'_1 + c'_2V \quad (31)$$

As can be seen from Eqs. 25 and 31 (see also Eqs. 22 and 29 for the interpretation of c'_3 and c'_2 , respectively), the equation for the conditional direct effect has the same form irrespective of whether a moderator affects path a and path c' or path b and path c' .

The total effect c is again the sum of the (conditional) direct effect and the conditional indirect effect. If V solely moderates the effect of M on Y , the total effect is denoted by

$$\text{Total effect} : c = \omega_{mv} + c' = a(b_1 + b_3V) + c' \quad (32)$$

The equations underlying the mediated moderation depicted in panel C in Fig. 10 are the following:

$$M = i_m + a_1X + a_2W + a_3XW + e_m \quad (33)$$

$$Y = i_y + bM + c'_1X + c'_2W + c'_3XW + e_y \quad (34)$$

As can be seen, Eqs. 20 and 33 as well as Eqs. 22 and 34 are identical. That is, mediated moderation as shown in Fig. 10 (panel C) is statistically identical to the case of moderated mediation in which a moderator W affects path a and path c' (Fig. 10, panel A) and corresponds to the statistical model shown in Fig. 11 (including the broken lines). Hence, the same relations apply for mediated moderation as for this case of moderated mediation.

The equations describing other conditional process models (e.g., in which a moderator Z affects path a and path b simultaneously) follow the same general principles applied in the examples outlined above. They can be found in publications by, for example, Edwards and Lambert (2007), Preacher et al. (2007), and Hayes (2018).

Statistical Inference for Conditional Process Models: Conditional Process Analysis

Testing for moderated mediation in conditional process models is somewhat different from testing for mediation in single or multiple mediator models, as no single numerical estimate for “the conditional indirect effect” can be tested against zero with the help of a bootstrap confidence interval. Instead, as mentioned above, there are multiple numerical estimates of the conditional indirect effect for different values of the moderator (see Fig. 12), and the question of interest is whether they differ significantly from each other. Different approaches have been introduced to answer this question (e.g., Edwards and Lambert 2007; Fairchild and MacKinnon 2009; Hayes 2015; Muller et al. 2005; Preacher et al. 2007). They can generally be referred to under the term *conditional process analysis* (Hayes 2018).

Earlier approaches (e.g., Muller et al. 2005) implicitly or explicitly rely on the assumption that to establish moderated mediation, one or more paths in the conditional process model need to be significantly moderated. These approaches test moderated mediation, for example, by testing the individual paths in a conditional process model and whether or not they are moderated (Muller et al. 2005) or by extending simple slopes analysis and the Johnson-Neyman technique to determine the significance of the conditional indirect effect at a few (similar to simple slopes analysis) or all (similar to the Johnson-Neyman technique) values of the moderator (Preacher et al. 2007).

It has been pointed out, however, that testing whether a constituent path in the conditional process model is significantly moderated or whether an indirect effect for one or more specific values of the moderator is different from zero is conceptually different from testing whether mediation, that is, the indirect effect, is moderated (Fairchild and MacKinnon 2009; Hayes 2015). In other words, testing moderation of the individual paths in a conditional process model or probing the indirect effect for different values of the moderator is useful for descriptive reasons, but it does not address the central question of interest in conditional process analysis.

Addressing this aspect, a further approach to testing moderated mediation is provided by Hayes (2015). He argues that moderated mediation can be demonstrated by testing the index of moderated mediation which, as noted above, corresponds to

the weight of the moderator in the equation for the conditional indirect effect (see Eqs. 24 and 30). If the index of moderated mediation is different from zero, the conditional indirect effect significantly varies as a linear function of the moderator. For example, if the index of moderated mediation in the free gift example ($a_3b = 0.126$) is different from zero, it could be concluded that the effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is significantly moderated by consumers' deal proneness: The indirect effect would be more positive for highly deal-prone consumers (i.e., consumers that score high on the moderator) as compared to low deal-prone consumers. Specifically, it would be 0.126 units more positive for every unit increase in deal-proneness.

To test the index of moderated mediation, a bootstrap confidence interval can be computed. Hayes (2015) shows that if this confidence interval excludes zero, it can be concluded that any two conditional indirect effects for different values of a moderator (e.g., plus and minus one standard deviation from the mean) significantly differ from each other. In the free gift example, the index of moderated mediation is found to be nonsignificant as the bootstrap confidence interval includes zero ($a_3b = 0.126$, bc.bci_{95%} [-0.014; 0.284]). Hence, it cannot be concluded that the indirect effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is moderated by consumers' deal-proneness.

The shift in focus in conditional process analysis from the moderation of a specific path in earlier approaches (e.g., Muller et al. 2005) to the moderation of the indirect effect in the approach by Hayes (2015) can lead to situations in which significant moderation of the indirect effect is found in the absence of significant moderation of an individual path of the indirect effect, and vice versa. Furthermore, it may happen that the index of moderated mediation is not significant, but there are differences in the significance of conditional indirect effects for different values of the moderator, and vice versa.

Take, for instance, the results from the analysis testing whether the effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is dependent on consumers' deal proneness (see Fig. 11): The regression coefficient denoting the moderating effect of W on path a , a_3 , is not significant ($a_3 = 0.165$, $p = 0.087$) and neither is the index of moderated mediation, as pointed out above ($a_3b = 0.126$, bc.bci_{95%} [-0.014; 0.284]). However, a significant indirect effect is found for highly deal-prone consumers (consumers one standard deviation above the mean, $ab_{\text{dealprone}+} = 0.230$, bc.bci_{95%} [0.056; 0.415]), but not for low deal-prone customers (consumers one standard deviation below the mean, $ab_{\text{dealprone}-} = -0.149$, bc.bci_{95%} [-0.469; 0.143]). Hence, it cannot be concluded that the effect of a large (vs. small) free gift on hedonic benefits (i.e., path a) is moderated by consumers' deal proneness (a_3 is nonsignificant). Moreover, it cannot be said that the conditional indirect effects of a large (vs. small) free gift on positive WOM intentions through hedonic benefits for differently deal-prone consumers differ from each other (a_3b , the index of moderated mediation is not significant). However, a positive conditional indirect effect of a large (vs. small) free gift on positive WOM intentions through hedonic benefits is found for highly deal-prone consumers ($ab_{\text{dealprone}+}$ is

significant), but not for low deal-prone consumers ($ab_{\text{dealprone}}$ is not significant). In situations like these, we argue that the test corresponding most closely to the specific research question investigated should be given the greatest weight in a researcher's judgment. Furthermore, we emphasize again that moderation of the indirect effect cannot be inferred from the mere observation that mediation occurs for some values of the moderator, but not for others (Hayes 2015).

Variable Metrics

A dichotomous or continuous moderator can easily be incorporated into regression-based mediation analysis. Multicategorical moderators can be included as a set of indicator variables (Hayes 2017), again, each representing a comparison of one category (or a set of categories) to another category (or a set of categories, see, e.g., Darlington and Hayes 2017).

Further Mediation Models

In the following section, further, more complex mediation models are described. These are models with multiple mediators and moderators, and with more than one predictor or outcome. Furthermore, we touch upon mediation analysis for longitudinal and multilevel data. Although these models generally go beyond the scope of a mere introduction to mediation analysis in an experimental context, we address them briefly as they, first, illustrate how the principles applied above can be extended to more complex mediation models and, second, account for assumptions in mediation analysis frequently not considered (omitted variables, timing of mediation, or nested data).

Notably, with some exceptions, the following models go beyond the scope of the OLS regression-based approach to mediation analysis. However, they can be analyzed with the help of other methods (e.g., structural equation modeling, see ► “Crafting Survey Research: A Systematic Process for Conducting Survey Research,” this volume). Furthermore, it should be emphasized that mediation models do not have to be complex in order to be of scientific value. Rather, the complexity of the model should be determined weighing the principle of parsimony against the premise to include all important causal variables in the model while keeping in mind that bias due to measurement error is a more serious issue in complex mediation models (Cole and Preacher 2014).

Mediation Models with Multiple Mediators and Moderators

Multiple mediator models can also include moderators (see conceptual diagrams in panel A and panel B in Fig. 13 depicting a moderated parallel multiple mediator model and a moderated serial multiple mediator model, respectively). Furthermore, in multiple mediator models with three or more mediators, serial and parallel mediation can be combined (Hayes 2018).

If a conditional process model includes more than one moderator (see panel C and panel D in Fig. 13 for examples), it can be distinguished between the concepts of

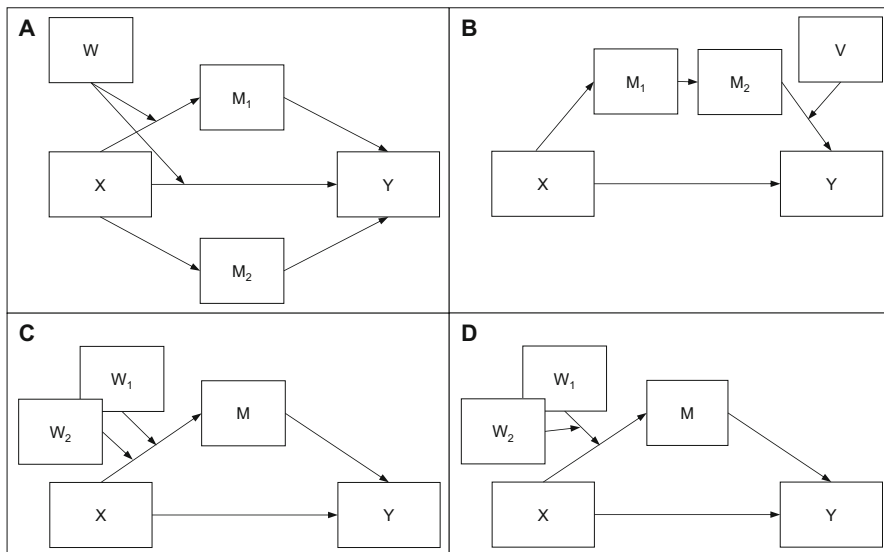


Fig. 13 Conceptual diagrams of mediation models with additional mediators as well as moderators

partial moderated mediation, *conditional moderated mediation*, and *moderated moderated mediation* (Hayes 2017). Partial moderated mediation refers to the question whether a conditional indirect effect of X on Y through M is moderated by a moderator W_1 while a second moderator W_2 affecting the same path is held constant (see panel C in Fig. 13). That is, partial moderated mediation denotes the moderating ability of W_1 independent of W_2 . In conditional moderated mediation and moderated moderated mediation, a second moderator W_2 affects the effect of W_1 (see panel D in Fig. 13 for an example). Hence, the moderation of W_1 is dependent on W_2 . In this scenario, two questions may be of interest: First, is the conditional indirect effect moderated by W_1 when W_2 takes a specific value? Second, does the moderating ability of W_1 change if W_2 changes? Conditional moderated mediation refers to the first question, that is, whether a conditional indirect effect of X on Y through M is moderated by a moderator W_1 at a specific value of W_2 . Moderated moderated mediation addresses the second question, that is, whether the moderation of a conditional indirect effect by a moderator W_1 changes if a second moderator W_2 changes. Details on these concepts, specifically, how to quantify and test them, are discussed by Hayes (2017).

Whether a mediator can be a moderator at the same time is a topic of debate. While some argue that one and the same variable can mediate as well as moderate a relationship between X and Y (e.g., Frazier et al. 2004), others differentiate more (e.g., Kraemer et al. 2002; Kraemer et al. 2008) or less strictly (e.g., Baron and Kenny 1986) between the conceptual definitions of a mediator and a moderator. Without taking a definite stand on the question, we think it is important to be aware of the conceptual and the statistical level of the debate: Differentiating clearly between a moderator and a mediator on a conceptual level does not mean that

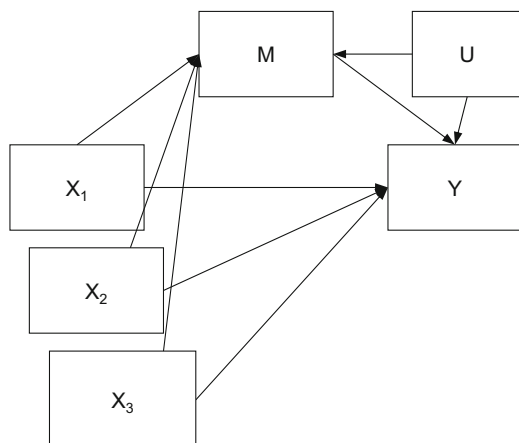
evidence for a mediating process cannot be provided by testing the significance of interaction effects (e.g., Jacoby and Sassenberg 2011; Kraemer et al. 2002, 2008; Pirlott and MacKinnon 2016; Spencer et al. 2005).

Moreover, while one might consider the same variable as a mediator in one model and as a moderator in another, it is another question to consider a variable simultaneously as mediator and moderator in the same model (Hayes 2018). Pieters (2017), for instance, finds no such model being tested in a review of $N = 138$ papers published in the *Journal of Consumer Research* between 2014 and 2016, which included $N = 166$ mediation analyses. We contend that one possible reason for the unusualness of models in which a variable acts as a mediator and moderator simultaneously (see Fig. 5 for such a case) is that it may be rather challenging (albeit likely not impossible) to theoretically deduct hypotheses for this situation. Such a model would, for instance, correspond to a hypothesis stating that the direct effect of a large (vs. small) free gift on positive WOM intentions, namely, the effect of a large (vs. small) free gift on positive WOM intentions controlling for the influence of hedonic benefits, is different for consumers perceiving high hedonic benefits from the sales promotion than for consumers perceiving low hedonic benefits from the sales promotion. However, we also see that it is mathematically possible and that researchers test such models (e.g., Kraemer et al. 2002, 2008).

Mediation Models with Multiple Predictors and Outcomes

Mediation models can be extended to include multiple predictor or outcome variables. These more extensive mediation models can be referred to as path analysis (mediation) models (MacKinnon 2008). Including more than one predictor in a mediation model (e.g., by adding covariates to the model or multiple indicator variables in case an independent variable is multicategorical) does not pose much difficulty to OLS regression-based mediation analysis. Each predictor X_i or covariate U_i is assigned a specific effect on M or Y , which is interpreted as the respective ability of X_i or U_i to predict M or Y , controlling for the effects of all other variables affecting M or Y (see Fig. 14).

Fig. 14 Conceptual diagram of a mediation model with several predictors X_i as well as a covariate U affecting M and Y



Including more than one outcome in a mediation model generally goes beyond what regression analysis can do, and the use of structural equation modeling becomes necessary to estimate the respective path coefficients (MacKinnon 2008). However, Hayes (2018) points out that, if no relationships between the dependent variables Y_i are modeled, regression analysis can be used to analyze mediation models in which X 's effect on multiple Y_i is transmitted through one or more M_i . Then, the estimated indirect effect on Y_i will be the same, regardless of whether the model was fitted simultaneously with structural equation modeling or by a series of separate regressions.

Incorporating Time and Nested Data in Mediation Analysis

Two emerging research fields in mediation analysis are longitudinal mediation analysis and multilevel mediation analysis. Describing them in detail is beyond the scope of this chapter (for overviews, see Preacher 2015 and MacKinnon 2008). However, we briefly address why mediation analysis for longitudinal and multilevel data is relevant.

Development in time is a crucial, albeit often implicit, aspect of mediation analysis: As a cause must precede an effect, time must elapse for X to cause M and for M to cause Y . Cole and Maxwell (2003) argue that mediation analysis based on cross-sectional designs (i.e., designs in which X , M , and Y are measured simultaneously and only once) provide accurate information about a mediation process unfolding over time only under rather limited conditions (see also Maxwell and Cole 2007; Maxwell et al. 2011). Time can be incorporated into mediation analysis through longitudinal designs, that is, repeated measurement of X , M , and Y (MacKinnon 2008; Preacher 2015). An exemplary longitudinal mediation model is depicted in Fig. 15. Longitudinal designs control for error resulting from individual differences and other unobserved variables. However, they are usually affected by common method bias.

Mediation analysis can also accommodate multilevel data, for example, consumers nested in different geographical regions or repeated measurements nested in one participant (Preacher 2015; Tofghi and Thoemmes 2014). An exemplary multilevel mediation model is depicted in Fig. 16. In nested datasets, the assumption

Fig. 15 An exemplary conceptual diagram of a longitudinal mediation model in which X , M , and Y are measured three times (at t_1 , t_2 , and t_3) and the variables are affected by their hypothesized cause as well as by themselves

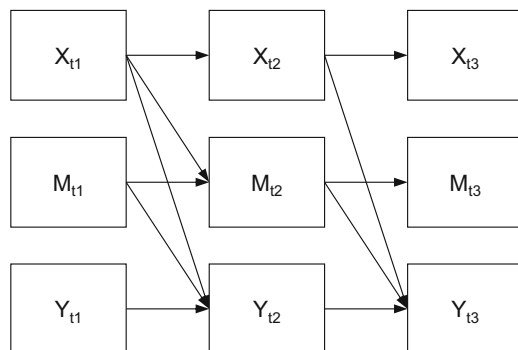
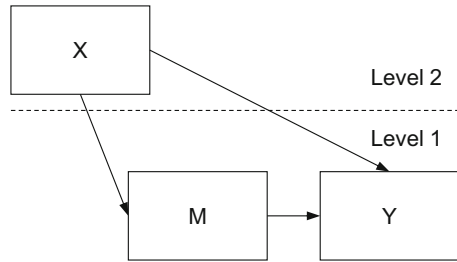


Fig. 16 An exemplary conceptual diagram of a multilevel mediation model in which X is a level-2 variable affecting M and Y which are level-1 variables. This model can be referred to as a 2-1-1 design



of independent error terms is violated (chapter ► “Multilevel Modeling” by Wieseke et al., this volume). If this is unaccounted for, that is, if the nesting is ignored, the risk of reporting a nonexistent indirect effect increases.

Strengthening Causal Inference in Mediation Analysis

Inferring causal relationships is the central idea of mediation analysis. Yet, the mere statistical results obtained from mediation analysis do not allow researchers to draw strong conclusions (if any) about the causal relationships between the variables included in the model (e.g., Baron and Kenny 1986; Bollen 1989). According to Cook and Campbell (1979), causality can be inferred when covariation, temporal precedence, and absence of bias due to confounding variables is given. Hence, if a significant indirect effect (i.e., covariation between X , M , and Y) is found based on simultaneously collected and nonexperimental data, alternative models in which the variables are differently ordered could explain the data as well (e.g., Y could precede M , which, in turn, could precede X , or M could precede X as well as Y). The observed relationships could also be a result of bias due to confounding variables. That is, there might be omitted variables causally influencing the variables in a mediation model. Furthermore, the variables assessed for X , M , and Y may only be correlates of the actual causes, mediators and consequences of an effect. Hayes (2018) refers to this as epiphenomenality.

The following section describes how causal inference in mediation analysis can be strengthened through design, the collection of further evidence, and statistical methods. Taken by themselves, none of these methods is sufficient to establish causality. In combination, however, they may strongly support a causal argument. Beyond that, the most crucial method to strengthen causal inference in mediation analysis is to provide strong theoretical support for one’s hypotheses tested.

Strengthening Causal Inference Through Design

Causal claims in mediation analysis can be strengthened through experimental methods (Spencer et al. 2005; Stone-Romero and Rosopa 2008). Specifically, randomizing X strengthens causal inference for the effect of X on M , as well as for

the total effect of X on Y . This is, first, because X clearly precedes M as well as Y if it is successfully manipulated, and, second, because possible confounding variables are controlled for through the random assignment of participants to different levels of X . Crucially, however, the randomization of X does not allow to causally interpret the effect of M on Y : As participants are not randomly assigned to a “level of M ,” path b is correlational. Furthermore, as the indirect and the direct effect take path b into account (see Eqs. 3 and 4), neither one of them can be interpreted causally.

Randomizing M has been suggested as a way to strengthen causal inference about the relationship between M and Y , for example, by conducting double randomized experiments (Spencer et al. 2005; Stone-Romero and Rosopa 2008; for an overview of different design approaches to manipulate mediators and mediating effects, see Pirlott and MacKinnon 2016). Double randomized experiments consist of a series of experiments. These are composed, first, of a study randomizing X and measuring M (and Y) and, second, of a study randomizing M and measuring Y (further studies investigating the hypothesized causal effect could be added to probe, e.g., whether Y does not cause M , as hypothesized). Evidently, data collected from a chain of experiments is not analyzed using mediation analysis (but, for instance, through analysis of variance, see chapter ▶ “Analysis of Variance” by Landwehr, this volume).

Establishing the causal chain from X to Y via M through multiple experiments provides strong causal evidence for the hypothesized relationships (Stone-Romero and Rosopa 2008). However, there are several difficulties associated with double randomized experiments (e.g., Bullock et al. 2010; Kenny 2008; MacKinnon 2008; Preacher 2015; Spencer et al. 2005; Stone-Romero and Rosopa 2008). Among other concerns, it has been noted that running double randomized experiments (just like experiments in general, see chapter ▶ “Field Experiments” by Valli et al., this volume) may not always be possible (e.g., because an active manipulation of X or M is not feasible, true randomization cannot be achieved, or a suitable control group cannot be identified) or desirable (e.g., because an active manipulation would be unethical). Furthermore, experimental designs may be somewhat artificial with regard to the operationalization employed (e.g., the manipulation of X or measurement of Y) and the setting chosen (e.g., the laboratory). This may cast doubt on the construct validity of the manipulations and measures, and on the external validity of the results obtained. In addition, double randomization assumes that measuring M (in a study in which X is randomized and M is measured) is equal to manipulating M (in a study in which the effect of M on Y is demonstrated), which may not be the case. Spencer et al. (2005) suggest running double randomized experiments only if the proposed mediating process is easy to manipulate and measure. If this does not apply (e.g., if the mediator is hard to measure but easy to manipulate), Spencer et al. (2005) recommend that other designs be employed.

If (double randomized) experiments cannot be conducted, it may seem reasonable to apply a sequential design to strengthen causal inference in mediation analysis, that is, to measure X , M , and Y on three subsequent points in time and hence “allow” X to “precede” M and M to “precede” Y . However, the logic of a sequential design is problematic as the measurement of a construct is independent of the conceptual

timing of a construct: Just because M is measured after X and Y is measured after X and M , one cannot conclude that X precedes M and M precedes Y (Cole and Maywell 2003). As a consequence, a sequential design by itself does not improve causal inference. Besides that, note that the independence of measurement and timing of a construct also implies that, unless the act of measuring Y affects M , Y may just as well be assessed before M without impairing causal inference (Lemmer and Gollwitzer 2017).

Generally, we recommend to apply experimental methods as often as possible while being aware of their limitations in establishing the causal claim that X affects Y through M . Moreover, we encourage readers to run multiple, methodologically diverse studies testing the proposed mediating process as this can compensate for methodological limitations arising when demonstrating mediation only through a single (type of) study. A selective overview of possible strategies to do so is given in the following section.

Strengthening Causal Inference Through the Collection of Further Evidence

Causal inference on mediation can be strengthened through the collection of further evidence after initial support for a proposed mediating process has been found. Specificity designs allow to investigate mediation in greater detail (Preacher 2015), either by isolating the relevant mediator among different possible mediators or by identifying conditions under which mediation is strengthened (enhancement designs) or weakened (blockage design). This can be done through analyzing mediation models with additional mediators or moderators that enhance or block the proposed mediating process (MacKinnon 2008; Pirlott and MacKinnon 2016).

In the free gift example described previously, for instance, testing the parallel multiple mediator model could be interpreted as an attempt to specify the proposed mediator, hedonic benefits. Simultaneously considering two possible mediating variables, hedonic and utilitarian benefits, allows to investigate whether it is specifically hedonic benefits and not consumer benefits more broadly (i.e., including utilitarian benefits) that transmit the effect of the large (vs. small) free gift on positive WOM intentions.

The generalizability of a mediating process can be demonstrated through consistency designs (Preacher 2015), that is, through replications of the initial study in different contexts or employing different conceptually related measures or manipulations (pattern matching, MacKinnon 2008). For instance, a causal claim for the proposed mediation in the free gift example would be strengthened if the effect could be replicated in a study that employs a tangible free gift instead of a gift card.

Further experimental designs to strengthen causal inference in mediation analysis have been suggested by Imai et al. (2013). For instance, strong causal evidence for a proposed mediating process can be provided by employing a parallel design which essentially combines a measurement-of-mediation design (Spencer et al. 2005), that is, a design in which only X is manipulated, but M and Y are measured, with a

concurrent double randomization design (Pirlott and MacKinnon 2016) in which both, X and M are simultaneously manipulated and only Y is measured. Furthermore, Preacher (2015) points out that employing a within-subject design, that is, an experiment in which participants are exposed to all levels of X , may be worthwhile to strengthen causal inference, because participants then serve as their own controls by simultaneously being in all experimental groups. Finally, MacKinnon (2008) notes that the mediating process may also be investigated from a qualitative perspective, as the focus in quantitative research may be too much on issues of statistical significance and less on the originally qualitative nature of a research question.

Strengthening Causal Inference Through Statistical Methods

Beyond experimental methods and collecting more data, statistical methods can also contribute to strengthen causal inference in mediation analysis (see also chapter ► “Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers” by Ebbes et al., this volume). Rooted in the potential outcomes framework, a formal (i.e., theoretical) framework which specifies assumptions for causal inference (see chapter ► “Field Experiments” by Valli et al., this volume), one group of statistical methods specifically addresses the issue of bias due to omitted confounding variables (for an overview, see Imai et al. 2011; MacKinnon and Pirlott 2015; Preacher 2015). These methods either assess the degree to which results of mediation analysis are affected by possible violations of the assumption that no confounding variable is omitted from the model (through sensitivity analysis, e.g., Imai et al. 2010), or aim to account for the influence of confounding variables (for an overview of different methods, see MacKinnon and Pirlott 2015).

Note that one implicit assumption of these methods is that the causal order of the variables in the mediation model is correctly specified which may not be the case. Importantly, as illustrated by Lemmer and Gollwitzer (2017), testing for a different causal order by comparing the size of indirect effects found in different possible mediation models (e.g., a model assuming that $X \rightarrow M \rightarrow Y$ and a model assuming that $X \rightarrow Y \rightarrow M$) while assuming that the mediation model for which the larger indirect effect is found must be the correct one (“reverse mediation testing”), is unlikely to be a helpful strategy to address this problem. Specifically, Lemmer and Gollwitzer (2017) show that the size of the indirect effect is affected by a factor completely unrelated to the true underlying causal order of the variables in a mediation model, namely, measurement error associated with M and Y . As a consequence, reverse mediation testing is rarely effective, likely ineffective, and potentially misleading (Lemmer and Gollwitzer 2017; see also Thoemmes 2015).

Questions Arising When Implementing Mediation Analysis

In the following, selected questions arising when implementing mediation analysis are addressed. Specifically, an overview of studies investigating required sample size and power in mediation analysis is given, and reasons for mean centering variables

in conditional process analysis as well as different coding schemes for categorical independent variables are discussed. Furthermore, limitations of the regression-based approach to mediation analysis are touched on and information is provided as to different software options to perform mediation analysis.

Sample Size and Power in Mediation Analysis

In studies performing mediation analysis, sample size is often too small to achieve sufficient power (Fritz and MacKinnon 2007). Hence, a note on optimal sample size in mediation analysis, that is, the required sample size to detect mediation with a power of $1 - \beta = 0.8$, seems warranted. Unfortunately, there is no easy rule of thumb. Overall, however, it has been shown that the effect with the greatest power in a single mediator model is the indirect effect and the power to detect both, the total and the direct effect, tends to be considerably smaller (Kenny and Judd 2014; Rucker et al. 2011).

Optimal sample size in mediation analysis is (sometimes counterintuitively) affected by the inferential approach, the size of the indirect effect and the individual paths, reliability of the measures, and the complexity of the mediation model. For instance, Fritz and MacKinnon (2007) investigate the required sample size to detect an indirect effect in a single mediator model comparing six different inferential approaches. They report that to detect a small indirect effect using a bias-corrected bootstrap confidence interval, the minimal sample size required to achieve a power of $1 - \beta = 0.8$ may be well above $N = 400$. If the indirect effect is large, the required sample size drops to approximately $N \geq 40$. The percentile bootstrap confidence interval is somewhat more conservative and, hence, requires slightly larger sample sizes to achieve sufficient power (Fritz and MacKinnon 2007; see also MacKinnon et al. 2004).

Taylor et al. (2008) provide evidence on the performance of different inferential approaches to test the three-path specific indirect effect in a serial multiple mediator model. They find that the bias-corrected bootstrap confidence interval performs well in samples larger than $N = 200$. Yet, Pieters (2017) describes an example in which a sample size of $N = 450$ is necessary to detect a three-path specific indirect effect of $a_1 d_{21} b_2 = 0.02$ with a power of $1 - \beta = 0.8$. Thoemmes et al. (2010) point out that in multiple mediator models, the power to detect individual paths and total as well as specific indirect effects can vary considerably. For example, a larger specific indirect effect may be sufficiently powered given a particular sample size, yet a smaller specific indirect effect may be greatly underpowered. Williams and MacKinnon (2008) report that, given a specific sample size, specific indirect effects consisting of three paths are harder to detect than are two-path specific indirect effects.

Fairchild and MacKinnon (2009) report that when performing conditional process analysis, the required sample size to reach a power of $1 - \beta = 0.8$ might be as large as $N = 500$ or even $N = 1,000$, depending on how much variance is explained by the model (the more, the better). These results are in accordance with results reported by Pieters (2017) who found a sample size of $N = 500$ to be needed to demonstrate moderated mediation. Morgan-Lopez and MacKinnon (2006) further

show that power to detect mediated moderation is impaired when the direct effect in the model is nonzero and the independent variable and the moderator are correlated.

Hoyle and Kenny (1999) point out that measurement error affecting the mediator can decrease power to detect mediation. Furthermore, they argue that large collinearity between X and M (corresponding to a large effect of X on M in an experimental context) negatively affects power in mediation analysis as it reduces the amount of variance in M that can contribute to predict Y (i.e., path b) and, consequently, the size of the indirect effect.

When data collection is complicated or expensive, sample size will likely be small. Increasing power in mediation analysis without increasing sample size can be achieved by maximizing variance in the independent variable, the mediator, or the dependent variable and by minimizing error variance (Fairchild and MacKinnon 2009; MacKinnon et al. 2013; Fritz et al. 2015). This can be achieved by design (e.g., by ensuring that the experimental manipulation is sufficiently strong), improved measurement (i.e., minimizing measurement error), or statistical methods (e.g., decreasing error through the inclusion of covariates or the use of structural equation modeling). Power could also be increased by using modern missing data techniques instead of listwise deletion of cases (Fritz et al. 2015), or simply by increasing alpha (Fairchild and MacKinnon 2009). However, Fritz et al. (2015) demonstrate that some measures intended to increase power may also have the opposite effect under specific circumstances (e.g., when there is not enough power to detect the effect of a covariate included to reduce error).

Mean Centering in Conditional Process Analysis

In conditional process analysis, it is sometimes recommended to mean center variables composing an interaction term (e.g., Muller et al. 2005). It has been argued that this reduces multicollinearity between the predictors (e.g., X and W) and interaction terms (e.g., XW) in the model, but this argument has been rebutted several times (Dalal and Zickar 2012; Echambadi and Hess 2007; Hayes 2018). However, mean centering variables without a meaningful zero point improves the interpretability of regression coefficients in conditional process analysis (see also chapter ► “Regression Analysis” by Skiera et al., this volume).

For instance, earlier, we discussed a conditional process model in which the effect of a large (vs. small) free gift (X) on positive WOM intentions (Y) through hedonic benefits (M) is dependent on consumers’ deal proneness (W). Assuming that W only affects path a , the equations describing the model look as follows (see Eqs. 20 and 21): $M = i_m + a_1X + a_2W + a_3XW + e_m$ and $Y = i_y + c'_1X + bM + e_y$. As the interaction between X and W is included in the model predicting M , the effect of X on M is modeled to depend on W . That is, the effect of X on M changes in value depending on the value of W . The regression coefficient a_1 denoting the effect of X on M must, hence, be tied to a specific value of W and this value is zero. If W is measured on a scale ranging from 1 (not at all deal-prone) to 7 (very deal-prone), zero is not meaningful and, as a consequence, neither is a_1 . However, if the

moderator is mean centered, zero indicates the mean deal proneness in the sample and a_1 denotes the effect of the large (vs. small) free gift on hedonic benefits for consumers with average deal proneness. Importantly, as interactions are symmetric, the same reasoning applies to the effect of W on M , a_2 , which is only meaningful if X has a meaningful zero point.

That is, centering affects the regression coefficients of the variables composing the interaction term. However, it does not affect the regression coefficient associated with the interaction term itself. Hence, to obtain meaningful regression coefficients for variables composing an interaction term, it is important to ensure that these variables have a meaningful zero point. Mean centering is one way of achieving this.

Coding of Categorical Independent Variables

When performing mediation analysis on data gathered in experiments, a coding scheme has to be selected to translate qualitative information about experimental groups into a set of quantitative indicator variables (Hayes and Preacher 2014). The coding scheme determines how the regression coefficients of the indicator variables are interpreted; hence, it should be selected carefully. There is a variety of different coding schemes of which an overview is presented, for example, by Darlington and Hayes (2017). To illustrate how coding affects the interpretation of regression coefficients, and thus, indirect effects, we briefly discuss two common coding schemes, dummy coding and effect coding.

When employing dummy coding (also called treatment coding or indicator coding), each experimental group G_i is compared to a reference group G_R . Accordingly, the regression coefficients associated with dummy coded indicator variables denote the mean difference between one group G_i and G_R , respectively. Dummy coded indicator variables consist of a series of zeros and ones representing these contrasts (see, e.g., Darlington and Hayes 2017). In fact, in all analyses conducted throughout this chapter, X is dummy coded and denoted by a variable taking the value of zero for participants in the small free gift condition and the value of one for participants in the large free gift condition. Consequently, the indirect effect is interpreted as the effect of the large (vs. small) free gift on positive WOM intentions through the respective mediator(s).

Effect coding (also called effects coding or sum coding) differs from dummy coding in that the individual groups G_i are compared to the grand mean, that is, the overall mean of the to-be-predicted variable across all groups. Accordingly, the regression coefficients associated with the indicator variables represent the mean deviation of a group G_i from the grand mean. In the simplest case, an effect coded indicator variable representing two experimental groups takes the values 1 and -1 (for more examples, see, e.g., Darlington and Hayes 2017). For instance, the independent variable from the free gift example could also be denoted by an indicator variable taking the value of 1 for participants in the large free gift condition and the value of -1 for participants in the small free gift condition. In this case, an indirect effect of $ab = 0.051$ indicates that, as compared to the mean positive WOM

intentions in the overall sample, a large free gift increases participants positive WOM intentions through hedonic benefits by 0.051 units.

Hair et al. (2006) argue that dummy coding is the most appropriate coding for an experiment in which there is a control group (G_R) that one or more experimental groups (G_i) are to be compared to. More generally, we would argue that whichever coding scheme should be employed depends on the research question investigated and the specific hypotheses to be tested. Notably, results obtained from analysis of variance can be replicated using regression analysis, namely, if the independent variable(s) are effect coded (chapter ► “[Analysis of Variance](#)” by Landwehr, this volume). This emphasizes the usefulness of regression analysis for analyzing data gathered in experiments.

Regression Analysis Versus Structural Equation Modeling

It was pointed out earlier that the linear equations describing a mediation model can be fitted sequentially with the help of regression analysis or simultaneously using structural equation modeling (SEM, ► “[Crafting Survey Research: A Systematic Process for Conducting Survey Research,](#)” this volume). There is a lively debate about which approach, regression analysis or SEM, is better when it comes to mediation analysis (e.g., James et al. 2006; Iacobucci et al. 2007; Hayes et al. 2017; Pek and Hoyle 2016). Reasonable arguments have been presented for either side referring to conceptual and statistical differences between the two approaches. Ultimately, we leave it up to the reader which approach to take. However, to enable an informed decision, we summarize important differences, shortcomings, and advantages of either approach in the following section.

On a conceptual level, the two approaches differ with regard to their focal mediation paradigm (James et al. 2006). Specifically, as a consequence of traditions specific to either approach, the default mediation model in regression-based mediation analysis assumes that X may directly affect Y even after controlling for M , as this is presumably likely the case in psychological research where the regression-based approach to mediation analysis originated (Baron and Kenny 1986). Within the SEM approach, however, whether or not the direct effect is included in the model depends on a priori considerations specific to the research question. As a consequence, James et al. (2006) argue that the SEM approach is more parsimonious, and hence, more in accordance with scientific principles.

With regards to statistical differences, it is important to consider that SEM encompasses regression analysis, meaning that any model that can be estimated with regression analysis can also be estimated with SEM. However, as mentioned above, the two approaches differ when it comes to fitting a mediation model. Whereas, in the regression-based approach, the equations describing the mediation model are sequentially fitted for each criterion M_i and Y , they are simultaneously fitted in the SEM approach. This has several consequences. First, it has been argued that simultaneously fitting the whole mediation model is closer to the conceptualization of mediation as one process as compared to a

causal chain of separate effects (Pek and Hoyle 2016). Second, the SEM approach is more flexible with regards to the complexity of the model fitted. For instance, while it is not possible to fit mediation models including multiple correlated mediators and outcomes with regression analysis (e.g., longitudinal mediation), such models can be analyzed within a SEM framework. At the same time, however, computational tools such as PROCESS, a macro for SPSS and SAS, relying on regression analysis (Hayes 2018) accommodate a variety of mediation models common to experimental research and will, hence, be sufficient in many cases. Third, it is possible in SEM, but not in regression analysis, to assess how well the mediation model fits the data. This allows to evaluate a specific mediation model as well as to compare multiple mediation models to each other. Hayes et al. (2017) argue, however, that information about a mediation model's fit carries little additional insight: First, fit for saturated models, that is, mediation models that include all possible paths, is likely perfect. Moreover, slightly different mediation models may fit equally well. Finally, testing the significance of specific coefficients is likely to carry more weight in a researcher's judgment than information about fit, as hypotheses generally refer to such specific coefficients (e.g., the index of moderated mediation or a specific indirect effect in a multiple mediator model).

Furthermore, while regression analysis is based on the assumption that latent constructs can be inferred from measured variables without measurement error, SEM estimates measurement error by statistically differentiating between manifest, that is, measured variables and latent variables. Provided that all assumptions underlying the estimation of latent variables are met, this accounts for the unreliability of measured variables. However, Pek and Hoyle (2016) note that there is a bias-efficiency trade-off to adding latent variables to a mediation model, as their inclusion to the model (just as the inclusion of manifest variables) makes it necessary to estimate more parameters which, all else being equal, may reduce the power of the analysis. Another issue with latent variables arises in conditional process analysis as the benefits of accounting for measurement error have to be weighed against considerable methodological uncertainty associated with estimating interactions between latent variables (Hayes et al. 2017).

Finally, Hayes et al. (2017) point out that SEM software may have more sophisticated options to deal with missing data than more basic statistical software. Furthermore, Hayes (2018) notes that with small samples, which are common in experimental research, SEM programs may be slightly biased as their standard errors may be underestimated in such conditions. However, Iacobucci et al. (2007) show that SEM performs well with samples as small as $N = 30$.

Overall, though, Hayes (2018) sees little justification in the general claim that SEM is the better approach to mediation analysis than regression analysis. Assuming that both approaches are suitable to analyze a mediation model, he argues that differences observed in the results from regressions and SEM (e.g., estimates of coefficients or boundaries of bootstrap confidence intervals) are indicative of computational characteristics of a specific SEM software rather than an actual difference between both methods in their ability to reveal mediation.

Software Tools for Mediation Analysis

An increasing variety of software tools enable commonly used statistic programs such as SPSS, SAS, or R to perform mediation analysis. For instance, *PROCESS*, the previously mentioned macro for SPSS and SAS (www.processmacro.org, Hayes 2018), allows researchers to analyze a considerable range of mediation models combining several inferential methods within the regression-based approach: *PROCESS* performs the causal steps procedure, runs the Sobel test, and computes different bootstrap confidence intervals. Other macros for SPSS and SAS allow to incorporate nonlinear effects in the mediation model (*MEDCURVE*, Hayes and Preacher 2010), perform mediation analysis in studies employing a two-condition within-subject design (*MEMORE*, Montoya and Hayes 2017), and use the distribution of the product approach to test the indirect effect (*PRODCLIN*, MacKinnon et al. 2007b). Many common SEM software options compute bootstrap confidence intervals for the indirect effect as well (e.g., Mplus, Muthén and Muthén 1998). With the help of so-called packages (i.e., shared code), general statistical software such as R can also be used for rather basic as well as advanced mediation analysis (e.g., *lavaan*, Rosseel 2012; *MBESS*, Kelley 2007; *mediation*, Tingley et al. 2014; *RMediation*, Tofghi and MacKinnon 2011; *psych*, Revelle 2016).

After having collected and analyzed the data, the next step is to report the results of one's mediation analysis. Excellent recommendations on how to do so comprehensively, comprehensibly, and convincingly are given, for example, by Hayes (2018) and Pieters (2017).

Summary

This chapter provides a regression-based introduction to mediation analysis with an emphasis on mediation analysis in an experimental context. Hence, the focus lies on the description and analysis of selected mediation models common to experimental research (the single mediator model, parallel and serial multiple mediator models, and conditional process models), while more complex mediation models are just briefly discussed. The chapter further addresses the question of how to strengthen causal inference in mediation analysis through design, the collection of additional data, and statistical methods, and closes with a discussion of topics frequently arising when implementing mediation analysis.

However, this chapter only represents a partial survey of the impressive progress made in mediation analysis over the last decade. Furthermore, many research questions in mediation analysis remain unsatisfactorily answered (for a recent summary, see Preacher 2015). Hence, we highly encourage readers to use the literature cited here as a starting point for further literature search. For instance, readers interested in a more detailed illustration of regression-based mediation analysis may refer to Hayes (2018), while readers coming from a structural equation background may find more information from MacKinnon (2008). Finally, an

illustration of recent developments in mediation analysis with a special emphasis on how to strengthen causal inference in mediation analysis is given by VanderWeele (2015).

Cross-References

- ▶ [Analysis of Variance](#)
- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Field Experiments](#)
- ▶ [Multilevel Modeling](#)
- ▶ [Regression Analysis](#)
- ▶ [Structural Equation Modeling](#)

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.
- Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. *Journal of Consumer Psychology*, *24*(4), 586–607.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*(1), 115–140.
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*(4), 550–558.
- Cavanaugh, L. A. (2014). Because I (don't) deserve it: How relationship reminders and deservingness influence consumer indulgence. *Journal of Marketing Research*, *51*(2), 218–232.
- Chandon, P., Wansink, B., & Laurent, G. (2000). A benefit congruency framework of sales promotion effectiveness. *Journal of Marketing*, *64*(4), 65–81.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 558–577.
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Boston: Houghton Mifflin.
- Dalal, D. K., & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods*, *15*(3), 339–362.
- Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. New York: Guilford Press.
- Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, *26*(3), 438–445.
- Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, *12*(1), 1–22.

- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185.
- Fairchild, A. J., & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science*, 10(2), 87–99.
- Frazier, P., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51(1), 115–134.
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233–239.
- Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, 47(1), 61–87.
- Fritz, M. S., Cox, M. G., & MacKinnon, D. P. (2015). Increasing statistical power in mediation models without increasing sample size. *Evaluation & the Health Professions*, 38(3), 343–366.
- Fritz, M. S., Kenny, D. A., & MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivariate Behavioral Research*, 51(5), 681–697.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle River: Pearson Prentice Hall.
- Hansen, W. B., & McNeal, R. B. (1996). The law of maximum expected potential effect: Constraints placed on program effectiveness by mediator relationships. *Health Education Research*, 11(4), 501–507.
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76(4), 408–420.
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.
- Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, 50(1), 1–22.
- Hayes, A. F. (2017). Partial, conditional, and moderated moderated mediation: Quantification, inference, and interpretation. *Communication Monographs*. <https://doi.org/10.1080/03637751-2017-1352100>.
- Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, 39(4), 709–722.
- Hayes, A. F., & Preacher, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research*, 45(4), 627–660.
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67(3), 451–470.
- Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis does method really matter? *Psychological Science*, 24(10), 1918–1927.
- Hayes, A. F., Montoya, A. K., & Rockwood, N. J. (2017). The analysis of mechanisms and their contingencies: PROCESS versus structural equation modeling. *Australasian Marketing Journal*, 25(1), 76–81.
- Hoyle, R. H., & Kenny, D. A. (1999). Sample size, reliability, and tests of statistical mediation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 195–222). Thousand Oaks: Sage.
- Iacobucci, D. (2012). Mediation analysis and categorical variables: The final frontier. *Journal of Consumer Psychology*, 22(4), 582–594.
- Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology*, 17(2), 139–153.
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309–334.

- Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765–789.
- Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 5–51.
- Jacoby, J., & Sassenberg, K. (2011). Interactions do not only tell us when, but can also tell us how: Testing process hypotheses by interaction. *European Journal of Social Psychology*, *41*(2), 180–190.
- James, L. R., Mulaik, S. A., & Brett, J. M. (2006). A tale of two methods. *Organizational Research Methods*, *9*(2), 233–244.
- Jose, P. E. (2013). *Doing statistical mediation and moderation*. New York: Guilford Press.
- Judd, C. M., & Kenny, D. A. (1981). Process analysis estimating mediation in treatment evaluations. *Evaluation Review*, *5*(5), 602–619.
- Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*(4), 979–984.
- Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods*, *11*(2), 353–358.
- Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science*, *25*(2), 334–339.
- Kisbu-Sakarya, Y., MacKinnon, D. P., & Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate Behavioral Research*, *49*(3), 261–268.
- Koschate-Fischer, N., & Schandlmeier, S. (2014). A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. *Journal of Business Economics*, *84*(6), 793–826.
- Koschate-Fischer, N., Stefan, I. V., & Hoyer, W. D. (2012). Willingness to pay for cause-related marketing: The impact of donation amount and moderating effects. *Journal of Marketing Research*, *49*(6), 910–927.
- Koschate-Fischer, N., Huber, I. V., & Hoyer, W. D. (2016). When will price increases associated with company donations to charity be perceived as fair? *Journal of the Academy of Marketing Science*, *44*(5), 608–626.
- Koschate-Fischer, N., Hoyer, W. D., Stokburger-Sauer, N. E., & Engling, J. (2017). Do life events always lead to change in purchase? The mediating role of change in consumer innovativeness, the variety seeking tendency, and price consciousness. *Journal of the Academy of Marketing Science*. <https://doi.org/10.1007/s11747-017-0548-3>.
- Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*(10), 877–883.
- Kraemer, H. C., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*(2S), 101–108.
- Lemmer, G., & Gollwitzer, M. (2017). The “true” indirect effect won’t (always) stand up: When and why reverse mediation testing fails. *Journal of Experimental Social Psychology*, *69*, 144–149.
- Lichtenstein, D. R., Netemeyer, R. G., & Burton, S. (1995). Assessing the domain specificity of deal proneness: A field study. *Journal of Consumer Research*, *22*(3), 314–326.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Routledge.
- MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, *17*(2), 144–158.
- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, *19*(1), 30–43.
- MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41–62.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*(4), 173–181.

- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83–104.
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39(1), 99–128.
- MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007a). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, 39(3), 384–389.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007b). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.
- MacKinnon, D. P., Kisbu-Sakarya, Y., & Gottschall, A. C. (2013). Developments in mediation analysis. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology: Volume 2: Statistical analysis* (pp. 338–360). New York: Oxford University Press.
- Mathieu, J. E., & Taylor, S. R. (2006). Clarifying conditions and decision points for mediational type inferences in organizational behavior. *Journal of Organizational Behavior*, 27(8), 1031–1056.
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12(1), 23–44.
- Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, 46(5), 816–841.
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110(1), 40–48.
- Montoya, A. K., & Hayes, A. F. (2017). Two condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, 22(1), 6–27.
- Morgan-Lopez, A. A., & MacKinnon, D. P. (2006). Demonstration and evaluation of a method for assessing mediated moderation. *Behavior Research Methods*, 38(1), 77–87.
- Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89(6), 852–863.
- Muthén, L. K., & Muthén, L. (1998). *Mplus [computer software]*. Los Angeles: Muthén & Muthén.
- Pek, J., & Hoyle, R. H. (2016). On the (in) validity of tests of simple mediation: Threats and solutions. *Social and Personality Psychology Compass*, 10(3), 150–163.
- Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, 44(3), 692–716.
- Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, 66, 29–38.
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(1), 825–852.
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3), 879–891.
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16(2), 93–115.
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, 6(2), 77–98.
- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1), 185–227.
- Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research (Version 1.6.12). <http://personality-project.org/r>, <http://personality-project.org/r/psych-manual.pdf>. Accessed 24 July 2017.
- Rossee, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

- Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, 5(6), 359–371.
- Savary, J., Goldsmith, K., & Dhar, R. (2014). Giving against the odds: When tempting alternatives increase willingness to donate. *Journal of Marketing Research*, 52(1), 27–38.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7(4), 422–445.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13, 290–312.
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89(6), 845–851.
- Stone-Romero, E. F., & Rosopa, P. J. (2008). The relative validity of inferences about mediation as a function of research design characteristics. *Organizational Research Methods*, 11(2), 326–352.
- Taylor, A. B., MacKinnon, D. P., & Tein, J.-Y. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, 11(2), 241–269.
- Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology*, 37(4), 226–234.
- Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling*, 17(3), 510–534.
- Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5), 1–38.
- Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods*, 43(3), 692–700.
- Tofighi, D., & Thoemmes, F. (2014). Single-level and multilevel mediation analysis. *The Journal of Early Adolescence*, 34(1), 93–119.
- Touré-Tillery, M., & McGill, A. L. (2015). Who or what to believe: Trust and the differential persuasiveness of human and anthropomorphized messengers. *Journal of Marketing*, 79(4), 94–110.
- Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18(2), 137–150.
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press.
- VanderWeele, T. J., & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1), 95–115.
- Wen, Z., & Fan, X. (2015). Monotonicity of effect sizes: Questioning kappa-squared as mediation effect size measure. *Psychological Methods*, 20(2), 193–203.
- Williams, J., & MacKinnon, D. P. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15(1), 23–51.
- Yuan, Y., & MacKinnon, D. P. (2014). Robust mediation analysis based on median regression. *Psychological Methods*, 19(1), 1–20.
- Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2), 197–206.

Part III
Applications



Measuring Customer Satisfaction and Customer Loyalty

Sebastian Hohenberg and Wayne Taylor

Contents

Introduction	910
Conceptual Background	911
The Relationship of Customer Satisfaction and Loyalty	911
Conceptualizing Customer Satisfaction and Loyalty	913
Measuring Customer Satisfaction	915
Survey Scales	915
Other Measurement Approaches	919
Measuring Customer Loyalty	921
Overview	921
Loyalty Intentions	922
Loyalty Behavior	925
The Future of Managing Customer Satisfaction and Loyalty	932
Concluding Remarks	933
References	934

Abstract

Measuring customer satisfaction and customer loyalty represents a key challenge for firms. In response, researchers and practitioners have developed a plethora of options on how to assess these phenomena. However, existing measurement approaches differ substantially with regard to their complexity, sophistication, and information quality. Furthermore, guidance is scarce on how firms can leverage and combine these approaches to implement a state-of-the-art satisfaction and loyalty measurement system. This chapter attempts to address this vacancy. The authors first define and conceptualize customer satisfaction

S. Hohenberg (✉)

McCombs School of Business, The University of Texas, Austin, TX, USA
e-mail: Sebastian.Hohenberg@mcombs.utexas.edu

W. Taylor

Cox School of Business, Southern Methodist University, Dallas, TX, USA
e-mail: wjtaylor@smu.edu

and customer loyalty. Next, the authors provide an overview of the different operationalization and measurement approaches that companies face when designing a customer satisfaction and loyalty measurement system. The authors also discuss some of the common modeling challenges associated with measuring loyalty, namely, dealing with self-selection bias. Finally, the authors project what the future holds in this area.

Keywords

Customer satisfaction · Customer loyalty · Measurement · Conceptualization · Operationalization · Scales · Loyalty programs

Introduction

Customer satisfaction and customer loyalty are key constructs in marketing management (Anderson et al. 1994; Howard and Sheth 1969). Due to their importance, research provides rich insights regarding their nature as well as regarding the determinants and consequences of both phenomena (Palmatier et al. 2006). Moreover, empirical evidence indicates that marketing managers conceive customer satisfaction and loyalty as important success factors (Aksoy 2013). Studies by Bain and Company (2013) and Anderson (2010) identify customer satisfaction and loyalty as top strategic priorities for firms. Furthermore, empirical results show that increasingly volatile customer and competitor behaviors in a digitalized economy will further increase the relevance of systematically managing customer satisfaction and loyalty in the upcoming years (Ernst and Young 2011; Reeves and Deimler 2011). Brooke (2016) recently summarized these issues: “the initial transaction between buyer and seller is but a prologue to the overall concern of marketing. Few businesses can be sure their customers will continue to engage with them (. . .). We live in an age of disruption” (p. 30).

For successfully managing customer satisfaction and loyalty, the basic requirement is the effective assessment of these constructs (Peterson and Wilson 1992; Watson et al. 2015). However, as Hayes (2008) points out, conceptualizing and measuring customer satisfaction and loyalty represent strong managerial challenges, especially for three main reasons. First, there is a wide range of different assessment approaches. Yet, evidence indicates that the existing conceptualizations and measurement approaches substantially differ with regard to their complexity, sophistication, and information quality (Fornell et al. 1996; Morgeson et al. 2011; Sheth 1970). Second, new marketing trends and technologies, such as Big Data or social media, provide various novel opportunities for marketers to gain insights on customer attitudes and behaviors (Homburg et al. 2015; Kozinets et al. 2010; Weinberg et al. 2015). These novel opportunities may have relevance for assessing customer satisfaction and loyalty in certain contexts. Third, empirical proof shows that the suitability of novel and existing customer satisfaction and loyalty approaches may substantially vary according to the specific application field and the consequences of mis-measurement may be severe (Aksoy 2013; Hayes 2008). For

instance, companies may mis-target customers based on an incorrect measurement, that is, dedicating resources to customers who don't need attention (e.g., those that are already satisfied and loyal or those that are a "lost case" for the firm) or failing to dedicate resources to key customers (e.g., customers who the firm is at risk of losing due to declining satisfaction).

Thus, given the lack of overviews, evaluations, and application guidelines of the different traditional and novel tools and measurement approaches, many firms struggle to design appropriate measurement systems for their particular needs and contexts (Hayes 2008). Against this background, this chapter introduces the reader to the constructs of customer satisfaction and loyalty. Moreover, actionable approaches and tools to measure customer satisfaction and loyalty information are described. Specifically, this chapter addresses the following questions:

- How can firms conceptualize customer satisfaction and customer loyalty?
- How can firms measure customer satisfaction?
- How can firms measure customer loyalty?

This chapter answers these questions across six sections. In this first section we have introduced the relationship between satisfaction and customer loyalty. The second section continues this discussion and explores the conceptual background of each construct. The third and fourth sections outline the common methods in which firms measure both satisfaction and customer loyalty. For satisfaction we discuss surveys, focus groups, and complaint analyses, and for customer loyalty we highlight surveys and databases, with additional attention to loyalty programs. We recognize that there are additional methods for measuring customer satisfaction and loyalty (e.g., social media) and that methods can apply to both topics (e.g., some databases can be used to gain insights on customer satisfaction). However, in the interest of brevity, we focus on the primary methods used within each subject in addition to the nuances of using each method for a given subject. The fifth section emphasizes the recent trends and future directions in measuring satisfaction and customer loyalty. Finally, the sixth section concludes.

Conceptual Background

The Relationship of Customer Satisfaction and Loyalty

Research has extensively analyzed customer satisfaction, customer loyalty, their relationship, as well as potential antecedents and consequences (see Palmatier et al. (2006) for an overview). Customer *satisfaction* is generally referred to as a postconsumption evaluation of perceived quality relative to prepurchase expectations about quality (Homburg et al. 2005, p. 85). In contrast, customer *loyalty* is defined as "a collection of attitudes aligned with a series of purchase behaviors that systematically favor one entity over competing entities" (Watson et al. 2015, p. 804). See section "[Conceptualizing Customer Satisfaction and Loyalty](#)" for a more detailed conceptualization of both constructs.

As prior investigators have pointed out, the existing knowledge on customer satisfaction and loyalty can be summarized along the “Customer Relationship Management (CRM)-Outcome Chain” (Anderson and Mittal 2000; Kumar and Reinartz 2012). More precisely, according to the CRM-Outcome Chain, firms’ marketing activities provoke customers’ psychological states (i.e., attitudes) as well as other loyalty reasons, which, in turn, result in diverse loyalty intentions and actual behaviors, eventually manifesting in economic outcomes (cf. Fig. 1). For instance, a firm’s investment in a customer loyalty program (i.e., a marketing activity) may enhance customers’ satisfaction (i.e., an attitude), which drives their loyalty intentions and is likely to result in additional future sales (e.g., actual loyalty behavior in terms of repurchases or cross-buying) and economic company success.

Thus, as shown by the CRM-Outcome Chain, customer satisfaction represents an important antecedent of customer loyalty. However, as the CRM-Outcome Chain also indicates, an increase in customer satisfaction does not necessarily result in a (equal) gain of customer loyalty (Anderson 1996; Woodruff et al. 1983). This is due to two main reasons. First, there are other factors besides customer satisfaction that can influence customer loyalty, for instance, other psychological states (e.g., trust, commitment), loyalty incentives (e.g., rewards for repurchases or cross-buying), contractual obligations (e.g., due to a legal contract, the customer must stay within a given relationship), technical causes (e.g., the customer depends on a system of a given provider), and economical causes (e.g., changing the supplier is relatively costly due to existing rebates or bonuses) (Hayes 2008; Kumar and Reinartz 2012; Watson et al. 2015). A meta-analysis has therefore found that customer satisfaction explains less than 25% of the variance of components of customer loyalty (Szymanski and Henard 2001). Second, the magnitude of the customer satisfaction-customer loyalty relationship is likely to depend on various situational and

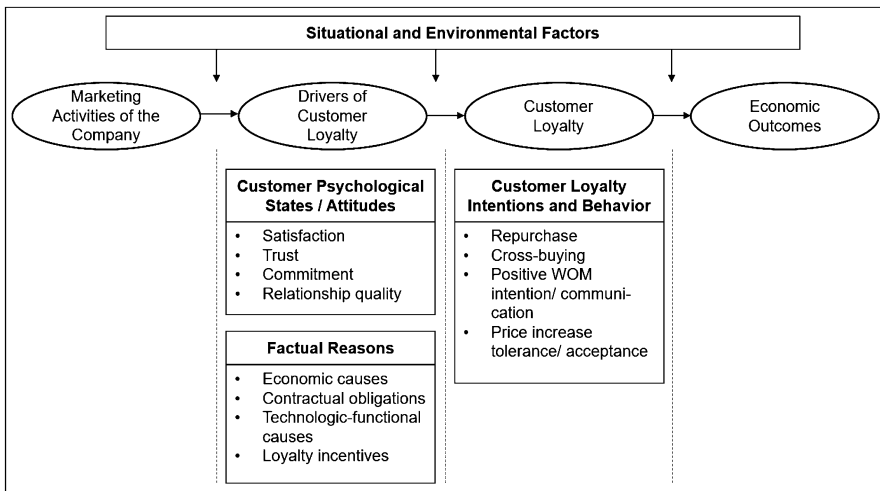


Fig. 1 CRM-Outcome Chain. (Adapted from Kumar and Reinartz 2012; Watson et al. 2015)

environmental factors (Aksoy 2013; Larivière et al. 2016; Shankar et al. 2003). For example, Bolton (1998) found that customer satisfaction is positively related to customer loyalty, yet the strength of this relationship depends on other factors, such as the level of satisfaction or the length of the prior customer-firm relationship. Similarly, Anderson et al. (2004) find the customer satisfaction-loyalty relationship is significantly weaker under high competition.

As a result of potential competing antecedents and various contingent conditions, Kumar et al. (2013) conclude in their review of the customer satisfaction-loyalty relationship: “the customer satisfaction-loyalty main effect is indeed weak and that customer satisfaction, by itself, can hardly change customer loyalty in a significant way. In fact, the systematic presence of moderators, mediators, and other predictors of loyalty introduce a high variability in the findings, thus reducing the role of satisfaction” (p. 247). In line with this point of view, rich evidence today indicates that even very satisfied customers can defect. For instance, Reichheld (1996) finds that more than 60% of satisfied customers may actually switch their providers. Likewise, evidence indicates that only about half of the households with service problems would remain loyal, even if their problems were satisfactorily resolved (Chandrashekar et al. 2007). These results emphasize that, in order to achieve sustainably high sales success with the existing customer base, firms need to systematically assess and manage customers’ loyalty as well as their satisfaction levels (Luo and Homburg 2007; Rust and Zahorik 1993). In embracing this comprehensive view, this chapter focuses in the following on the conceptualization and measurement of both, customer satisfaction and customer loyalty.

Conceptualizing Customer Satisfaction and Loyalty

To conceptualize customer satisfaction, prior research has distinguished (1) a transaction-specific perspective and (2) a cumulative perspective (Anderson et al. 1994). The former perspective conceives customer satisfaction as the buyer’s cognitive state, resulting from the evaluation of being adequately rewarded for a particular sacrifice she has undergone (Churchill Jr. and Surprenant, 1982; Howard and Sheth 1969; Oliver 1981). In contrast, the latter perspective comprehends customer satisfaction as the cognitive state resulting from the evaluations of the entire interactions with a firm over time (Hunt 1977; Verhoef 2003). Hence, the transaction-specific and the cumulative perspective provide a highly similar understanding of customer satisfaction that essentially differs with regard to the reference object (i.e., a single transaction versus the entire relationship).

Thus, by drawing on both of these perspectives, this study conceptualizes customer satisfaction as the result of a cognitive process during which the customer compares her prior expectations regarding the product’s performance with the actually perceived performance (Gupta and Zeithaml 2006). This conceptualization builds on the “Confirmation-Disconfirmation Paradigm” (Oliver 1980). According to this paradigm (cf. Fig. 2), customers compare the perceived performance of the product or service to an expected performance standard (e.g., based on prior

experiences or desires) (Halstead 1999). If the customer perceives the actual performance as higher (equal) relative to her expectations, the expectations will be positively discontinued (confirmed), thus resulting in customer satisfaction. In contrast, if the expected performance is greater than the actually perceived performance, customers will experience a negative discontinuation of their expectations, which results in dissatisfaction (McCollough et al. 2000).

The conceptualization of customer loyalty is more complicated. Although customer loyalty has been in the focus of marketing research and practice for a long time (Oliver 1999), there is no consensus among researchers on how to define customer loyalty (Aksoy 2013; Kumar and Reinartz 2012; McAlexander et al. 2003). Yet, most prior studies agree that customer loyalty is a complex, multidimensional construct.

More precisely, prior research has often conceptualized customer loyalty by differentiating two theoretical elements, i.e., loyalty (future) intentions and the (current) loyalty behavior (McAlexander et al. 2002; Oliver 1999; Watson et al. 2015). Furthermore, as Fig. 3 shows, these theoretical elements have both been defined in terms of four dimensions (i.e., repurchase, cross-buying, positive WOM/recommendation, and price increase acceptance/tolerance). Recent empirical

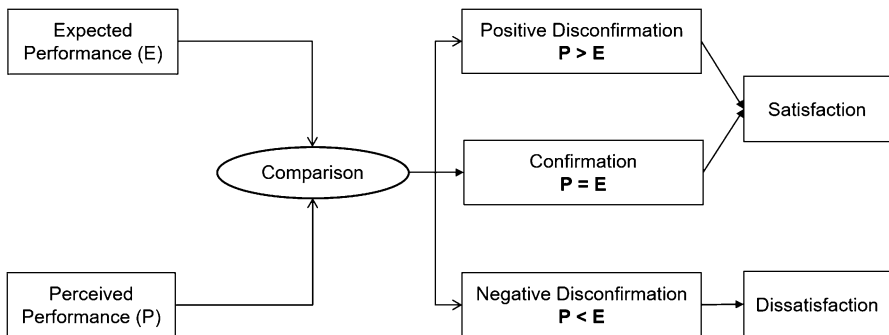


Fig. 2 Conceptualization of customer satisfaction: the Confirmation-Disconfirmation Paradigm. (Adapted from Boshoff 1997)

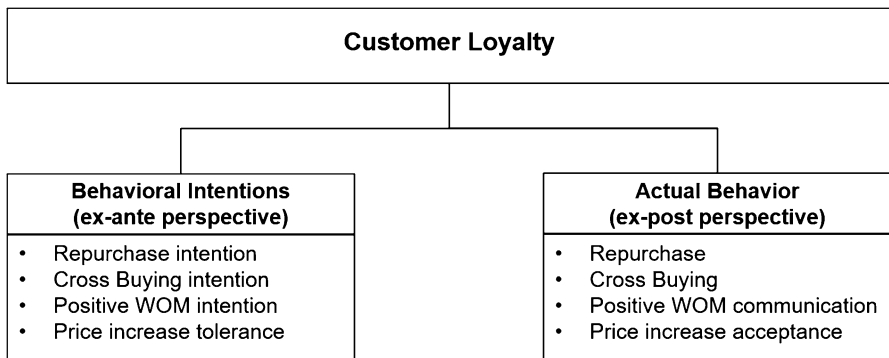


Fig. 3 Conceptualization of Customer Loyalty. (Adapted from Bruhn 2016)

findings showing that loyal customers in many situations tend to pay less, not more, illustrate the relevance of conceptualizing customer loyalty via multiple dimensions and differentiating between intentions and actual behavior, instead of merely looking at repurchase-related variables (see Umashankar et al. (2017) or Wieseke et al. (2014) for detailed overviews).

Measuring Customer Satisfaction

Customer satisfaction can be measured in a variety of ways (Homburg and Fuerst 2010). Due to several advantages in terms of flexibility and comprehensiveness, most academics and practitioners today measure customers' satisfaction through surveys, using and adapting established scales (Hayes 2008; Zairi 2000). Against this background, we discuss in this section how customer satisfaction can be assessed via scales in surveys and then review alternative approaches which can improve upon traditional customer satisfaction scales.

Survey Scales

As previously outlined, there are two conceptual perspectives to customer satisfaction that differ with respect to their reference object: the transaction-specific perspective focuses on individual transactions, whereas the cumulative perspective centers on the entire relationship. Research has shown that customer satisfaction can be operationalized according to both perspectives in surveys which have resulted in the creation of a variety of scales (Hayes 2008; Peterson and Wilson 1992). In the following, we will provide an overview of the leading customer satisfaction scales and present guidelines of how to adapt and choose between them.

Adopting a transaction-specific perspective, Oliver (1980) developed one of the first multi-item customer satisfaction scales. As Table 1 shows, this scale encompasses six reflective items, which have been used and adapted at various times in the marketing literature. For instance, Bearden and Teel (1983) adapted Oliver's (1980) six items to their research context, and then, due to problems with the scale's psychometric properties, reduced the scale to four main items. More recently, Homburg et al. (2005) and Homburg et al. (2006) drew on Oliver (1980) and Bearden and Teel (1983) to measure customers' satisfaction with a specific transaction with 4 items on an 11-point Likert scale. Moreover, various recent measurements of a transaction-specific customer satisfaction in marketing research use further reduced, adapted, and more efficient forms of the Oliver (1980) scale; some even draw on single-item scales (e.g., Chandrashekar et al. 2007).

Adopting a cumulative perspective, Cannon and Perreault Jr. (1999) provided a multi-item customer satisfaction scale measuring the satisfaction with the entire relationship (see Table 1). This scale was further developed by Homburg and Stock (2004) and Homburg et al. (2011). In synthesizing this literature analysis, findings reveal that customers' satisfaction with a particular transaction and satisfaction with the entire relationship are generally measured with a multi-item

Table 1 Examples of leading customer satisfaction scales

Authors (year)	Reference object	Type of scale	Items (item reliabilities, if specified)	Other psychometric properties
Oliver (1980)	Transaction	Not specified	I am satisfied with my decision to get or not to get a flu shot If I had it to do all over again, I would feel differently about the flu shot program (R*) My choice to get or not to get a flu shot was a wise one I feel bad about my decision concerning the flu shot (R*) I think that I did the right thing when I decided to get or not to get the flu shot I am not happy that I did what I did about the flu shot (R*)	CA = 0.82
Homburg et al. (2005)	Transaction	11-point Likert scale	All in all, I would be satisfied with this restaurant [experience] The restaurant [experience] would meet my expectations The earlier scenario compares to an ideal restaurant experience Overall, how satisfied would you be with the restaurant visit just described?	CR = 0.98 (study 1) CA > 0.94 (study 2)
Cannon and Perreault (1999)	Relationship	7-point Likert scale	Our firm regrets the decision to do business with this supplier (R*) (0.50) Overall, we are very satisfied with this supplier (0.80) We are very pleased with what this supplier does for us (0.87) Our firm is not completely happy with this supplier (R*) (0.59) If we had to do it all over again, we would still choose to use this supplier (0.59)	CA = 0.84 AVE = 0.67
Homburg and Stock (2004)	Relationship	5-point Likert scale	We are very pleased with the products and services that this company delivers (0.74) We enjoy collaborating with this company (0.71) On an overall basis, our experience with this company has been positive (0.84) This company is first choice for us for the purchase of these products and services (0.53) On an overall basis, we are satisfied with this company (0.85)	CA = 0.91 CR = 0.92 AVE = 0.55

(continued)

Table 1 (continued)

Authors (year)	Reference object	Type of scale	Items (item reliabilities, if specified)	Other psychometric properties
Homburg et al. (2011)	Relationship	7-point Likert scale	We are very pleased with the products and services of company X (0.61) We intensively enjoy collaborating with company X (0.76) On an overall basis, our experience with company X has been very positive (0.85) On an overall basis, we are very satisfied with this company (0.92)	CA = N/A CR = 0.94 AVE = 0.78
Fuerst (2012)	Relationship	7-point semantic differential	How satisfied are you with firm XYZ? How advantageous do you consider the relationship with firm XYZ? How well does firm XYZ fulfill your expectations?	N/A

Note: R* = reverse coded; N/A = not available

reflective Likert scale with an uneven amount of scale points (e.g., seven scale points ranging from “strongly disagree” to “strongly agree”). The scales provided in Table 1 can serve as a source for item selection, but the items obviously need to be adapted to the specific transaction of interest. See Jarvis et al. (2003) for guidance of how to choose between uneven and even scale points and select the exact number of scale points in different situations, and Bergkvist and Rossiter (2007) for potential problems and advantages of single-item scales (“Design and Process of Survey Research”).

In addition, previous work across the two conceptual perspectives often recommends measuring customer satisfaction at two different levels: the overall level and the detailed level (Churchill Jr. and Surprenant 1982; Homburg and Fuerst 2010; Rust and Zahorik 1993). This recommendation is because an entity (e.g., a transaction or an overall relationship) and customers’ satisfaction with it may encompass various aspects. Hence, knowledge regarding which of the single aspects account for how much of the overall satisfaction score may provide actionable implications for enhancing customers’ future satisfaction levels (Hayes 2008).

For instance, according to the cumulative perspective, customer satisfaction at the overall level may refer to the customer’s total satisfaction with the entire customer-firm relationship. (Note that customer satisfaction could also be measured at different levels according to the transaction-specific perspective. For instance, the overall level could relate to the satisfaction with the entire transaction (e.g., “how satisfied are you with the purchase of the new iPad overall?”) and the detailed level

could refer to the functionalities of the product (e.g., the features and the usability) or the purchase process (e.g., consulting by the sales rep, payment modes, financing options, etc.) As Table 1 shows, overall customer satisfaction is typically measured on a Likert scale with four to five reflective items. Due to the reflective nature of the items, a customer's satisfaction with the overall customer-firm relationship is specified as the average of all chosen scale items (see chapter ► [“Crafting Survey Research: A Systematic Process for Conducting Survey Research”](#) by Klarmann and Homburg in this handbook for details). In aggregating the overall-level satisfaction of all survey participants, firms can compute the customer satisfaction index (i.e., the average overall customer satisfaction). To enhance the comparability of customer satisfaction indices between different satisfaction measurements, it is recommended to keep the items measuring overall-level customer satisfaction as consistent as possible across measurements and time. Moreover, as previous work has pointed out, firms may choose to rescale customer satisfaction indices to a scale ranging from 0 to 100 to facilitate interpretation and discussion of results (Fuerst 2012; Griffin et al. 1995).

Customer satisfaction at the detailed level in this example refers to the customer's satisfaction with specific performance aspects of the firm (e.g., customer service, complaint handling, or satisfaction with a product) (also see Grigoroudis and Siskos 2009; Rust and Zahorik 1993). Customer satisfaction at the detailed level could also be assessed via reflective multi-item scales (which could lead to a higher validity and reliability of the assessment). However, many customer satisfaction measurements interested in assessing the detailed level draw on single items to ensure parsimony of the measurement and to increase response rates (Fuerst 2012; Hayes 2008). Similar to the overall level, customers' satisfaction with each performance aspects can be aggregated across all survey participants once the individual responses are collected. Comparing the overall customer satisfaction index with the indices of the detailed performance aspects may provide important explanations for the level of the customer satisfaction index and, potentially, indicate first levers for improving the customer satisfaction index in the future (Diamantopoulos 2011; Rust and Zahorik 1993; Homburg and Klarmann 2012).

Measuring customer satisfaction at the detailed level can be conducted in two steps (Fuerst 2012). In the first step, firms should carefully analyze their offerings and identify all major functionalities that may influence the customers' satisfaction in order to design a comprehensive measurement (Griffin et al. 1995; Homburg and Klarmann 2012; Rust and Zahorik 1993). The relevant functionalities may vary substantially according to the product type, the industry, or company-specific factors (Homburg and Fuerst 2010). For example, a limousine transportation service provider might want to assess the quality and response time of the service center, punctuality of service delivery, as well as integrity and commitment of the drivers, whereas a car manufacturer might rather put emphasis on product quality, brand reputation, and satisfaction with after sales service. In the second step, firms should then specify all determinants of the identified functionalities. For instance, if the aforementioned car manufacturer has identified product quality, brand reputation, and after sales service as the critical functionalities, the manufacturer now needs to specify the drivers of these functionalities (e.g., for the functionality after sales

service: speed of service, behavior of service personal, possibilities to complain, and quality of service results). Figure 4 demonstrates how customer satisfaction can be operationalized at an overall level and the detailed level, using a different example of a private bank.

Finally, in addition to assessing the level of customer satisfaction, it is a focal aim of customer satisfaction measurements to identify the most critical drivers of customer satisfaction (Morgan et al. 2005). As Gustafsson and Johnson (2004) show, a customer satisfaction measurement encompassing the overall and detailed level can be easily used to make this identification. Figure 5 summarizes one of the advanced methods to conduct such an evaluation: structural equation modeling (see also Gustafsson and Johnson (2004) and Homburg and Klarmann (2012) for more details and overviews of alternative methods). As this figure further indicates, by using structural equation modeling, firms can receive insights regarding the strength of the satisfaction drivers by looking at the standardized path coefficients (see the chapter “► Structural Equation Modeling” by Hans Baumgartner and Bert Weijters in this handbook for details).

Other Measurement Approaches

As prior work has demonstrated, there are various other approaches to assess customer satisfaction (Brandt and Reffett 1989; Bruhn 2003; Van Doorn and Verhoef

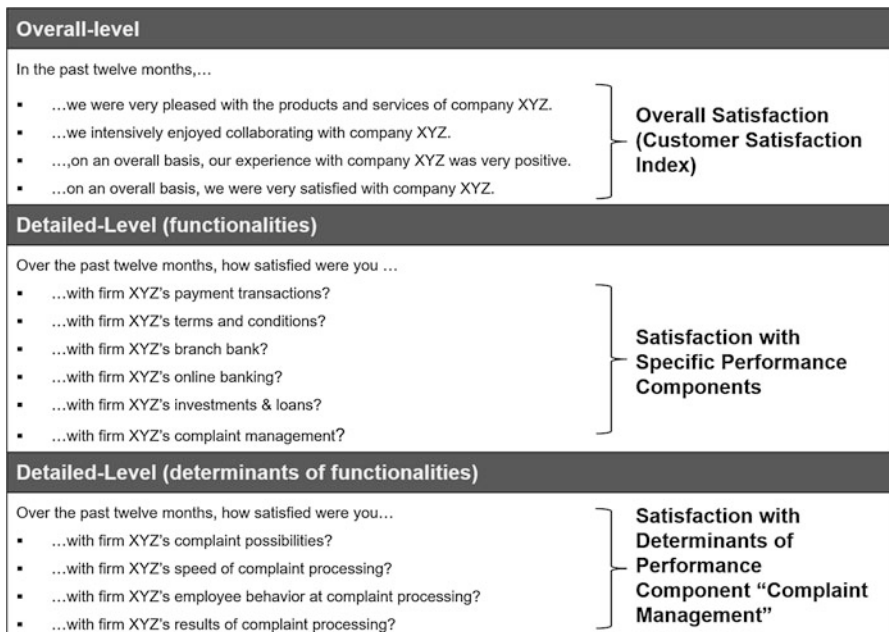


Fig. 4 Measuring customer satisfaction – example of a private bank. (Adapted from Fuerst 2012, p. 134 ff)

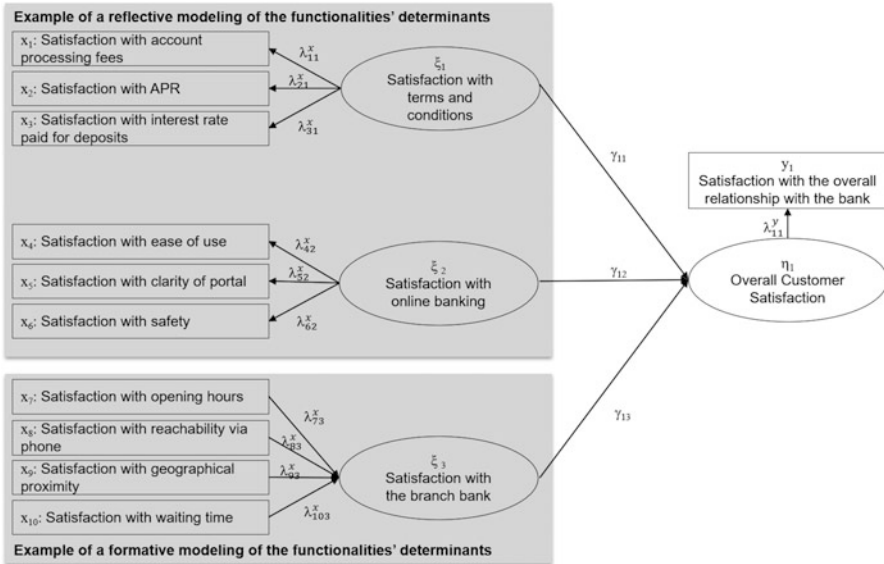


Fig. 5 Simplified structural equation model for determining the strength of satisfaction drivers. (Adapted from Homburg and Klarmann 2012, p. 204)

2008). The following paragraphs will focus on two of the most popular measurement approaches: complaint analysis and focus groups (see Bruhn (2003) for the depiction of further and more rare methods). However, because these approaches generally relate to a specific event or transaction (McCollough et al. 2000; Smith and Bolton 1998), these approaches have various disadvantages as compared to the above-described direct satisfaction ratings via survey scales, e.g., in terms of measuring customer satisfaction comprehensively. Thus, firms generally amend their customer satisfaction surveys with these approaches and use results from the complaint analysis and focus groups for a selective improvement of their offerings (Fuerst 2012).

Prior work has shown that customer suggestions and complaints can be a valuable implicit source for analyzing the underlying factors that determine customer satisfaction levels (Homburg and Fuerst 2005; Singh and Pandya 1991). For instance, if a firm receives many complaints concerning the availability of its call center, it may start to systematically analyze the complaints and the performance data of the call center in order to detect and ultimately eliminate the underlying cause of the problem. Likewise, if a company receives an increased amount of suggestions regarding the functionalities of their products, these suggestions may provide valuable information on how to improve the products' handling in the future, adding to the future satisfaction of customers. An advantage of analyzing customer suggestions and complaints is that the company can utilize existing data to get insights on latent problems that affect customer interactions (Grigoroudis and Siskos (2009); see also Homburg and Fuerst (2005) for a detailed discussion of complaints in the

context of customer satisfaction and customer loyalty). However, a major drawback of this method is that usually only a few customers actually complain and, thus, various existing deficits may not become obvious with this method (Richins 1983). Thus, we recommend that firms use the systematic analysis of customer suggestions and complaints in addition to the above-described direct customer satisfaction measurement via scales.

Moreover, firms can use focus groups or other more qualitative research approaches to gain more detailed insights into the underlying reasons for (the lack of) customer satisfaction. More precisely, such qualitative approaches focus on specific customer experiences (i.e., incidents), such as the contact with the service helpdesk of the firm, being advised by a salesperson, or using the product for the first time (Bruhn 2003). Firms can employ the critical incident technique in order to analyze these decisive moments in the customer-firm relationship: these incidents constitute deviations from the customers' "business-as-usual" mindset, which might affect their evaluation of entire business relationship (Gremler 2004; Van Doorn and Verhoef 2008). However, as the incident-related approaches focus on one specific touch point or transaction with the firm, they can hardly be used to assess the cumulative or overall satisfaction (Brandt and Reffett 1989; Bruhn 2003). Thus, in line with recommendations in the methodological research (Davis et al. 2011), we recommend that firms utilize focus groups in addition to customer satisfaction surveys to gain additional in-depth insights on selected issues.

Measuring Customer Loyalty

Overview

As shown in section "[Conceptual Background](#)" above, customer loyalty is generally conceptualized as encompassing two theoretical elements (i.e., loyalty intentions and actual loyalty behaviors). Moreover, both theoretical elements can be operationalized alongside four dimensions: repurchase, cross-buying, positive WOM/recommendation, and price increase acceptance/tolerance. The researcher can use both objective data – e.g., from the company's CRM system – and subjective data – e.g., from customer surveys – in order to gather information on customer loyalty. However, assessing actual loyalty behavior in surveys is problematic, because, as Sheppard et al. (1988) demonstrate, stated behavioral intentions are often a weak predictor of actual behavior (e.g., due to unexpected events and factors of the social environment). Thus, prior research recommends focusing on objective data to assess customers' actual loyalty behavior (Hayes 2008; Kumar and Shah 2004; Peters et al. 2010). Moreover, prior work also recommends to focus on subjective data (e.g., surveys) to assess loyalty intentions, because these intentions are latent and subjective in nature (Hayes 2008). In adopting this view, we created Fig. 6, which provides an overview of the operationalization of attitudinal and

Dimension	Behavioral Intentions (<i>surveys</i>)	Actual Behavior (<i>databases</i>)
Repurchase	In how far... ...are you planning to buy the product/service from firm XYZ again? ...are you intending to stay customer at firm XYZ? ...are you planning to increase the share of products/services purchased from firm XYZ?	<ul style="list-style-type: none"> ▪ Repurchase rate ▪ Share of wallet ▪ Larger cart
Cross-buying	In how far... ...are you planning to buy additional/other products/services from firm XYZ? ...is it an option for you to buy product categories/service categories from firm XYZ?	<ul style="list-style-type: none"> ▪ Range of product categories purchased ▪ Dollar amount spent on additional product categories
Positive WOM intention/ communication	In how far... ...are you intending to recommend firm XYZ to your business partners? ...will you recommend firm XYZ to your acquaintances and friends?	<ul style="list-style-type: none"> ▪ Number of referrals ▪ Valence of referrals
Price Increase Tolerance/ Acceptance	<ul style="list-style-type: none"> ▪ Would you tolerate a 2/5/10/20 percent price increase and still buy products/services from firm XYZ? ▪ By how much could the prices of firm XYZ increase such that you would still buy from it in the future? 	<ul style="list-style-type: none"> ▪ Number of customers leaving/staying at the firm after a 2/5/10/20 percent price increase ▪ Number of discount requests ▪ Number of pricing complaints

Fig. 6 Measuring customer loyalty intentions and behavior. (Adapted from Homburg and Fuerst 2010)

behavioral customer loyalty. In the following section, we discuss approaches to measuring customer loyalty in greater detail.

Loyalty Intentions

Firms should employ subjective approaches to derive insights with respect to customers' loyalty intentions (Watson et al. 2015). This recommendation is because subjective approaches are particularly suitable to uncover latent constructs and underlying subjective factors (Hayes 2008; Klarmann 2008). Similar to measuring customer satisfaction, there are two different approaches to measuring customer loyalty intentions in surveys (see Table 2). First, there is the aggregated approach, which tries to assess the overall customer loyalty with up to five or more reflective items (e.g., Watson et al. 2015). Applications of the aggregated approach assess the customer's loyalty intentions through indicators that essentially target the same underlying latent loyalty construct (i.e., collection of attitudes aligned with a series of purchase behaviors that systematically favor one entity over competing entities). Second, as also shown in Table 2, there is the disaggregated approach. This approach builds more directly on the above-described conceptualization of customer loyalty via multiple dimensions. Scales following this approach therefore aim to assess customer loyalty intentions for each of its disaggregated dimensions. For instance, Homburg et al. (2011) focus on three dimensions of the customer loyalty concept and measure each dimension via two items.

Table 2 Examples of leading customer loyalty intention scales

Authors (year)	Approach of scale	Type of scale	Items (item reliabilities, if specified)	Other properties
Brakus et al. (2009)	Overall loyalty	7-point Likert scale	In the future, I will be loyal to this brand I will buy this brand again This brand will be my first choice in the future I will not buy other brands if this brand is available at the store I will recommend this brand to others	N/A
Watson et al. (2015)	Overall loyalty	N/A	I prefer [target] over competitors I enjoy doing business with [target] I consider [target] my first preference I have a positive attitude toward [target] I really like [target]	N/A
Zeithaml et al. (1996)	Loyalty dimensions	7-point likelihood scale	Say positive things about XYZ to other people Recommend XYZ to someone who seeks your advice Encourage friends and relatives to do business with XYZ Consider XYZ your first choice to buy services Do more business with XYZ in the next few years Do less business with XYZ in the next few years (R) Take some of your business to a competitor that offers better Continue to do business with XYZ if its prices increase somewhat Pay a higher price than competitors charge for the benefits you currently receive from XYZ Switch to a competitor if you experience a problem with XYZ's service Complain to other customers if you experience a problem with XYZ's service Complain to external agencies, such as the Better Business Bureau, if you experience a problem with XYZ's service Complain to XYZ's employees if you experience a problem with XYZ's service	N/A
Homburg et al. (2011)	Loyalty dimensions	7-point Likert scale	Customer intentions to repurchase We consider company X as our first choice for the purchase of	CR = 0.81 AVE = 0.60

(continued)

Table 2 (continued)

Authors (year)	Approach of scale	Type of scale	Items (item reliabilities, if specified)	Other properties
			<p>such products and services (0.49) We intend to stay loyal to company X (0.71)</p> <p>Customer intentions to increase share of wallet We intend to do more business with company X in the future (0.77) We intend to additionally purchase other products and services from company X in the future (0.51)</p> <p>Customer word of mouth We recommend company X to other people (e.g., customers, business partners, friends) (0.64) We say positive things about company X to other people (e.g., customers, business partners, friends) (0.82)</p>	
<p>Homburg and Fuerst (2010)</p>	<p>Loyalty dimensions</p>	<p>7-point semantic differential</p>	<p>Repurchase. In how far.are you planning to buy the product/service from firm XYZ again? . . .are you intending to stay customer at firm XYZ? . . .are you planning to increase the share of products/services purchased from firm XYZ?</p> <p>Cross-buying. In how far.are you planning to buy additional/other products/services from firm XYZ? . . .is it an option for you to buy product categories/service categories from firm XYZ?</p> <p>Recommendation. In how far.are you intending to recommend firm XYZ to your business partners? . . .will you recommend firm XYZ to your acquaintances and friends?</p> <p>Price increase acceptance/ tolerance Would you tolerate a 2/5/10/20 percent price increase and still buy products/services from firm XYZ? By how much could the prices of firm XYZ increase such that you would still buy from it in the future?</p>	<p>N/A</p>

Note: (R) = reverse coded; N/A = not available

As an alternative to customer loyalty scales, many companies today have adopted the so-called Net Promoter Score (Reichheld 2003). As Keiningham et al. (2007) explain, the “Net Promoter is a metric derived from survey responses to a recommend likelihood question. Respondents who provide a rating of 9–10 are classified as ‘promoters’; respondents who provide a rating of 6 or lower are classified as ‘detractors.’ Net Promoter is calculated by subtracting the proportion of a firm’s detractors from its proportion of promoters (i.e., Net Promoter = promoters – detractors)” (p. 39). Although the Net Promoter Score is widely spread in practice and appears to be intuitive and efficient, research results regarding the Net Promoter Score are mixed (Keiningham et al. 2007; Morgan and Rego 2006). For instance, in their extensive analysis of various customer loyalty measures, Keiningham et al. (2007) find “no support for the claim that Net Promoter is the single most reliable indicator of a company’s ability to grow” (p. 45). Hence, we warn firms to exclusively rely on the Net Promoter Score for assessing their customers’ loyalty but instead recommend to integrate it as one component to their measurement system (e.g., as an alternative measure for the recommendation facet of loyalty).

Finally, previous research indicates that customers’ loyalty intentions and customers’ satisfaction levels are usually assessed within the same survey. Figure 7 provides an overview of decisions that need to be taken when designing a customer survey with sections on satisfaction and loyalty. See the chapter ► “Crafting Survey Research: A Systematic Process for Conducting Survey Research” by Vomberg and Klarmann in this handbook for more detailed guidance on the design, process, and evaluation of customer surveys. For interesting application examples of customer satisfaction and loyalty surveys, see Hayes (2008).

Loyalty Behavior

As prior research shows, firms should generally employ objective approaches to measure customer loyalty behavior (Peters et al. 2010; Mellens et al. 1996; Kumar and Shah 2004). This tendency is because the objective approaches draw on directly observable numbers that are not biased due to subjective perceptions, incomplete memory of events, or unexpected events (McNeal 1969; Sheppard et al. 1988). Thus, the objective approaches allow a more valid, reliable, and timely assessment of actual customer behavior (Mellens et al. 1996). Due to the increasing availability of objective data (e.g., from CRM systems), the objective loyalty measurement approaches have gained importance for firms over the past couple of years (Sarstedt and Mooi 2019). Hence, firms can use such data nowadays to gain insights on most of the different dimensions of customer loyalty behavior (i.e., repurchase, cross-buying, price increase acceptance, and positive WOM behavior) by developing and monitoring appropriate key performance indicators (KPIs). In the following, we first explain how firms can use several general databases to generate information on customer loyalty behavior before we discuss loyalty programs, which are for many companies the most valuable data source for loyalty behavior information.

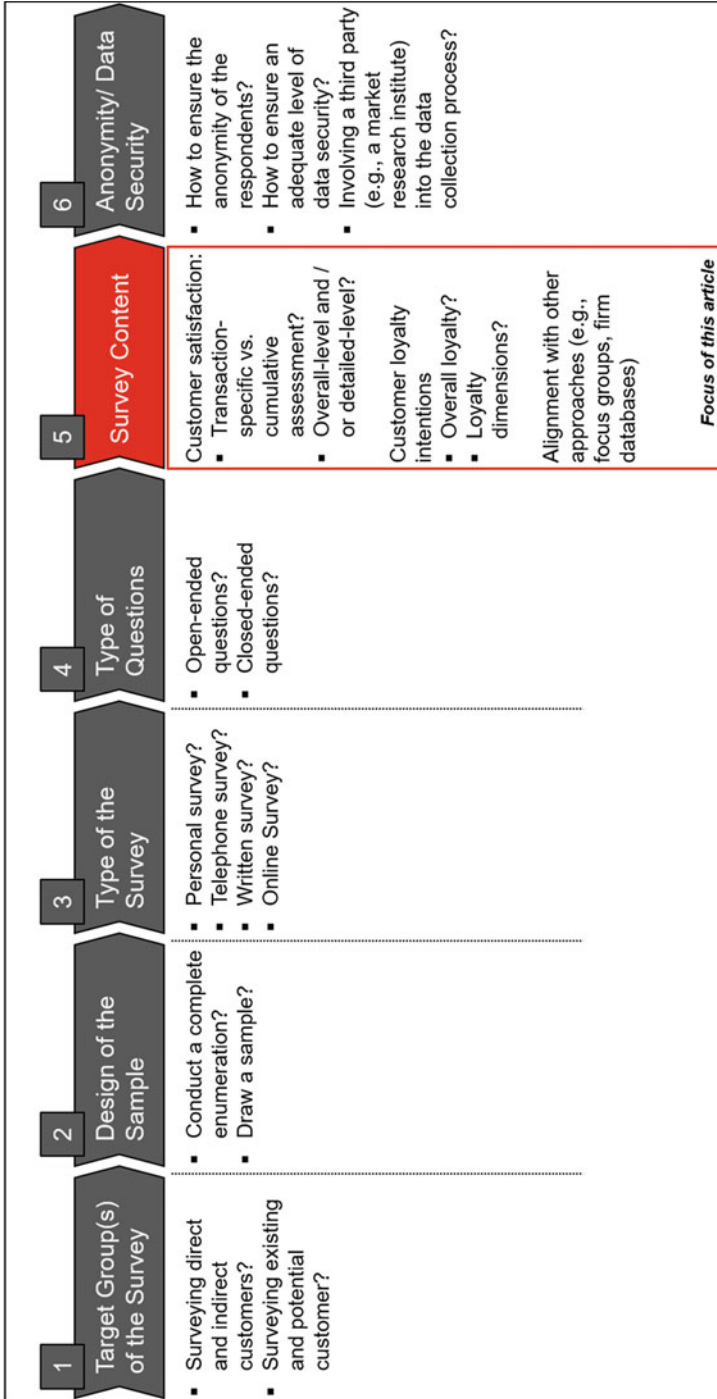


Fig. 7 Overview of decisions and process of a customer satisfaction and loyalty survey

General Databases

As prior work has pointed out, there are two types of objective measurement approaches which differ according to the orientation of the data source (i.e., internal objective and external objective approaches) (Fuerst 2012; Homburg and Fuerst 2010). While the former draw on company-internal data sources (such as CRM systems, central contact management databases, or ERP systems), the latter approaches exploit company-external data sources (such as external panel data). Internal objective approaches provide especially valuable information on actual customer loyalty behavior, since they generally capture individual customer behavior and therefore allow for very detailed and precise evaluations (Sarstedt and Mooi 2019). Thus, these approaches can be used to gain transparency on three of the four behavioral dimensions of customer loyalty (i.e., repurchase behavior, cross-buying behavior, and price increase acceptance). Moreover, depending on the CRM system configuration, firms may also be able to gain first insights regarding the fourth behavioral loyalty dimension (i.e., positive WOM communication), e.g., by monitoring the proportion of new customers that came through recommendations of existing customers. Figure 8 illustrates how firms can derive meaningful KPIs for the different dimensions of customer loyalty behavior by utilizing various objective data sources (e.g., CRM data, central contact management data, or scanner data).

In addition, firms may want to amend the internal objective techniques by external objective techniques to gain more nuanced and comprehensive insights in all customer loyalty behavior dimensions. There are two different kinds of external data that have relevance for the measurement of customer loyalty behavior: panel data (Sarstedt and Mooi 2019) and social media data (Ma et al. 2015). First, when using panel data, firms usually generate information at an aggregated level, such as the market, segment, or company level. This aggregated information on customer

Dimension to be Measured:	Repurchase Behavior	Cross-Buying Behavior	Price Increase Acceptance	Positive WOM Communication
Key Question:	How many customers do actually buy again?	How many of the existing customers start buying from other sales units?	Are the customers willing to continue buying from the firm despite price increases?	How likely is it that a customer recommends the firm to family/ friends/ acquaintances?
Name of KPI:	<i>Repurchase Rate</i>	<i>Cross Buying Extent</i>	<i>Price Elasticity of Demand</i>	<i>Recommendation extent</i>
Definition of KPI:	Number of Repurchases / Number of Initial Purchases	<ul style="list-style-type: none"> ▪ Number of sales units a customer buys from ▪ Dollar amount a customer spends in additional sales units 	$(dQ / Q) / (dP / P)$	<ul style="list-style-type: none"> ▪ Amount of recommendations ▪ Positive WOM score
Data Source:	e.g., Scanner Data, Loyalty Program Data, further Data from the CRM Tool of the Sales Department, Social Media, or Panel Data			

Fig. 8 KPIs to measure actual loyalty behavior

loyalty behavior can be valuable. For instance, firms can use this information as benchmarks for their more specific analyses or to detect general trends and tendencies (Fornell et al. 1996). Most panel data with relevance for measuring customer loyalty are collected by the leading market research agencies (e.g., Nielsen, GfK) and need to be purchased. Second, to gain further insights on customer loyalty behavior, such as on product or brand level, firms can use social media analyses. More precisely, firms can employ a social media crawler (e.g., Brandwatch, Social Crawlytics, spirm3r, or PromptCloud) to retrieve all relevant discussions as entries from various social media, such as blogs, communities (e.g., Twitter, Facebook), business networks (e.g., LinkedIn, XING), photo sharing (e.g., Instagram, Flickr), as well as products and services reviews (e.g., Amazon.com). The social media crawler is able to deliver the relevant social media entries on a regular basis (e.g., weekly or monthly) such that firms can systematically evaluate these entries to gain further insights on the recommendation behavior of their customers, at an aggregated level or a more disaggregated level depending on the abilities of the hired social media crawler. See Fig. 9 for an example of automatic outputs of a social media crawler (e.g., daily mentions, author profession, and brand sentiment). These automatic outputs do not provide direct results on customer loyalty. Instead, to gain more direct insights on customer loyalty behavior, more sophisticated text analysis tools (e.g., dictionaries or machine learning algorithms) need to be employed (see chapter ► “Automated Text Analysis” by Humphreys in this handbook for more details on text analysis). As Fig. 9 also indicates, with adequate methods, such as sentiment analysis (Homburg et al. 2015), firms could even gain objective insights on customers’ attitudes, such as satisfaction and loyalty intentions.

Loyalty Programs

Overview of loyalty programs One of the most common ways that firms measure customer loyalty behavior is through their loyalty programs. Since their beginnings with airlines in the early 1980s, loyalty programs (LPs) as marketers know them today have seen a tremendous increase in participation rates and are now prevalent in a variety of industries. At their core, LPs offer consumers rewards in exchange for providing firms with detailed transaction data to be used to develop marketing strategies that are designed to optimize customer engagement with the firm. Given their importance in measuring loyalty behavior, we briefly review the design and modeling issues associated with loyalty programs.

The design of loyalty programs encompasses five key components: (1) membership requirements, (2) program structure, (3) point structure, (4) reward structure, and (5) program communication (Breugelmans et al. 2015). These components are common to all types of LPs, but the variations in program design across firms are immense. Of the five components, reward structure has been given the most attention in the marketing literature. Currently, there are two types of rewards schemes in loyalty programs: customer tier and frequency based. Customer-tier LPs, which are popular in the airline, hotel, and casino industry, group customers into segments according to their actual or potential purchase volume or profitability, with higher tiers receiving some form of preferential treatment (Blattberg et al. 2008; Kumar and

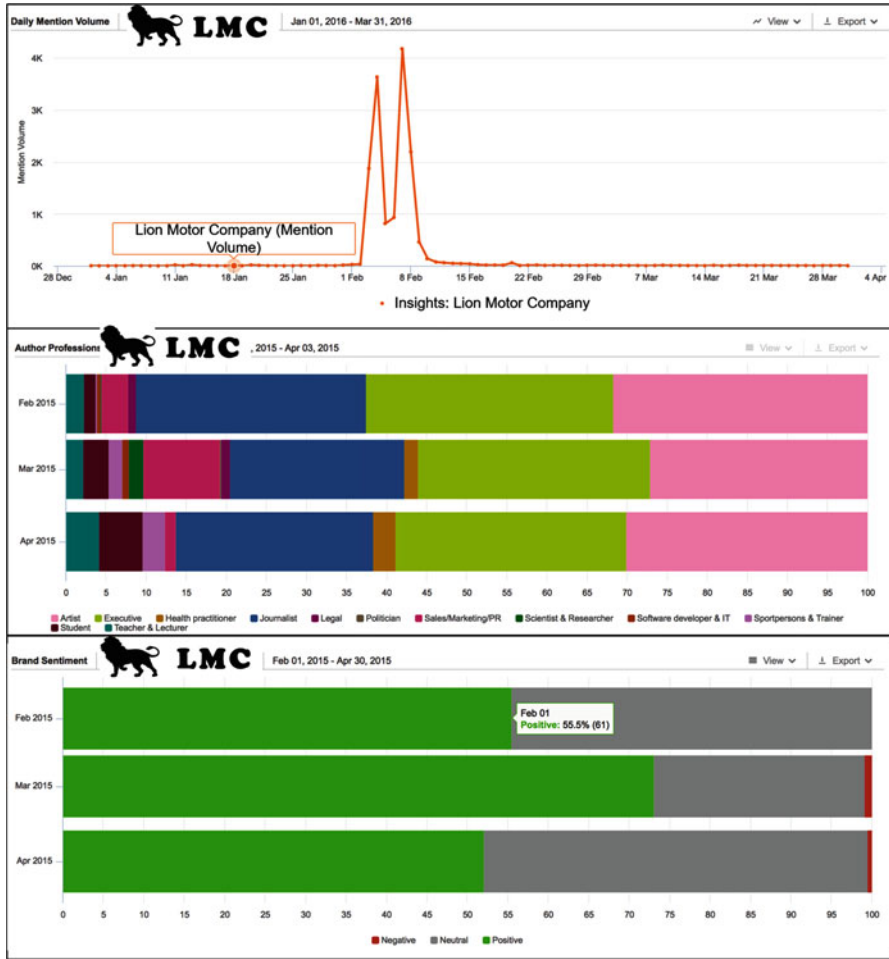


Fig. 9 Example report from a social media crawler. (Fictitious example based on a sample output provided by Brandwatch)

Shah 2004). In frequency-based LPs, customers earn rewards as a function of purchase volume.

In spite of the popularity of loyalty programs, their effectiveness has long been subject to debate. This debate has centered on the costs of giving discounts and perks to the loyal customers, as well as the costs of administering the program itself, and whether these costs are justified by increases in spending by those customers. Loyalty programs have the potential to increase profits by increasing switching costs for existing customers, stealing business from rivals, or through second-degree price discrimination. They may also indirectly increase profits by increasing customers perceptions of the firm, by generating customer data that can be used for targeted promotions or CRM, or by exploiting agency issues such as flights booked

by business travelers and paid for by their employers (Roehm et al. 2002; Dreze and Nunes 2008; Verhoef 2003; Shugan 2005). Empirical studies of whether loyalty programs actually do increase profits have found mixed results. Verhoef (2003) finds that the effects are positive but very small, DeWulf et al. (2001) find no support for positive effects of direct mail, Shugan (2005) finds that firms gain short-term revenue at the expense of longer-term reward payments, and Hartmann and Viard (2008) find no evidence that loyalty programs create switching costs. For a more complete review on loyalty programs, see Bijmolt et al. (2011); Liu and Yang (2009); and McCall and Voorhees (2010). Regardless of the profit impact, one clear advantage of loyalty programs is that it allows firms to objectively track and monitor customer interactions with the firm.

Challenges in measuring customer loyalty through loyalty programs As discussed earlier, conceptualizing and measuring customer loyalty can be challenging. In this section, we review the common modeling issues encountered when attempting to measure customer loyalty behavior using an LP and discuss some of the common solutions where applicable. Broadly, these issues can be categorized as follows: endogeneity, complexity, and attribution.

From a modeling perspective, endogeneity is likely the most prevalent issue in measuring customer loyalty. At a high level, it must be recognized that the design of the LP itself is not a random outcome. While this is technically an endogeneity issue, it tends to be abstracted away in most research. In addition, while endogeneity related to price or targeted marketing decisions is a concern, these are relatively simple for a firm to address through randomized experiments. For a detailed discussion of endogeneity issues, see the chapter ► [“Dealing with Endogeneity: A Non-technical Guide for Marketing Researchers”](#) by Ebbes, Papies, and van Heerde in this handbook.

A more challenging, and common, scenario to address is when a firm attempts to measure customer loyalty behavior by assessing customer engagement with the loyalty program. This endogeneity issue stems from the fact that customers self-select into loyalty programs, which can lead to biased estimates of LP effectiveness. For example, if a firm compares spending behavior between customers who are enrolled in the loyalty program with those who are not, they will likely find that spending is higher from those who are enrolled. While this difference may be caused by the program itself, the more likely explanation is simply that customers who spend more with the firm are more likely to benefit from the LP and hence are more likely to enroll. A less obvious example is when a firm observes spending for the same customer both before and after joining the loyalty program. Here the firm may be tempted to conclude that changes in spend for each individual eliminate the endogeneity issue. However, again we have a selection issue: it is possible that customers joined the LP due to *anticipated* changes in spend. This will again lead to biased estimates of LP effectiveness.

A common approach to correct selection bias is to employ a Heckman two-stage correction method (Heckman 1979). In the classic bivariate selection problem, we have the following:

$$y_1^* = X_1\beta_1 + \varepsilon_1$$

$$y_2^* = X_2\beta_2 + \varepsilon_2$$

where y_2^* is the outcome of interest (e.g., customer spending) and is only observed if $y_1^* > 0$ (e.g., whether the customer decides to enroll in the LP). The stars indicate that these are latent variables – for example, a firm may not observe customer-level spending unless they are enrolled in the loyalty program. Problems arise in estimating β_2 if there is a nonzero correlation between ε_1 and ε_2 .

Under this specification, it can be shown that when

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim N \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

we have

$$\mathbb{E}[y_2|y_1^* > 0] = X_2\beta_2 + \rho\lambda(X_1\beta_1)$$

where $\lambda(z) = \phi(z)/\Phi(z)$, or the inverse Mills ratio. Clearly, failure to account for any nonzero correlation ρ will result in a biased estimate of β_2 .

To correct this bias, Heckman’s two-step estimator first runs a probit regression of y_1 onto $X_1\beta_1$ to get $\hat{\beta}_1$ in the first step and then runs an OLS of y_2 on $X_2\beta_2 + \rho\lambda(X_1\hat{\beta}_1)$ in the second stage. The estimated coefficient of the inverse Mills ratio, $\hat{\rho}$, may indicate whether or not sample selection correction is needed (so long as certain model assumptions are met). For more details on model assumptions, estimation, and interpreting the two-step estimator coefficients, see Heckman (1979) and Certo et al. (2016).

Building on this model, later research found that more flexible control functions could be implemented. For example, rather than use the inverse Mills ratio, instead include higher-order polynomials of $\hat{\beta}_1$. The intuition is that the flexible function will essentially replicate the inverse Mills ratio without the need for a probit in the first stage. See Ahn and Powell (1993) for more details or Ellickson and Misra (2012) for a recent application.

Besides endogeneity, another challenge in the analysis of loyalty program effectiveness is the complexity of the environment in which LPs typically operate. Analysts must recognize that LPs have numerous moving components, each which may influence customer loyalty, for example, the mere offering of a LP, the design of the rewards structure regardless of customer earnings and redemption activity (e.g., a relatively complicated versus simple tier or earnings structure), or strategic reactions from competition (Breugelmans et al. 2015). These are only a few of many potential issues that arise when attempting to model and quantify the effectiveness of a loyalty program and how it influences customer loyalty. Typically, researchers circumvent these challenges by either (1) conducting field experiments to isolate the causal effect of interest, (2) identify a natural experiment in situations where a field

experiment is infeasible, or (3) specify a structural model and use economic theory to identify the causal effect of interest.

Related to the complexities distinct to the LP itself is that of attribution among other firm activities. Loyalty programs do not manage customer engagement in a vacuum. Multiple touchpoints make it difficult to pin down exactly what is driving customer engagement, whether it be billboards, TV, print or radio advertising, or the LP itself. It is still not clear the extent to which the LP design interacts with these other components, and more work is needed in this area. These difficult attribution issues have not yet been solved and are only beginning to receive serious attention in research. For a short discussion of current work in this area, see Ascarza et al. (2017). If implemented properly, loyalty programs can be a powerful tool for firms to increase customer engagement with a firm. However, firms must carefully balance the costs of its implementation and management against the potential gains from additional customer information and more refined targeting abilities.

The Future of Managing Customer Satisfaction and Loyalty

The practice of measuring and managing customer satisfaction and loyalty is in a constant state of change (with “satisfaction” and “loyalty” now often referred to using the more general term “customer engagement”). However, these changes have occurred at an increasingly rapid pace, in part because the degree to which a firm can manage the customer engagement process is very much dependent on computational technology (e.g., processing speed, data storage costs, along with the labor to extract and analyze the data, to name a few).

These changes are most visible in the speed, customization, and rate in which customer satisfaction and loyalty is managed. For example, some firms are moving away from static marketing campaigns and ad hoc analyses in favor of systematic, real-time evaluation and tuning of the engagement process (see Schwartz et al. 2017; Hauser et al. 2009). Part of this has been driven by the speed and ease in which firms can customize offers, conduct A/B testing, and in general customize the customer’s experience with the firm.

Related, there has been a growing interest in applying machine learning methods to solve marketing problems. Even though machine learning has been around for decades, their use has accelerated in tandem with increases in computational power. A byproduct is that there has been, to some extent, a shift that emphasizes scalability and predictive accuracy in customer behavior rather than understanding causality of marginal effects. However, there has been strong interest in understanding how machine learning can work with econometrics (e.g., Athey 2018) or as a way to augment traditional marketing analyses (Ascarza 2018; Liu and Toubia 2018). In addition, some of the machine learning methods have allowed managers to deal with ultrafine, unstructured data such as text (e.g., social media content, as previously discussed) or voice (such as reviews or call center recordings) which would be cumbersome to process using traditional methods. Although it is a general consensus today that the potential of these methods for assessing customer satisfaction and

loyalty is very high, researchers are still in the early phase of determining how best to use these powerful tools and to the boundaries of their limitations.

Given the current environment, what does the future hold? It is not unreasonable to expect continued computation performance increases to drive much of the innovations in marketing and the customer engagement process. We are still a long way from optimizing a customer's lifetime marketing activity, which evolves over time and may be a function of competitor reactions. Relatively, there has been increased attention to how the sequence of outcomes itself influences customer engagement (e.g., reinforcement learning). It is simply a matter of time before even the most complex interactions can be modeled and processed fast enough to be of use to both practitioners and academics.

The key for marketing managers is to not allow advances in technology cloud the primary objective of increasing customer satisfaction and loyalty. Firms should implement technology to augment the customer relationship, rather than force the relationship process around current limitations in technology. More generally, a firm should avoid imposing any part of their management structure onto the relationship process (e.g., sending a customer complaint to a different department simply because that is how the organization is structured). As the field of customer engagement advances, the one unchanging factor is that the customers' experience is of paramount importance.

Concluding Remarks

Customer satisfaction and customer loyalty represent key constructs in marketing research and management. The measurement of customer satisfaction and loyalty is the basic requirement of a successful satisfaction and loyalty management. However, as there is a lack of transparency and guidance regarding the different traditional and novel measurement approaches, many firms currently struggle to identify the most appropriate tools and approaches for their particular situation. Against this background, this chapter has aimed at answering the following research questions: (1) How can firms conceptualize customer satisfaction and customer loyalty? (2) How can firms measure customer satisfaction? (3) How can firms measure customer loyalty?

To address these questions, this chapter has first summarized how customer satisfaction and customer loyalty are generally conceptualized and operationalized. Next, this chapter has outlined various important approaches and issues that relate to the measurement of customer satisfaction and loyalty on the basis of the previously introduced conceptual background. The chapter also discussed the current state and future of measuring both customer satisfaction and loyalty as a function of advances in technology and growing competitive pressure.

There is strong evidence that firms will need to focus on customer satisfaction and loyalty management with even greater emphasis in the upcoming years. Firms should continually reevaluate their approaches and tools for measuring customer satisfaction and loyalty and update when needed. Challenging your firm's current

measurement approaches using the ideas and tools presented in this chapter may serve as a first step toward implementing such a continuous improvement process.

References

- Ahn, H., & Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1–2), 3–29.
- Aksoy, L. (2013). How do you measure what you can't define? The current state of loyalty measurement and management. *Journal of Service Management*, 24(4), 356–381.
- Anderson, E. W. (1996). Customer satisfaction and price tolerance. *Marketing Letters*, 7(3), 265–274.
- Anderson. (2010). Meng marketing trends report 2010.
- Anderson, E. W., & Mittal, V. (2000). Strengthening the satisfaction-profit chain. *Journal of Service Research*, 3(2), 107–120.
- Anderson, E. W., Fornell, C., & Lehmann, D. R. (1994). Customer satisfaction, market share, and profitability: Findings from Sweden. *Journal of Marketing*, 58(3), 53–66.
- Anderson, E. W., Fornell, C., & Mazvanchery, S. K. (2004). Customer satisfaction and shareholder value. *Journal of Marketing*, 68(4), 172–185.
- Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1), 80–98.
- Ascarza, E., Fader, P. S., & Hardie, B. G. (2017). Marketing models for the customer-centric firm. In *Handbook of marketing decision models* (pp. 297–329). New York: Springer.
- Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda*. Chicago: University of Chicago Press.
- Brakus, J. J., Schmitt, B. H., & Zarantonello, L. (2009). Brand experience: what is it? How is it measured? Does it affect loyalty?. *Journal of Marketing*, 73(3), 52–68.
- Bain & Company (2013). Management Tools & Trends 2013. http://www.bain.de/Images/BAIN_BRIEF_Management_Tools_%26_Trends_2013.pdf. Accessed 25 July 2016.
- Bearden, W. O., & Teel, J. E. (1983). Selected determinants of consumer satisfaction and complaint reports. *Journal of Marketing Research*, 20(1), 21–28.
- Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44(2), 175–184.
- Bijmolt, T. H., Dorotic, M., Verhoef, P. C., et al. (2011). Loyalty programs: Generalizations on their adoption, effectiveness and design. *Foundations and Trends® in Marketing*, 5(4), 197–258.
- Blattberg, R., Kim, B., & Neslin, S. (2008). *Database Marketing: Analyzing and Managing Customers*. New York: Springer.
- Bolton, R. N. (1998). A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing Science*, 17(1), 45–65.
- Boshoff, C. (1997). An experimental study of service recovery options. *International Journal of Service Industry Management*, 8(2), 110–130.
- Brandt, D. R., & Reffett, K. L. (1989). Focusing on customer problems to improve service quality. *Journal of Services Marketing*, 3(4), 5–14.
- Brugelmanns, E., Bijmolt, T. H., Zhang, J., Basso, L. J., Dorotic, M., Kopalle, R., Minnema, A., Mijnlief, W. J., & Wunderlich, N. V. (2015). Advancing research on loyalty programs: A future research agenda. *Marketing Letters*, 26(2), 127–139.
- Brooke. (2016). Rewards, returns, and ringside seats. *Marketing News*, 50(6), 28–35.
- Bruhn, M. (2003). *Relationship marketing: Management of customer relationships*. Harlow: Pearson Education.
- Bruhn, M. (2016). *Kundenorientierung: Bausteine fuer ein exzellentes Customer Relationship Management (CRM)* (Vol. 50950). München: CHE Beck.
- Cannon, J. P., & Perreault, W. D., Jr. (1999). Buyer-seller relationships in business markets. *Journal of Marketing Research*, 36(4), 439–460.

- Certo, S. T., Busenbark, J. R., Woo, H.-s., & Semadeni, M. (2016). Sample selection bias and heckman models in strategic management research. *Strategic Management Journal*, 37(13), 2639–2657.
- Chandrashekar, M., Rotte, K., Tax, S. S., & Grewal, R. (2007). Satisfaction strength and customer loyalty. *Journal of Marketing Research*, 44(1), 153–163.
- Churchill, G. A., Jr., & Surprenant, C. (1982). An investigation into the determinants of customer satisfaction. *Journal of Marketing Research*, 19(4), 491–504.
- Davis, D. F., Golicic, S. L., & Boerstler, C. N. (2011). Benefits and challenges of conducting multiple methods research in marketing. *Journal of the Academy of Marketing Science*, 39(3), 467–479.
- DeWulf, K., Odekerken-Schröder, G., & Iacobucci, D. (2001). Investments in consumer relationships: A cross-country and cross-industry exploration. *Journal of Marketing*, 65(4), 33–50.
- Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *MIS Quarterly*, 35, 335–358.
- Dreze, X., & Nunes, J. C. (2008). Feeling superior: The impact of loyalty program structure on consumers' perceptions of status. *Journal of Consumer Research*, 35(6), 890–905.
- Ernst & Young (2011). The digitisation of everything: How organizations must adapt to changing consumer behaviour, available at: [http://www.ey.com/Publication/vwLUAssets/The_digitisation_of_everything_-_How_organisations_must_adapt_to_changing_consumer_behaviour/\\$FILE/EY_Digitisation_of_everything.pdf](http://www.ey.com/Publication/vwLUAssets/The_digitisation_of_everything_-_How_organisations_must_adapt_to_changing_consumer_behaviour/$FILE/EY_Digitisation_of_everything.pdf). Retrieved on 25 Jul 2016.
- Ellickson, P. B., & Misra, S. (2012). Enriching interactions: Incorporating outcome data into static discrete games. *Quantitative Marketing and Economics*, 10(1), 1–26.
- Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: Nature, purpose, and findings. *Journal of Marketing*, 60(4), 7–18.
- Fuerst, A. (2012). Verfahren zur Messung der Kundenzufriedenheit im Ueberblick. *Kundenzufriedenheit: Konzepte–Methoden–Erfahrungen*, 8, 123–154.
- Gremler, D. D. (2004). The critical incident technique in service research. *Journal of Service Research*, 7(1), 65–89.
- Griffin, A., Gleason, G., Preiss, R., & Shevenaugh, D. (1995). Best practice for customer satisfaction in manufacturing firms. *MIT Sloan Management Review*, 36(2), 87.
- Grigoroudis, E., & Siskos, Y. (2009). *Customer satisfaction evaluation: Methods for measuring and implementing service quality* (Vol. 139). Springer Science & Business Media.
- Gupta, S., & Zeithaml, V. (2006). Customer metrics and their impact on financial performance. *Marketing Science*, 25(6), 718–739.
- Gustafsson, A., & Johnson, M. D. (2004). Determining attribute importance in a service satisfaction model. *Journal of Service Research*, 7(2), 124–141.
- Halstead, D. (1999). The use of comparison standards in customer satisfaction research and management: A review and proposed typology. *Journal of Marketing Theory and Practice*, 7(3), 13–26.
- Hartmann, W., & Viard, B. (2008). Do frequency reward programs create switching costs? A dynamic structural analysis of demand in a reward program. *Quantitative Marketing and Economics*, 6(2), 109–137.
- Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website morphing. *Marketing Science*, 28(2), 202–223.
- Hayes, B. E. (2008). *Measuring customer satisfaction and loyalty: Survey design, use, and statistical analysis methods*. ASQ Quality Press.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Homburg, C., & Fuerst, A. (2005). How organizational complaint handling drives customer loyalty: An analysis of the mechanistic and the organic approach. *Journal of Marketing*, 69(3), 95–114.
- Homburg, C., Fürst, A. (2010). Überblick über die Messung von Kundenzufriedenheit und Kundenbindung. In M. Bruhn, & C. Homburg (Eds.), *Handbuch Kundenbindungsmanagement* (pp. 599–634). Wiesbaden: Gabler.
- Homburg, C., & Klarmann, M. (2012). Die indirekte Wichtigkeitsbestimmung im Rahmen von Kundenzufriedenheitsuntersuchungen: Probleme und Loesungsansatze. In C. Homburg

- (Eds.), *Handbuch Kundenzufriedenheit: Konzepte - Methoden - Erfahrungen*. Wiesbaden: Gabler.
- Homburg, C., & Stock, R. M. (2004). The link between salespeoples job satisfaction and customer satisfaction in a business-to-business context: A dyadic analysis. *Journal of the Academy of Marketing Science*, 32(2), 144.
- Homburg, C., Koschate, N., & Hoyer, W. D. (2005). Do satisfied customers really pay more? A study of the relationship between customer satisfaction and willingness to pay. *Journal of Marketing*, 69(2), 84–96.
- Homburg, C., Koschate, N., & Hoyer, W. D. (2006). The role of cognition and affect in the formation of customer satisfaction: A dynamic perspective. *Journal of Marketing*, 70(3), 21–31.
- Homburg, C., Mueller, M., & Klarmann, M. (2011). When should the customer really be king? On the optimum level of salesperson customer orientation in sales encounters. *Journal of Marketing*, 75(2), 55–74.
- Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and managing consumer sentiment in an online community environment. *Journal of Marketing Research*, 52(5), 629–641.
- Howard, J. A. & Sheth, J. N. (1969). *The theory of buyer behavior*. New York: John Wiley and Sons.
- Hunt, H. K. (1977). *Conceptualization and measurement of consumer satisfaction and dissatisfaction* (pp. 77–103). Cambridge, MA: Marketing Science Institute.
- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30(2), 199–218.
- Keiningham, T. L., Cooil, B., Andreassen, T. W., & Aksoy, L. (2007). A longitudinal examination of net promoter and firm revenue growth. *Journal of Marketing*, 71(3), 39–51.
- Klarmann, M. (2008). *Methodische Problemfelder der Erfolgsfaktorenforschung: Bestandsaufnahme und Empirische Analysen*. Ph.D. thesis.
- Kozinets, R. V., De Valck, K., Wojnicki, A. C., & Wilner, S. J. (2010). Networked narratives: Understanding word-of-mouth marketing in online communities. *Journal of Marketing*, 74(2), 71–89.
- Kumar, V., & Reinartz, W. (2012). *Customer relationship management: Concept, strategy, and tools*. Berlin: Springer Science & Business Media.
- Kumar, V., & Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century. *Journal of Retailing*, 80(4), 317–329.
- Kumar, V., Dalla Pozza, I., & Ganesh, J. (2013). Revisiting the satisfaction–loyalty relationship: Empirical generalizations and directions for future research. *Journal of Retailing*, 89(3), 246–262.
- Larivière, B., Keiningham, T. L., Aksoy, L., Yalgin, A., Morgeson, F. V., III, & Mithas, S. (2016). Modeling heterogeneity in the satisfaction, loyalty intention, and shareholder value linkage: A cross-industry analysis at the customer and firm levels. *Journal of Marketing Research*, 53(1), 91–109.
- Liu, J., & Toubia, O. (2018). A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6), 930–952.
- Liu, Y., & Yang, R. (2009). Competing loyalty programs: Impact of market saturation, market share, and category expandability. *Journal of Marketing*, 73(1), 93–108.
- Luo, X., & Homburg, C. (2007). Neglected outcomes of customer satisfaction. *Journal of Marketing*, 71(2), 133–149.
- Ma, L., Sun, B., & Kekre, S. (2015). The squeaky wheel gets the grease - an empirical analysis of customer voice and brand intervention on twitter. *Marketing Science*, 34(5), 627–645.
- McAlexander, J. H., Schouten, J. W., & Koenig, H. F. (2002). Building brand community. *Journal of Marketing*, 66(1), 38–54.
- McAlexander, J. H., Kim, S. K., & Roberts, S. D. (2003). Loyalty: The influences of satisfaction and brand community integration. *Journal of Marketing Theory and Practice*, 11(4), 1–11.

- McCall, M., & Voorhees, C. (2010). The drivers of loyalty program success: An organizing framework and research agenda. *Cornell Hospitality Quarterly*, 51(1), 35–52.
- McCullough, M. A., Berry, L. L., & Yadav, M. S. (2000). An empirical investigation of customer satisfaction after service failure and recovery. *Journal of Service Research*, 3(2), 121–137.
- McNeal, J. U. (1969). Consumer satisfaction-measure of marketing effectiveness. *MSU Business Topics-Michigan State University*, 17(3), 31–35.
- Mellens, M., Dekimpe, M., & Steenkamp, J. (1996). A review of brand-loyalty measures in marketing. *Tijdschrift voor economic en management*, 4, 507–533.
- Morgan, N. A., & Rego, L. L. (2006). The value of different customer satisfaction and loyalty metrics in predicting business performance. *Marketing Science*, 25(5), 426–439.
- Morgan, N. A., Anderson, E. W., & Mittal, V. (2005). Understanding firms customer satisfaction information usage. *Journal of Marketing*, 69(3), 131–151.
- Morgeson, F. V., Mithas, S., Keiningham, T. L., & Aksoy, L. (2011). An investigation of the cross-national determinants of customer satisfaction. *Journal of the Academy of Marketing Science*, 39(2), 198–215.
- Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4), 460–469.
- Oliver, R. L. (1981). Measurement and evaluation of satisfaction processes in retail settings. *Journal of Retailing*, 57, 25.
- Oliver, R. L. (1999). Whence consumer loyalty? *Journal of Marketing*, 63(4 suppl 1), 33–44.
- Palmatier, R. W., Dant, R. P., Grewal, D., & Evans, K. R. (2006). Factors influencing the effectiveness of relationship marketing: A meta-analysis. *Journal of Marketing*, 70(4), 136–153.
- Peters, L. D., Pressey, A. D., & Greenberg, P. (2010). The impact of CRM 2.0 on customer insight. *Journal of Business and Industrial Marketing*, 25, 410.
- Peterson, R. A., & Wilson, W. R. (1992). Measuring customer satisfaction: Fact and artifact. *Journal of the Academy of Marketing Science*, 20(1), 61.
- Reeves, M., & Deimler, M. (2011). Adaptability: The new competitive advantage. *Harvard Business Review*, 89(4), 134–141.
- Reichheld, F. F. (1996). Learning from customer defections. *Harvard Business Review*, 74, 56–69.
- Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, 81(12), 46–55.
- Richins, M. L. (1983). Negative word-of-mouth by dissatisfied consumers: A pilot study. *Journal of Marketing*, 47(1), 68–78.
- Roehm, M., Pullins, E., & Jr, H. R. (2002). Designing loyalty-building programs for packaged goods brands. *Journal of Marketing Research*, 39(2), 202–213.
- Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. *Journal of Retailing*, 69(2), 193–215.
- Sarstedt, M., Mooi, E., (2019). A concise guide to market research: The process, data, and methods using IBM SPSS Statistics (3rd edition). Springer
- Schwartz, E. M., Bradlow, E. T., & Fader, R. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4), 500–522.
- Shankar, V., Smith, A. K., & Rangaswamy, A. (2003). Customer satisfaction and loyalty in online and offline environments. *International Journal of Research in Marketing*, 20(2), 153–175.
- Sheppard, B. H., Hartwick, J., & Warshaw, P. R. (1988). The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of Consumer Research*, 15(3), 325–343.
- Sheth, J. N. (1970). Measurement of multidimensional brand loyalty of a consumer. *Journal of Marketing Research*, 7(3), 348–354.
- Shugan, S. (2005). Brand loyalty programs: Are they shams? *Marketing Science*, 24(2), 185–193.
- Singh, J., & Pandya, S. (1991). Exploring the effects of consumers' dissatisfaction level on complaint behaviours. *European Journal of Marketing*, 25(9), 7–21.

- Smith, A. K., & Bolton, R. N. (1998). An experimental investigation of customer reactions to service failure and recovery encounters: Paradox or peril? *Journal of Service Research*, 1(1), 65–81.
- Szymanski, D. M., & Henard, D. H. (2001). Customer satisfaction: A meta-analysis of the empirical evidence. *Journal of the Academy of Marketing Science*, 29(1), 16–35.
- Umashankar, N., Bhagwat, Y., & Kumar, V. (2017). Do loyal customers really pay more for services? *Journal of the Academy of Marketing Science*, 45(6), 807–826.
- Van Doorn, J., & Verhoef, P. C. (2008). Critical incidents and the impact of satisfaction on customer share. *Journal of Marketing*, 72(4), 123–142.
- Verhoef, P. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing*, 67(4), 30–45.
- Watson, G. F., Beck, J. T., Henderson, C. M., & Palmatier, R. W. (2015). Building, measuring, and profiting from customer loyalty. *Journal of the Academy of Marketing Science*, 43(6), 790–825.
- Weinberg, B. D., Milne, G. R., Andonova, Y. G., & Hajjat, F. M. (2015). Internet of things: Convenience vs. privacy and secrecy. *Business Horizons*, 58(6), 615–624.
- Wieseke, J., Alavi, S., & Habel, J. (2014). Willing to pay more, eager to pay less: The role of customer loyalty in price negotiations. *Journal of Marketing*, 78(6), 17–37.
- Woodruff, R. B., Cadotte, E. R., & Jenkins, R. L. (1983). Modeling consumer satisfaction processes using experience-based norms. *Journal of Marketing Research*, 20(3), 296–304.
- Zairi, M. (2000). Managing customer satisfaction: A best practice perspective. *The TQM Magazine*, 12(6), 389–394.



Market Segmentation

Tobias Schlager and Markus Christen

Contents

Introduction to the Concept	940
Market Segmentation: Key Considerations	942
Heterogeneity and Homogeneity	942
Segment-of-One	942
Concluding Thoughts	943
Market Segmentation: Process	943
Step 1: Characterizing the Ideal Market Segment	944
Step 2: Determining the Segmentation Criteria	945
Step 3: Collecting and Evaluating Data	948
Step 4: Forming Segments	951
Step 5: Evaluating the Final Segment Solution	957
Step 6: Implementing the Market Segmentation	960
Conclusions and Managerial Implications	962
Cross-References	964
References	965

Abstract

Market segmentation describes the practice of grouping consumers that are alike concerning specific characteristics. The idea is that firms can better identify and target attractive segments and customize marketing actions for each segment. Equally important, segmentation allows firms to avoid consumers that are unprofitable or otherwise incompatible with its marketing strategy. Like other marketing concepts, market segmentation has changed over the years with increasing globalization and digitalization. But the concept of market segmentation has been and will continue to be one of the key concepts in marketing practice. In this chapter, we define market segmentation along with its key characteristics, describe the process by which it unfolds, outline the main traps to avoid, and

T. Schlager (✉) · M. Christen

Faculty of Business and Economics (HEC) University of Lausanne, Lausanne, Switzerland

e-mail: Tobias.Schlager@unil.ch; Markus.Christen@unil.ch

provide an outlook into the future. A key concern of this chapter is also to reflect the key challenges in business environments, such as the abundance of data, globalization, as well as the acceleration of different trends.

Keywords

Heterogeneity · Market segmentation · Segmentation basis · Segmentation process · Segment-of-one · Social media

Introduction to the Concept

A key goal of firms is to allocate resources to their best opportunities and efficiently reach their organizational objectives. This requires firms to identify *groups of customers* who would most likely respond favorably to their offers and marketing actions. Market segmentation serves this purpose and is one of the most fundamental concepts in marketing management.

Market segmentation goes back to Smith (1956) and is defined as the identification of groups of customers with similar characteristics or needs who therefore are likely to exhibit similar behavior and reactions (i.e., are homogeneous) and that are distinct from other groups of consumers (i.e., are heterogeneous) in ways that are relevant for the firm. These groups should be mutually exclusive and collectively exhaustive. In other words, every customer should be allocated to exactly one segment. This definition has remained largely unchanged. For example, to Dolnicar et al. (2018), segmentation is “*the process of grouping consumers into naturally existing or artificially created segments of consumers who share similar product preferences or characteristics*” (Dolnicar et al. 2018, p. 11).

Conceptually, market segmentation is a compromise between, on the one hand, considering all customers as unique entities, with their idiosyncratic needs and preferences, enabling a firm to fully *customize* its marketing actions, and, on the other hand, considering the entire population of customers as similar enabling the firm to address them with a set of *standardized* marketing actions. By identifying subgroups or segments that are sufficiently homogeneous and different from other subgroups within a heterogeneous population, a firm can standardize its marketing actions for the subgroup only and still customize marketing actions across subgroups.

For instance, a car manufacturer could group its customers into five segments. In the best case, the car manufacturer should then be able to customize its marketing activities and products to these segments or a subset of them and still benefit from economies of scale. Thanks to market segmentation, General Motors was known to offer a “car for every purse and purpose,” which contrasted with Ford’s one-size-fits-all Model T, which was famously available only in black.

The key goals of market segmentation are therefore the following (Kotler 1989; Mahajan and Jain 1978):

1. Understanding the range of customer differences
2. Simplifying a market through grouping customers

3. Selecting target segments
4. Developing segment-tailored marketing actions
5. Efficiently allocating the firm’s resources towards their target segments and marketing actions

Market segmentation is not only a theoretical concept but also one with high managerial relevance as it serves to develop *marketing strategy and actions* (Kotler 1997; Wedel and Kamakura 2012). The concept has become one of the key pillars of any given marketing strategy (Dolnicar et al. 2018) and significantly contributes to the success of marketing within a firm. Unsurprisingly, market segmentation is one of the tools that entails the largest influence on marketing decisions (Roberts et al. 2019). Among a large set of marketing tools, the respondents rated market segmentation as the most impactful tool or concept. From a strategic perspective, market segmentation allows a firm to capitalize on a superior market position as well as to identify niche segments (Beane and Ennis 1987; Weinstein 1987, 2004).

The logical extension of market segmentation is the segmentation, targeting, positioning (STP) framework (DeSarbo et al. 2008; Lilien and Rangaswamy 2004). The real business value of market segmentation follows from the targeting and positioning decisions. Positioning comprises the development and implementation of marketing actions to communicate a firm’s image relative to the competition (Ries and Trout 1980). The STP framework and this chapter’s focus are illustrated in Fig. 1.

Market segmentation has come a long way. In a seminal article, Daniel Yankelovich (1964) urged marketing managers to abandon simplistic segmentations based on demographic information and introduced psychographic and values-based segmentation. Despite the importance professed by managers, he argued, 40 years on, that the practice of market segmentation had significant room for improvement (Yankelovich and Meer 2006). Our discussions with leaders in strategy consulting confirm this conclusion. Moreover, novel sources of data nowadays allow creating more elaborated market segmentations than ever.

The rest of this chapter is organized into three parts. We first discuss the key considerations of any market segmentation and the question of whether market segmentation is still relevant in the light of developments such as product customization and personalized communication. We then provide a detailed description of the segmentation process. Finally, we conclude by discussing the effects of newer

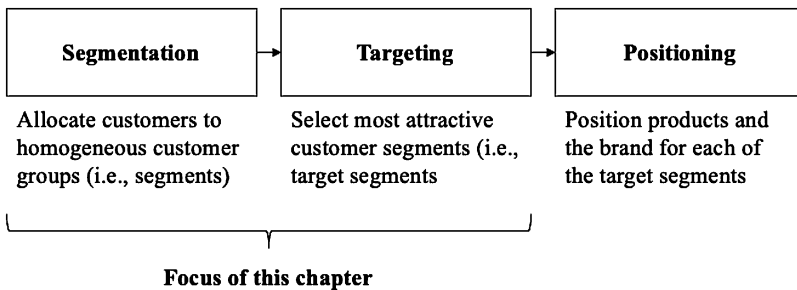


Fig. 1 The segmentation-targeting-positioning (STP) framework and the focus of this chapter

developments on this fundamental marketing concept, managerial implications, and highlight the main traps of market segmentation.

Market Segmentation: Key Considerations

Heterogeneity and Homogeneity

Market segmentation aims at allocating customers into groups based on how similar they are to each other. In technical terms, “similar” refers to homogeneous and “dissimilar” to heterogeneous. Thus, any segmentation, while motivated by the presence of customer differences, has the objective of reducing heterogeneity in the target market by identifying similarities, eventually enabling establishing a marketing strategy.

It would be tempting to start with the premise that customers are different. However, how different consumers are and whether differences are relevant for a firm is a matter of perspective. For example, humans share 99.9% of DNA, i.e., are homogeneous. For many medical problems, therefore, a standardized treatment suffices. For some diseases, however, the remaining 0.1% matter greatly. Likewise, many consumers have similar basic needs, seek similar benefits, but they can differ in terms of their specific preferences for quality, their decision-making process. We all must eat to satisfy our need for calories, but we can still have different preferences concerning the type of food, the way of preparation, and the time, place and form of consumption, creating many opportunities for segmentation.

Moreover, customers’ preferences are not fixed and so is the heterogeneity. Firms can influence customer preferences with their marketing strategies. Whether a market consists of substantial heterogeneity and whether recognizing this heterogeneity by a firm is useful depends therefore on its business strategy and competitive situation and its ability to turn it into an advantage.

Segment-of-One

Instead of segmenting customers, a firm could also consider each customer to form her or his own segment. This perspective is often referred to as “segment-of-one,” or that each customer forms her or his own segment (Bailey et al. 2010; Peppers and Rogers 1997; Winger and Edelman 1989) challenges the concept of market segmentation, which aims at simplifying the market.

The digital revolution is a key driver of the discussion around the relevance of segmentation because it has a huge impact on the way firms can identify, target, and engage with customers. The data generated by the digital revolution allow for insights into individuals at an increasingly granular level.

To better understand the impact of the digital revolution on the value of market segmentation, it is useful to distinguish between strategic segmentation and *operational* segmentation, i.e., a difference between defining target customers, customizing some marketing actions to individual customers. In particular, product

customization (Gilmore and Pine 1997; Kotler 1989; von Hippel 1998, 2001), as well as personalized communication (Ansari and Mela 2003; Arora et al. 2008; Postma and Brokke 2002) allow shifting key components of a firm's marketing program from an aggregated segment level to an individual level. Firms nowadays offer products that can be individualized or customized in response to consumers' idiosyncratic needs, which has been reflected under the term "mass customization" (Dellaert and Dabholkar 2009). Mass customization is a viable strategy, but successfully implementing mass customization requires organizational flexibility, e.g., in the production process (Piller 2004). Related to market segmentation, mass customization can provide the basis for reflecting the heterogeneity between customers by addressing their product preferences on an individual basis. The previously mentioned activities come together under what can be described as one-to-one marketing, which is defined as tailoring one or more dimensions of a firm's marketing mix to individual customers (Arora et al. 2008).

These ideas, however, do not challenge the value of "traditional" segmentation as much as they highlight the importance of distinguishing between strategic segmentation (for hard-to-change and hard-to-customize marketing actions) and operational segmentation (for easy-to-change and easy-to-customize marketing actions). The former makes marketing actions more efficient, while the latter enables customization and makes marketing actions more effective. The digital revolution has certainly shifted a number of marketing actions in many industries from the former to the latter and the combination of the two – a strategic segmentation and an operational segmentation – is key to success (e.g., Bailey et al. 2010). Table 1 outlines the differences between different levels of segmentation:

Concluding Thoughts

The value of any market segmentation ultimately depends on the heterogeneity in the market *and* the firm's marketing strategy. While some firms might even be able to customize their products and personalize their communication efforts, for many firms, "segment-of-one" strategies may be beyond their capacities. Those primarily refer to operational segmentations; however, any firm can benefit from a strategic segmentation.

Market Segmentation: Process

Different suggestions regarding the process of market segmentation exist (e.g., Dolnicar et al. 2018; Wedel and Kamakura 2012). We propose six key phases in market segmentation, ranging from *characterizing the ideal segment* to *implementing the market segmentation* (see Fig. 2).

We will illustrate the process in the domain of the automotive industry and therefore point out the critical factors in each of the phases. For this illustration purpose, we use survey data on 250 consumers that include variables related to:

Table 1 Unsegmented marketing, differentiated marketing, and the segment-of-one

	Unsegmented marketing	Differentiated marketing	Segment-of-one marketing
Unit of analysis	All customers	Groups of customers (segments)	Individual customers, consumption occasions
Target customers	Everyone	Single or multiple segments	Individual customers
Market characteristics	Little heterogeneity	Large heterogeneity	Each customer is unique
Information needs	Low, occasional	High, periodical	Very high, real-time
Granularity of data	Aggregated	Differentiated	Individual
Product	Mass production, physical products	Variety with mass production, physical products, and services	Services and digital products, mass customization
Marketing mix	Same marketing mix for all customers	Several alternative marketing mixes	Each customer receives a different marketing mix
Firm objective	Competitive advantage from low costs	Competitive advantage from differentiation	Competitive advantage from personalization and mass customization
Major disadvantage	Competitor may identify and create segments	Higher complexity and cannibalization	High cost of variety and complexity, no economies
Examples	Industrial commodities (salt)	Most consumer products (detergents)	Most capital goods, luxury goods (yachts)

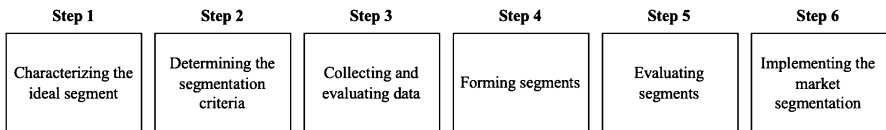


Fig. 2 The process of market segmentation

- Demographics (e.g., gender, age, income)
- Attitudes (e.g., attitudes towards cars as well as the key factors by which they choose a car; e.g., “Image is not important to me in a car”)
- Preferences (e.g., concerning the focal product)

Step 1: Characterizing the Ideal Market Segment

The first step in the process of market segmentation is to anticipate and describe the ideal segment as well as the ideal segment solution (Dolnicar et al. 2018).

As market segmentation is characterized by its exploratory nature (i.e., firms normally do not know the final market segments before conducting the

segmentation), this step might seem counterintuitive. Omitting this step might lead to firms ending up with a solution that does not fit firm strategy. For instance, while one firm may focus on smaller segments with fewer customers (i.e., niche segments), other firms – the market leaders – tend to only focus on segments that include a greater number of consumers (i.e., mass segments). The decision about which segment to focus on likely depends on multiple dimensions, such as the segments’ profitability, the number of consumers firms can cater to, their marketing activities, as well as their value proposition (e.g., for some firms, consumers’ response to price elasticity may be more important than for others).

Accordingly, the description of the ideal segment should be done before the actual segmentation and the ideal segment should be characterized by the same criteria that will be used to evaluate it later in the process (see *Step 5: Evaluating Segments*).

In the example of the choice of the car model “Ford Ka,” the ideal segment should allow Ford to identify the potential buyers of that new model, which is characterized by a unique design and sporty driving, as opposed to the traditional small car buyer who was a single, first-time or income-constrained buyer. Table 2 illustrates the means and medians of the variables that describe consumers depending on whether they prefer the Ford Ka (listing the model among the top three car choices), no preference for the Ford Ka (listing the model among the bottom three car choices), or middle (none of the previous categories).

Step 2: Determining the Segmentation Criteria

After determining the ideal segment, firms need to select the variables allowing them to identify that segment by distinguishing between consumers. These variables are called “segmentation criteria.”

While the literature has proposed many variables that can serve as a basis for segmentation (e.g., Bock and Uncles 2002), this chapter characterizes them using two dimensions (see Wedel and Kamakura 2012):

- *Observability*: The extent to which a variable can be observed without interaction with consumers
- *Consumption-specificity*: The extent to which a variable is related to the specific consumption decision

Table 2 Car preferences and means across variables

<i>Preference</i>	<i>Size</i>	<i>Gender (% female)</i>	<i>Married (yes)</i>	<i>Single (yes)</i>	<i>Age (average)</i>	<i>Income category (median)</i>
<i>Ford Ka top 3</i>	45.6%	53%	56.1%	31.6%	36.8	4.0
<i>Middle</i>	29.6%	51%	48.6%	43.2%	38.1	3.5
<i>Ford Ka least 3</i>	24.8%	35%	43.5%	43.5%	33.4	3.5

Observability. One key distinction between segmentation criteria is whether the segmentation criteria can be easily observed or not (Dolnicar et al. 2018; Wedel and Kamakura 2012). Observability describes a firm's ability to assess a dimension for individual customers without communicating with their consumers. Highly observable variables are visible. The best observable variables are typically related to the "who." Examples of observable segmentation variables include many demographic variables like gender, age, location of the customer, household size, and often culture (Wedel and Kamakura 2012). The combination of these measures can also result in socioeconomic status classifications (i.e., upper-middle class, lower-middle class, etc.). Another advantage is that these variables are stable (Wedel and Kamakura 2012); they do not change much or at all over time. In business markets, an easily observable variable is the industry or the size of the customer. Such observable variables also allow firms to identify any new customers as members of a specific segment. For instance, using age as a segmentation variable allows allocating new customers to specific segments even when they had not been subject to the initiation market segmentation. Accordingly, observable segmentation criteria have various advantages.

Nevertheless, observable variables have disadvantages. One is that segments that are formed based on observable segmentation criteria typically differ less in their responsiveness to firm activities (Frank et al. 1972; McCann 1974). Marketing actions are more effective when they aim at such underlying motivational drivers, the goals customers pursue (Kruglanski and Szumowska 2020). Although this renders observable variables less valuable, firms nowadays still widely apply observable variables as segmentation criteria and their popularity as segmentation criteria is unbroken (Wedel and Kamakura 2012). In sum, unobservable variables related to needs, goals, and benefits are typically more meaningful for the segmentation as they are more closely related to preferences (e.g., attitude towards specific products) or future behavior (e.g., purchase behavior), but it is oftentimes easier for firms to use observable segmentation criteria.

The question "*who does what, when, where, how, and why?*" provides another classification of segmentation variables into *needs* or *benefits sought* ("why") and the *decision-making process* ("how") of the consumer, *demographic* variables ("who"), and *behavior-specific* information ("what," "when," and "where"). The "why" and the "how" are more related to motivational drivers of behavior but typically unobservable. On the other hand, the "who" and "what" are much easier to observe but are often not closely related to motivational drivers.

Assume using age as segmentation criterion when considering our example of the Ford Ka: Using age, which is observable, as segmentation criterion would unlikely lead to a meaningful segmentation given that reportedly the consumers interested in that model ranged across all age categories. By contrast, consumers' preference for driving pleasure and their ambition to "be different" from others, which both are unobservable to Ford, would be better suited in identifying the segments that are eventually meaningful to Ford.

Accordingly, firms need to assess whether observable variables are sufficient for their market segmentation, or whether they need deeper insights into their customers and therefore need unobservable variables.

Consumption-specificity. The second dimension by which segmentation criteria can be classified is the extent to which a criterion is specific to the consumers' buying and consumption behavior. Examples of criteria with a high consumption specificity are product usage frequency, store loyalty, adoption stage, usage situation, and brand loyalty (Wedel and Kamakura 2012). While some of the variables, and in particular the behavioral variables related to the "what" (e.g., store patronage or online presence visits), are more readily available as they can be observed (e.g., using a CRM system or weblog systems that track consumers' online behavior), others might be more difficult to collect (Payne and Frow 2005; Verhoef et al. 2010). For instance, attitudinal variables, such as ambition to "be different" from other consumers, need to be revealed by consumers in surveys (Wedel and Kamakura 2012). Even more so, consumption-specific variables are often unobservable, especially those related to the "how." Among those unobservable characteristics are variables such as the decision-making process (Blattberg and Sen 1974; Dhalla and Mahatoo 1976), different types of elasticities (and in particular price elasticity; Wedel and Kamakura 2012), and customer perceptions of brand attitudes (Yankelovich 1964).

Using behavioral variables related to the "what" can seem problematic because the exact purpose of segmentation is to identify segments whose behavior can potentially be shaped through marketing actions. While segmentation based on consumption or usage volume, brand loyalty, or customer profitability, a typical segmentation in services or B2B markets to allocate sales effort is relatively easy to implement, and it requires an existing market, available transaction data, stable, hard to change customer behavior, and good coverage of all potential customers. If these conditions do not hold, it can easily result in self-fulfilling "prophecies" (e.g., low users remain low users or brand switchers remain brand switchers), because there is no effort to change behavior, in overlooking a significant part of the market because customers do not appear in existing databases, or in getting blind-sided by competitors, when their actions can change customer behavior. In the end, there is still a need to understand the drivers of the observed behavior to understand customers and guard against these threats.

The variety of the to-date examined measures is so broad that "almost every consumer behavior variable has been proposed for segmenting markets" (Bock and Uncles 2002). The advantage of variables with a high consumption specificity is that they are more predictive of consumer's attitudes and behavior towards a firm's products. However, one downside of high consumption-specificity is that its reliability is questionable (Lastovicka et al. 1990) and may even be negatively related to the stability (Wilkie and Cohen 1977). In particular, consumption-specific variables are often subject to more significant changes given that they depend on characteristics as the rate of innovation or the change of a competitors' marketing communications. By contrast, general, or consumption unspecific, characteristics are typically more stable than consumption-specific characteristics. Among general variables count, for instance, customers' general values, cultural values (Steenkamp and Ter Hofstede 2002), as well as lifestyle (Wedel and Kamakura 2012). Thus, firms need to trade-off whether they use variables that are high in terms of consumption specificity (and thus be more informative for a firm) or that might be more stable (and thus longer lasting).

Table 3 Classification of segmentation variables (see Wedel and Kamakura 2012)

	Consumption specific	Consumption unspecific
<i>Observable</i>	Purchase frequency Store loyalty Online behavior Time spent on the website Purchase amount	Demographics (age, gender) Culture Public social media profile information Industry application
<i>Unobservable</i>	Attitudes toward product/brand Benefits Needs Price sensitivity Decision-making unit and process	Values Lifestyle Opinions Strategic objectives

Building on these dimensions, one can establish a matrix that classifies the different segmentation criteria and thus allows firms to select any of those (see Table 3).

A pitfall of market segmentation is to either *only* use observable (e.g., demographic) or unobservable variables as segmentation variables. A good market segmentation often requires a *combination* of these variables, which allow a firm to form *and* identify actionable, responsive segments.

In the Ford Ka example, the following list is available as segmentation criteria:

- *Demographic variables*: Age, marital status, children, first car, age category, children category, income category, residence, single, marital status, household size
- *Attitudinal variables related to cars and driving* (examples):
 - When it comes to cars, my heart rules my head.
 - I want a car that is nippy and zippy.
 - I prefer cars with high performance.
 - For me, a car is a symbol of freedom and independence.

While the demographic criteria can be (typically) easily observed, the attitude variables require a survey to be evaluated. For the segmentation analysis, we selected three factors that (Factor 1: Value of *small* cars as means of transportation, Factor 2: Seeing a car as a statement, Factor 3: Seeing driving as more than transportation) that summarized the attitudinal variables (an inventory of 100 different questions about transportation, cars, and driving). This approach should allow us to identify segments that have distinct attitudes toward cars and driving.

Step 3: Collecting and Evaluating Data

In this step, firms need to assess the data that are available, as well as identify the need for data that may complement these yet available data (Dolnicar et al. 2018). Firms then collect these data. This step of the segmentation process has significantly

changed in the past decades due to new data sources and new data collection methods.

Classical sources of market research remain the most important source of data. To the pillars of classical market research count among others:

- Survey-based methods (e.g., paper-back surveys, web-based surveys)
- Conjoint analysis (e.g., choice-based conjoint analysis, adapt conjoint analysis)
- Other methods (e.g., data from CRM databases, consumers' expressions on social media)

Surveys. Surveys are a method to collect self-reported data and are the most common source of market segmentation also, because they are easy and cheap to collect (Dolnicar et al. 2018). Current and prospective consumers respond to questions on their needs, attitudes (e.g., brand evaluation, customer satisfaction), attitudes and preferences (for product attributes, products, brands), and behaviors (e.g., information search behavior, purchases of the company and its competitors). Notably, those questions should allow for a precise assessment of the dimensions later used for the segmentation. The key advantage of surveys is that they allow to reveal unobservable variables and thus information about the consumers that cannot be easily collected differently. When conceptualizing consumer surveys for segmentation, it is key to allow the collection of unbiased information. This requires considerations about how to recruit respondents (e.g., to avoid biases related to self-selection and representativity), the number of respondents that are recruited (i.e., sample size), as well as how the survey is designed (e.g., to avoid response biases and the effort required of respondents to fill out the survey; Dolnicar et al. 2018).

Conjoint analysis. Conjoint analysis extends these surveys by making a systematic assessment of how individual attributes can affect consumers' choices; eventually, these methods allow to estimate consumers' willingness to pay and thus price sensitivity. Conjoint analysis follows the idea that consumers can assess and compare products as a whole better than individual attributes. Oftentimes, such conjoint analysis starts by requiring respondents to indicate options that they would not consider at all. This allows for more precisely assessing the remaining potentially attractive options. Respondents then evaluate those offers as a package against each other, iteratively comparing different offers. While many methods related to conjoint analysis have been developed, the traditional ones are choice-based conjoint analysis and adaptive conjoint analysis. While the latter only requires respondents to select between two offers at a time while accounting for their previous choices, the former provides multiple profiles of offers (e.g., consisting of a product, a price, a package). Choice-based conjoint analysis is thus typically more demanding. Given that price sensitivity is one of the key segmentation criteria and the importance of individual product attributes allow for actionable insights, conjoint analysis is particularly attractive for collecting data for segmentations.

Other methods. The increase in digitalization and the proliferation of customer relationship management (CRM) databases allow to enrich classical sources with additional segmentation-relevant data from manual or automated recording of *customer observations*. In CRM databases information on a company's current

customers is archived and readily accessible, and they typically consist of demographic data, interaction data (previous interactions, such as sales interactions), as well as transaction data (e.g., historic purchases, shopping baskets). Relatedly, web-log systems (e.g., Google Analytics) track consumers in real-time generate detailed behavioral data on individual consumers. Moreover, online social networks nowadays have been established as a source that allows to provide significant amounts of data on both individual consumers as well as their social context (e.g., Casteleyn et al. 2009; Russell 2011). What is more, data exchanges oftentimes allow firms to access information about consumers prior to their arrival on the website.

With an increasing amount of available data, the question for firms shifts from “How can firms access the data?” to “How attractive are the data?” (Cai and Zhu 2015), and ultimately “How important are the data for effectively differentiating between groups of consumers?” (e.g., Haustein 2016; Liu et al. 2016; Tufekci 2014). This section discusses novel data and how firms can use them for market segmentation (Öztaysi and Onar 2013; Netzer et al. 2012) and proposes that these novel data can allow firms to assess previously unobservable consumer characteristics.

A key source of data today are online social networks. Two data types can be usefulness for segmentation:

- Standardized expression of attitudes (e.g., Facebook/Twitter likes)
- Textual data (e.g., Facebook/Twitter comments)

The first, expression of attitudes describe the various ways of expressing oneself on online social networks that are standardized. Examples of those are, for instance, the act of one’s liking of a fellow user’s comment, of a topic, or even of a brand. While each online social network offers its own labeling of this expression (e.g., Facebook uses so-called “likes”), they often reveal preferences and are easy for firms to collect. The revealed preferences can allow inferring people’s personalities, which can be valuable as segmentation criteria.

A large amount of information is posted daily on social in the form of texts (Netzer et al. 2012), for instance, when consumers write status updates (i.e., information about their current activities), comment on others’ activities, or engage in conversations. Such data are unstructured, which leaves advantages and disadvantages. While the data potentially contain rich information, the complexity of extracting information may not be trivial and often requires skills and resources (Netzer et al. 2012). Among the common methods of extracting insights from textual data are sentiment analysis (which analyzes the emotional profile of a message; Dhaoui et al. 2017) or advanced natural language processing (Tsai and Chiu 2004), and for more detailed information, please see chapter “Automated Text Analysis” of this handbook. Moreover, to translate this textual data into segmentation criteria, firms also need to deal with multiple key challenges, such as (1) matching people’s profiles on online social networks to actual persons and (2) identifying whether consumers state their honest opinion. Despite the previous issues, marketers can take advantage of these new data and also use them as a basis for market segmentation – oftentimes to render observable previously unobservable variables.

In the Ford Ka example, data were collected via surveys. Consumers had to indicate their preferences concerning the car, against different attitudes, as well as whether they listed the Ford Ka among the three most preferred cars, three least preferred cars, or in the middle. Moreover, this survey allowed obtaining information about non-observable variables, such as the consumers' attitudes. However, one challenge related to this survey is the relatively small sample of consumers and the assessment of the representativity. This points to a general problem in data collection for segmentation purposes: creating a sample that is representative to estimate the size of segments, when the segmentation is not yet known. This problem can be addressed with additional market research *after* the segmentation is complete. Alternatively, a commonsense segmentation, which we elaborate on in the next section, can guide the determination of the sample and the sample size *before* the survey is conducted.

Step 4: Forming Segments

The formation of segments, i.e., the actual act of market segmentation, is fundamentally a quantitative task requiring specialized statistical methods. However, any firm can benefit from a market segmentation, even when that segmentation is informal (i.e., lacks sophisticated methods). Dolnicar et al. (2018) propose to distinguish between *commonsense* and *data-driven* market segmentation. While the commonsense segmentation allocates consumers to different groups iteratively using different segmentation variables, data-driven segmentation uses those criteria at the same time along with model-based estimations (Dolnicar et al. 2018). A commonsense segmentation might be as simple as using just paper and pencil by brainstorming about *why* a customer would buy a product or service, derive value from it, and *how* a customer is likely to acquire it.

For using more quantitative market segmentations, multiple methods have evolved over the past decades (Wedel and Kamakura 2000). Initially, researchers used classical multivariate statistical methods such as cluster analysis, discriminant analysis, and regression analysis. More recently, the emphasis has shifted to model-based segmentation methodologies involving more complex optimization and numerical methods, finite mixtures, and Bayesian approaches given the various criteria established for effective market segmentation.

The purpose of this section is to give the reader sufficient insights into different methods to understand their value and offer a guideline to identify the appropriate method for their market segmentation challenge. In particular, it discusses the assumptions of different methods about segments. For detailed technical descriptions of segmentation methods, we refer the reader to other chapters in this handbook (see chapter ► [“Cluster Analysis in Marketing Research”](#)) and to other sources that focus on the description of statistical methods and processes. Forming market segments that are useful for business decisions can be broken into three different but related steps:

- *Creation of market segments*: How many segments do properly represent the customer heterogeneity of a market? What is the unifying need or behavior within a segment and what are the key differences between the segments?

- *Profiling of market segments*: How can firms classify customers into existing segments? What are the characteristics of customers within a segment?
- *Sizing of market segments*: How many customers belong to a segment?

Creation of market segments. Market segmentation is required as a result of the heterogeneity of customers in a market (Kotler 1997). We, therefore, need first a description and quantification of this heterogeneity. The “classic” case of segmentation occurs when the company has no or only limited knowledge of customer heterogeneity, making segmentation an exploratory task. The main goal is to *determine the number of segments* based on the collected data and which segments each customer is assigned to. The most appropriate methods for this task are various existing clustering methods. Cluster analysis is often seen as synonymous with segmentation. It is, however, important to note that clustering methods are most useful for the creation of segments but do not suffice for market segmentation overall.

Different clustering methods make different assumptions about the nature of customer heterogeneity. A first assumption is whether or not a customer belongs to one and only one segment. This leads to the distinction between *nonoverlapping*, *overlapping*, and *fuzzy* clusters (Hruschka 1986). The first type follows the idea that segments ought to be mutually exclusive and collectively exhaustive (MECE) while the other two allow for customers to belong to multiple segments. Fuzzy clustering provides a probability vector for segment membership and can thus be seen as an intermediate solution between overlapping and nonoverlapping clusters.

If a customer can belong to multiple segments, then the customer can be exposed to different, potentially conflicting, marketing actions. On the other hand, it is well known that customers can belong to different segments, especially when we extend market segmentation beyond the grouping of customers to a classification of consumption situations or occasions (Arabie 1977). For example, the same customer can consume beer for different reasons and at different locations. So, the distinction between nonoverlapping and overlapping clustering is closely related to the question of what exactly is the “object” that should be grouped in a market segmentation. This is an important decision that relates to the distinction between strategic and operational segmentation. From a strategic perspective, it is more appropriate to assume a customer belongs to only one segment and therefore assume clusters are nonoverlapping. From an operational perspective, when customer activation requires a combination of marketing actions, forming overlapping clusters can be very useful.

Nonoverlapping clustering methods are the most commonly used methods in marketing for market segmentation. Because the key question when creating segments concerns the number of segments, the distinction between *hierarchical* clustering and *nonhierarchical* clustering methods is very important. Hierarchical methods do not require the specification of several segments. They start with each customer forming a single-subject cluster. These clusters are then linked in successive steps until all customers are in the same cluster. This forms a tree-like structure, hence the term hierarchical clustering. Nonhierarchical methods start from a random initial division of customers, which is then changed until an optimization criterion is

achieved given the a priori specified number of segments. Typical optimization criteria involve some kind of distance measure to account for the within and the between variances. Hierarchical methods can therefore be seen as even more exploratory than nonhierarchical methods. Their disadvantage is the lack of a conceptual basis to justify a hierarchical structure among customers to characterize heterogeneity. This structure also implies that a customer stays in the same cluster irrespective of how many clusters are formed as the number of clusters is changed through successive steps of combining clusters.

Determining the appropriate number of clusters is one of the most difficult problems in the creation of segments. The goal, especially for strategic market segmentation, is to obtain the smallest number of segments that makes sense for the firm. A useful approach for this problem is to use multiple methods. Hierarchical clustering gives an initial estimate of the number of clusters which then can be applied to nonhierarchical clustering to refine and verify the segmentation. In the third step, the number of clusters can be systematically increased and decreased. The cost of increasing the number of segments is increased complexity. For example, if an increase in the number of segments yields another segment that would not be targeted, there is little value in increasing the number of segments. Similarly, the size of the additional segment can be too small to be attractive or meaningful. On the other hand, a reduction of the number of segments can eliminate a potentially attractive target segment or an otherwise valuable market insight.

There are statistical methods to determine the optimal number of clusters (e.g., Calinski and Harabasz 1974), but it is important to understand that there is no theoretically correct market segmentation, and the final number of segments is a subjective decision that is based on statistical metrics and business considerations. The final criterion is the utility of a final segmentation for the business and its marketing actions – the business value. For more details about clustering methods, we refer the reader to chapter “Cluster Analysis in Market Research” in this handbook.

Profiling of market segments. This business value also depends on the ability to properly sort customers into the created segments, including customers who were not part of the data used to create the segments in the first place. In other words, market segmentation requires a description or profiling of segment members. For that, we need *observable* variables as outlined the section “Step 2: Determining the Segmentation Criteria.”

Clustering and classification methods can be confused because both methods allocate subjects into several groups or segments. In classification tasks, we know the number of groups and the membership of existing subjects to those groups. The objective of clustering methods is to reduce a larger number of items into a smaller number of homogeneous clusters based on collected data. The objective of classification is to assign an item to the appropriate group with the help of a classification model. Using the terminology of machine learning, classification is a typical task of directed knowledge discovery while clustering is an example of undirected knowledge discovery. In a classification task, we have a dependent variable – the existing segments – and independent or predictor variables and we have predictive methods.

The two most common methods are discriminant analysis and logistic regression. Both are used when the dependent variable is categorical. Discriminant analysis creates discriminant function(s) to maximize the difference between the groups on the function and is only used for categorization. Logistic regression works like ordinary least squares regression but on the logit of the dependent variable. It can be used for categorization but more also provides the odds ratio for each variable. Cluster analysis without a subsequent discriminant or regression analysis does not yield properly formed *and* profiled customer segments. An alternative method to clustering plus discriminant analysis is latent class analysis (LCA; see also chapter ► “[Finite Mixture Models](#)”), which allows for a segmented analysis of customer reactions to various marketing actions, especially price, and an integration of segment creation and profiling (Grover and Srinivasan 1987; Kamakura and Russell 1989).

Predictive methods are also useful in the context of CRM systems when segments can be formed based on past purchase behavior or customer value estimates, which then can be linked to available predictor variables. With today’s computing power, very sophisticated predictive models using Markov Chain Monte Carlo (MCMC) methods can be deployed to directly approximate the posterior distribution of a parameter of interest, thereby making the customer heterogeneity visible. Moreover, the possibility to link it to observable predictors greatly facilitates the description or profiling of segments.

Decision trees also provide a powerful and easy-to-implement classification method. The goal is to build a tree that will allow us to predict the dependent variable based on the values of attributes or independent variables. Decision trees differ from logistic regression in the way they generate the boundaries between to separate different classes. Regression “fits” a line to divide the space, whereas decision trees bisect the space into smaller and smaller regions in a nonlinear fashion. While decision trees, especially today’s high-power machine learning methods, have superior classification performance, logistic regression accounts for the simultaneous effects of all predictors and is usually less costly in terms of sample size. For a further discussion of the advantages and disadvantages of decision trees, see Berry and Linoff (1997) or Murthy (1998).

Sizing of market segments. Segmentation schemes offered by consulting companies, like the VALS™ segmentation, or simple observable variables like zip codes. The value of these segmentation methods is somewhat limited today. Conceptually, however, these methods play an important role in the segment formation process. First, the selection of *target* segments depends on the evaluation of the attractiveness of the segments. One important determinant of segment attractiveness is the size of the segments (see “[Step 5: Evaluating the Final Segment Solution](#)”). An accurate estimation of the size critically depends on the proper sampling of the underlying population, but without some a priori knowledge of the segment structure, it is all but impossible to ensure a representative sample for creating the segments. When Renault launched the Twingo and Ford the Ka, the market segmentation revealed the presence of design-sensitive buyers of small cars. But the size of these segments

could not be reliably estimated because the respondents were not recruited with this type of segmentation in mind.

Second, any overall average market research result calculated across a heterogeneous customer group is useless, if not dangerous. The calculation of the average customer satisfaction across all Starbucks customers depends among other things on customer acquisition. The rapid store expansion attracted a lot of convenience buyers who did not value Starbucks as much as the experience buyers, the brand lovers. It is possible that within both segments, customer satisfaction increases, but a faster increase in the number of generally less satisfied convenience buyers would result in a decrease in the average customer satisfaction. Similarly, if Coca-Cola had segmented people before doing taste tests to develop New Coke, for example, on simple variables like their brand preference and attitudes, marketing history might be one disaster shorter.

The third key point is that for *strategic* market segmentation, the key segment formation decision is setting the number of clusters. This is a managerial judgment task guided by statistics. For an *operational* market segmentation, especially in a setting where marketing decisions are automated, statistical methods, like machine learning methods, are essential. They are very powerful, but the daily online experience with retargeting illustrates the danger of very powerful methods: they can be precisely wrong.

It should also be clear that the formation of segments is not a simple sequential process. Segmentation is an iterative process. In particular, the number of segments and the choice of profiling variables also depends on the organizational applicability of the segments. The next step reviews how to evaluate the final segment solution. In sum, it is important to see different methods as complementary and not as substitutes. A good segmentation process requires *descriptive* and *predictive* methods, and without at least some a priori idea of segmentation (see “[Step 1: Characterizing the Ideal Market Segment](#)”), any single statistical method can yield incomplete results.

Returning to the Ford Ka example, we conducted a k-means cluster analysis (see chapter ► “[Cluster Analysis in Marketing Research](#)”) to reveal different segments. In a k-means cluster analysis, the so-called k-means algorithm iteratively partitions the data set into a predefined number of groups that are distinct and nonoverlapping. Each data point is then allocated to a specific group.

In a first step, we created the segments using the k-means algorithm and the R package. This step was done to determine the number of segments. We simulated different numbers of clusters and examined to what extent the heterogeneity within the segments was reduced by adding additional clusters with the elbow method (the number of segments was chosen at the kink in the line plot, i.e., at four clusters).

Next, we visualized the segments to provide a profile of those segments (Fig. 3a–c). Specifically, we plotted the segments using the three-, four-, and five-segment solution. As can be seen, the five-segment solution was overlapping for the first two factors, while the three- and four-segment solutions were discriminating well between the consumers on those segments.

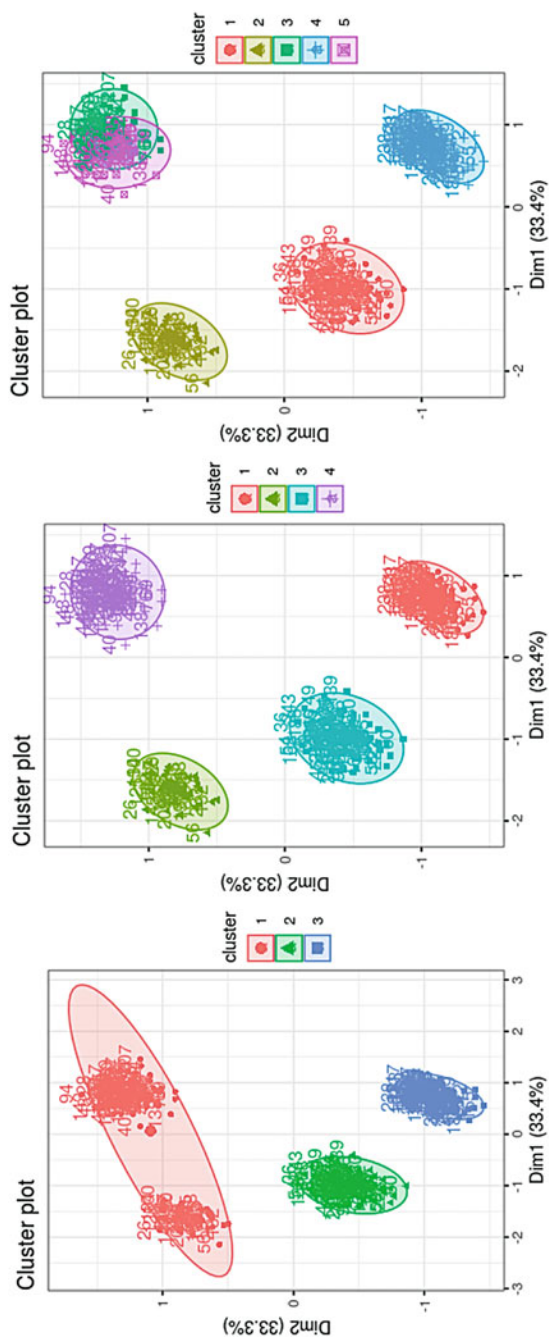


Fig. 3 (a–c) The segments illustrated

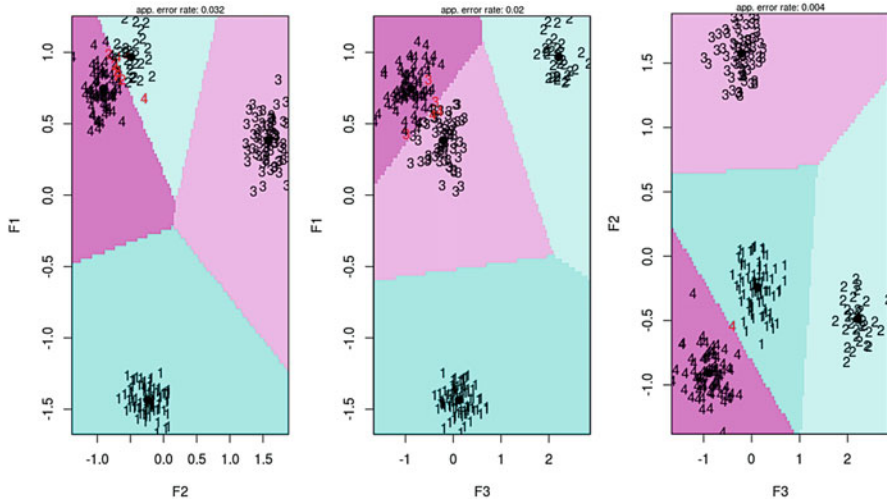


Fig. 4 a–c The discriminatory power of the factors for the four-segment solution

We next used a discriminant analysis to test whether each of the factors contributes to the separation between the factors. The factors discriminated well between segments as the following figure illustrates (Figs. 4a–c):

Step 5: Evaluating the Final Segment Solution

After forming segments, firms need to evaluate the viability of the segmentation. For doing so, several commonly accepted criteria have been established (see Table 4; Pires et al. 2011).

Identifiability. A key criterion for evaluating a market segmentation is the identifiability of the segments (Blattberg and Sen 1974; Dolnicar et al. 2018; Wedel and Kamakura 2012). Identifiability of the segments describes a marketer’s ability to identify the segments. Typically, firms can use observable criteria to identify those segments. For instance, when segments differ in terms of age or gender, it would likely be easy to identify the market segments. By contrast, the identifiability is low when nearly no, or no observable criteria are considered in the segmentation. If only unobservable variables such as attitudes are used to distinguish segments, firms are challenged in distinguishing between segments. In practice, firms often either use a combination of observable and non-observable variables (to be able to identify segments without trading off forming meaningful segments) or even entirely use observable variables to ensure the resulting segments are identifiable.

Accessibility. A second criterion is whether a consumer segment can be effectively reached by a firm or its “accessibility” (Kotler 1997; Pires et al. 2011). While the proliferation of new media increases the accessibility of segments and thus this

Table 4 Criteria to evaluate final segmentation

Criterion	Explanation	Ford Ka
<i>Identifiability</i>	Firm's ability to identify the segments using a specific set of observable criteria	Difficult
<i>Accessibility</i>	Extent to which a segment can be effectively reached by a firm	Good
<i>Responsiveness</i>	Extent to which a segment responds to a particular marketing program or product position	Good
<i>Stability</i>	Extent to which a segment remains similar as well as meaningful over time over a specific period of time	Probably good
<i>Substantiality</i>	Extent to which segments will yield enough profits such that they justify being targeted	Unclear

criterion seems to be of less practical relevance, it is by no means given that all segments are equally easy to be reached. Nowadays, consumers can effectively limit the extent to which they are reached by specific communication efforts and thus might be less easily accessible by companies. One way of how access to consumer segments becomes more difficult is the use of ad blockers. Ad blockers restrict the possibility of firms to effectively reach consumers via online advertising (Goldfarb and Tucker 2011). More importantly, some segments are more prone to deter display advertising that they are approached with by a company, i.e., the technology-savvy consumers who are aware of the latest ad blockers. As a result, firms may not be able to reach those consumer segments and thus a market segmentation that results in those segments may create some hardly accessible consumer segments. While this is only one example of how accessibility to segments can vary across segments, multiple others exist.

Responsiveness. "Responsiveness" implies that market segments respond to a firm's marketing activities, such as response to a particular marketing communication campaign, to variations in price, or even the product's features (Myers 1996). Moreover, the segments' responses should also differ from each other, which is sometimes considered to be an extra criterion called "differentiability." Based on the segments' responses, firms can more efficiently specify their marketing activities. For instance, assume two segments differ in terms of their media habits. While the first segment predominantly uses traditional media such as TV and radio, the second segment uses new media and focuses on receiving all information via social media. The difference between the two segments' media usage allows firms to align their media mix in a way that they can reach each of the segments most efficiently. The firm can target the first segment with classical TV campaigns, while it can target the second segment by specifying characteristics for their advertisements on social media. By contrast, if the segments do not differ concerning how they respond to various marketing activities, the market segmentation has failed to attain its goal of more efficiently allocating its resources.

Stability. A criterion of market segmentation is the extent to which the segments are stable (Dolnicar et al. 2018; Hassan et al. 2003; Pires et al. 2011). Put differently, this criterion examines whether the market segmentation as well as the segments it formed remain meaningful over time. Segmentation stability is insofar important for

firms as it allows to craft longer-term strategies based on those segments. Dolnicar and Leisch (2010) propose repeating the segmentation when a firm is concerned about changes to assess its stability. Assessing the stability includes evaluating whether the *same* segments will form in a similar structure and size over time. While overall segments might reproduce in different segments over time, individual segments might form less well over time. This segment level stability refers to “how often a market segment with the same characteristics is identified across different repeated calculations of segmentation solutions with the same number of segments” (Dolnicar et al. 2018; p. 167). Notably, the stability of segmentation solutions depends on the criteria chosen for segmentation. A market segmentation that relies on consumers’ values, which are characterized to be stable over time, eventually results in a more stable segmentation than one that builds on criteria that vary significantly, e.g., consumers’ sentiment expressions on social media.

Substantiality of segmentation. The final and key criteria for evaluating market segments is the “substantiality” of the formed segments. Substantiality describes the extent to which a segment is attractive enough such that it justifies being targeted – which is typically derived by a segment’s profitability potential. Substantiality thus depends on a combination of measures, such as the number of consumers assigned to a segment, economic variables as their customer lifetime value, their purchasing power, or their price sensitivity (Hassan et al. 2003). While the number of consumers in a segment is not the only measure, it still remains important as it relates to the firm’s strategy: Large segments are typically targeted by market leaders (and thus characterized by more competition), whereas smaller segments form so-called niche segments, which may be more attractive for smaller firms. It is worth noting, however, that recent developments have challenged this idea. Production techniques, as mass customization and personalized communication and dynamic pricing nowadays allow firms to cater to needs of individual consumers. Accordingly, on an operational level, firms are nowadays able to target much smaller segments than previously and to individualize their activities.

We next illustrate the segments using the demographic variables reported at the outset and added the preference for the Ford Ka. In terms of the established criteria for segmentation, it will be difficult to easily identify the segments given that they barely differ in terms of observable attributes such as gender or age. The accessibility to each of the segments is given as follow-up analyses concluded: They can be reached using classical communication instruments. As can be seen by the classification (see Table 5), the segments differ in their preferences for the Ford Ka, thus,

Table 5 Segment descriptions

<i>Segment</i>	<i>Size</i>	<i>Pref Ka</i>	<i>Gender (% female)</i>	<i>Married</i>	<i>Single</i>	<i>Age</i>	<i>Income category</i>
<i>Seg. 1</i>	31%	43.6%	41%	53.8%	35.9%	37.0	3.77
<i>Seg. 2</i>	13%	56.3%	50%	53.1%	37.5%	35.1	3.59
<i>Seg. 3</i>	26%	41.5%	62%	50.8%	40%	37.1	3.28
<i>Seg. 4</i>	30%	46.7%	43%	46.7%	38.7%	35.6	3.97

they seem to be differently responsive, with Segment 2 being the one that has the strongest preference for the Ford Ka. While we did not illustrate the stability over time (as this was a cross-sectional survey), one can assess the substantiality of the segments: The smallest segment of the four-segment solution seems to be most likely to consider the Ford Ka among their top three choices. The segment is bolded. This last problem could potentially be alleviated by extending the targeting to include, for example, all or part of Segment 4.

Step 6: Implementing the Market Segmentation

The final step is the implementation of a market segmentation. In this phase, firms need to consider their resources as well as their organization. The key objective of this section is to emphasize that a market segmentation is a means to an end, which is to more effectively serve segments of consumers). Three points require attention:

- Allocating sufficient resources to the implementation (in the short- *and* long-run)
- Closely aligning marketing and sales departments
- Regularly evaluating and updating the market segmentation

To materialize the value of any market segmentation demands to fully embrace it. This may as well imply substantial organizational changes (Croft 1994) and if firms are not willing to bear these consequences and thoroughly implement the market segmentation, even a potentially successful market segmentation may fail (Dibb and Simkin 1997, 2001). While firms may reorganize and align their marketing activities in the short run, they should also consider the long run. For example, in our Ford example, Segment 4 not only seems to be the biggest segment but also the one with the greatest income. However, this potentially profitable consumer segment does not prefer the Ford Ka as much and thus does not match the product well. Investing resources to develop a product that meets that segment's needs might not be worth the effort.

One of the challenges when implementing a market segmentation is the alignment of the marketing department and the sales department. Typically, creating a market segmentation is the task of the former, while the implementation primarily concerns the latter. This requires a close alignment between both departments, in terms of its organizational routines or IT systems. The knowledge of the market structure that the marketing department reveals by means of the market segmentation need to be internalized into a firm's CRM system to grant the sales department easy access. What is more, the operationalization of the market segmentation can be updated in real-time in case a firm's IT systems allow for doing so. For instance, depending on the information that firms have about visitors of their websites (e.g., obtained by Google Analytics), firms could display personalized information. The information though also flows in the opposite direction. A successful segmentation depends on the availability of data within the CRM system, which are primarily generated or provided by sales or customer service.

Besides these formal implementation challenges, the results of a market segmentation should also be actively communicated to a firm's employees (e.g., product developers or customer contact personnel) in a way that it is useful. For illustration purpose, firms can use of so-called "personas" that are actual or fictional profiles (Pires et al. 2011) that portrait the members of each market segment in a representative manner. These personas are typically focused on the attitudes and motivations of that segment and given that they illustrate segments they translate the mere numbers and descriptive patterns of a market segment into useful visualizations. While using actual consumers as personas can be more believable (Judge et al. 2012), fictional ones may fit more precisely and may be dynamically changed to reflect changes in of a given market segment (An et al. 2017). Such illustrations may support the firm in various considerations. For instance, product developers might consider such personas, and their projected motivations and goals, as basis for developing products that appeal to multiple segments (Pires and Stanton 2005). Moreover, personas and their visual characterization may support customer contact personnel in identifying members of individual market segments (Pires et al. 2011) and thus be decisive about whether the market segmentation will be used in practice. While no personas are indicated, we labeled the segments according to their psychographic profiles (see Table 6).

Many firms see market segmentation as something you do "once in a lifetime." Market segments are based on consumer characteristics and these characteristics change over time. As outlined in *Step 5: Evaluating the Final Segment Solution*, a market segmentation should be stable over an extended period of time as it creates the basis for long-term strategic decisions that are hard to adapt or even reverse. At the same time, firms should not ignore the implications of consumer trends for market segmentation. A market segmentation needs to be regularly evaluated against these trends. When new consumers enter a market, the variables used to identify segments may need to be adapted. When new consumer needs emerge, the variables used to create the segments need to be adapted. When both change at the same time, it is time to change the entire market segmentation.

For example, for a long-time, small cars created customer value through a lower price. As a result, all manufacturers segmented consumers based on socioeconomic

Table 6 Segment names and descriptions

	<i>Segment 1</i>	<i>Segment 2</i>	<i>Segment 3</i>	<i>Segment 4</i>
Description	Attention seekers	Freedom lovers	No-nonsense neutrals	Sensible classics
Size	31%	13%	26%	30%
Car preference	Unique car	Funky car	Basic car	Sensible car
Factor 1	Medium	High	Low	High
Factor 2	High	Medium	Medium	High
Factor 3	Medium	High	Medium	Low
Other descriptors	Fashion conscious	Skeptical of government regulation	Can only afford small car	Value conscious

variables and focused on reducing development and production costs. Consider the rising popularity of small cars. Various technological, environmental, and social trends made small cars attractive for other reasons than price (e.g., urban mobility, environmental footprint). These benefits also attracted new consumer segments to small cars. It was not until the introduction of the Renault Twingo and the Ford Ka, that car manufacturers realized the need for new market segmentations.

Conclusions and Managerial Implications

This chapter described the concept and process of market segmentation. The value of market segmentation has been discussed in light of recent developments, like the greater ability of firms to deliver individualized offerings, personalized communication, as well as a much greater degree of price discrimination.

This chapter ascertains that segmentation is as relevant today as it was when the concept was first introduced into management practice. In fact, technology trends have further increased the importance of distinguishing between strategic market segmentation, which guides a firm's long-term decisions (e.g., positioning or innovation and new product strategies), and operational market segmentation, which relates to real-time decisions like personalized communication (Jenkins and McDonald 1997; Sausen et al. 2005). While strategic segmentation is important no matter the firm's strategy or products, operational segmentation intuitively depends on the extent to which firms can adapt to their consumers in real time on an individual basis, which is facilitated by a greater level of digitalization (e.g., of communication). Accordingly, the digitalization provides firms with different abilities to customize their offerings and to personalize their communications and in future, customers will be able to benefit from additional offerings. Nevertheless, developments in the opposite direction exist, too, such as the increased sensitivity for privacy issues (Aguirre et al. 2016; Awad and Krishnan 2006) and data protection (as newer developments like the General Data Protection Regulation). These developments challenge the proliferation of real-time segmentation.

Despite its undeniable value to any business, market segmentation is not without controversy. Because the fundamental goal is to customize marketing actions, at least to some degree, in other words to treat different people differently, segmentation can be viewed as discrimination with another name. By choosing a target segment, firms implicitly choose to serve and focus on some customers and not the entirety of their customers. At a time when society is increasingly sensitive to all forms of discrimination, marketers can face a dilemma. By creating the basis for customization, segmentation creates value for consumers. At the same time, it creates the basis for differentiation and monopoly power, which can be detrimental to customer value. Moreover, marketing actions based on segmentation can reinforce stereotypes that at the root of gender inequality or ethical discrimination.

Even though segmentation is a well-established and fundamental concept in marketing, many firms still struggle with it. Like marketing, successful market segmentation is science and art; it requires judgment on part of market researchers

and managers. We conclude this chapter with a list of seven segmentation traps to avoid.

1. *No strategic market segmentation at all.* There is still the belief that market segmentation is something only for big consumer goods firms. Even if the number of customers is small or real-time customization is possible, segmentation improves the quality of marketing and business decisions. All firms can benefit from a strategic market segmentation because the value of segmentation derives not only from the firm's own marketing strategy but importantly from a more detailed understanding of consumers. At least, this can improve the firm's understanding of their consumers, reduce complexity, and simplify internal communications.
2. *Seeing market segmentation as a statistical exercise.* Developing a marketing strategy and selecting target customers are fundamental to the strategy processes of any business. Data alone cannot decide how to segment a market, especially not for strategic market segmentation. Management judgment is part of the process. As a result, market segmentation is always a strategic decision and should not be reduced to a statistical exercise.
3. *Confusing segments and product categories.* Market segmentation groups consumers and leads to strategies and activities that are specific to the needs and behaviors of target segments. This includes adapting products and developing a portfolio of products for the target segments. While a close relationship between target segments and product categories is important, one should not confuse them. There is no "small car" segment; there are consumers – a segment – whose needs are best addressed with a small car – a product category.
4. *Market segmentation solely with demographic variables.* A specific market segment should have a uniform reaction to a marketing action or a set of marketing actions. Consumer behavior is primarily driven by deep motivational and attitudinal factors and not demographic variables. Thus, only using demographic segmentation variables is not promising and likely fails to uncover segments that respond unambiguously to marketing activities.
5. *Segments are not identifiable.* The flipside of the fourth trap is a market segmentation that fails to assign consumers to the created segments. Deep motivational and attitudinal factors are not observable; oftentimes, they are at best revealed over time through consumers' behavior. To implement marketing actions, the target segment must be identifiable, which requires the description of segments with observable variables.
6. *Segment attractiveness is segment size.* The first goal of segmentation is to structure the market into internally homogeneous market segments. The second goal is to prioritize these segments. Segment attractiveness is not just determined by size and growth. The profitability of a market segment also depends on the competition – bigger markets attract more competition – factors as the segment's price sensitivity and whether a firm is able to efficiently serve a segment. Thus, bigger segments might not always be the more attractive ones.

7. *Considering segmentation as a static process.* A key success factor of market segmentation is to keep it relevant over the years while considering how environments and consumers change. It is thus imperative to constantly reevaluate and potentially update the market segmentation.

To conclude, market segmentation is at the core of any market strategy and is rightly considered to be one of the most important marketing tools. Newer developments shape and affect the value of market segmentation and allow firms to customize and individualize their offerings, prices, and communications. The rise of digital marketing certainly had and will continue to have a significant impact on market segmentation, but “big data” and “marketing automation” carry the danger of turning market segmentation into a technical and operational issue instead of cementing its place as the foundation of all marketing strategy and actions. Customers will continue to exhibit similarities and differences and it is up to firms to understand them and harness them for their business strategies and thus the value of market segmentation remains undisputed.

Cross-References

- ▶ [Analysis of Variance](#)
- ▶ [Applied Time-Series Analysis in Marketing](#)
- ▶ [Assessing the Financial Impact of Brand Equity with Short Time-Series Data](#)
- ▶ [Automated Text Analysis](#)
- ▶ [Bayesian Models](#)
- ▶ [Challenges in Conducting International Market Research](#)
- ▶ [Choice-Based Conjoint Analysis](#)
- ▶ [Cluster Analysis in Marketing Research](#)
- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Crafting Survey Research: A Systematic Process for Conducting Survey Research](#)
- ▶ [Experiments in Market Research](#)
- ▶ [Exploiting Data from Field Experiments](#)
- ▶ [Field Experiments](#)
- ▶ [Finite Mixture Models](#)
- ▶ [Fusion Modeling](#)
- ▶ [Logistic Regression and Discriminant Analysis](#)
- ▶ [Measuring Customer Satisfaction and Customer Loyalty](#)
- ▶ [Measuring Sales Promotion Effectiveness](#)
- ▶ [Mediation Analysis in Experimental Research](#)
- ▶ [Modeling Customer Lifetime Value, Retention, and Churn](#)
- ▶ [Modeling Marketing Dynamics Using Vector Autoregressive \(VAR\) Models](#)
- ▶ [Multilevel Modeling](#)
- ▶ [Panel Data Analysis: A Non-technical Introduction for Marketing Researchers](#)
- ▶ [Partial Least Squares Structural Equation Modeling](#)
- ▶ [Regression Analysis](#)

- ▶ [Return on Media Models](#)
- ▶ [Social Network Analysis](#)
- ▶ [Structural Equation Modeling](#)
- ▶ [Willingness to Pay](#)

References

- Aguirre, E., Roggeveen, A., Grewal, D., & Wetzels, M. (2016). The personalization-privacy paradox: Implications for new media. *Journal of Consumer Marketing*, 33(2), 98–110.
- Ansari, A., & Mela, C. (2003). E-customization. *Journal of Marketing Research*, 40(2), 131–140.
- Arabic, P. (1977). Clustering representations of group overlap. *Journal of Mathematical Sociology*, 5(1), 113–128.
- Arora, N., Drèze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., Joshi, Y. V., Kumar, V., Lurie, N. H., Neslin, S., Sajeesh, S., Su, M., Syam, N. B., Thomas, J., & Zhang, Z. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters*, 19(3), 305–321.
- Awad, N. F., & Krishnan, M. S. (2006). The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *Management Information Systems Quarterly*, 30(1), 13–28.
- Bailey, C., Baines, P., Wilson, H., & Moira, C. (2010). Segmentation and customer insight in contemporary services marketing practice: Why grouping customers is no longer enough. *Journal of Marketing Management*, 25(3–4), 227–252.
- Beane, T. P., & Ennis, D. M. (1987). Market segmentation: A review. *European Journal of Marketing*, 21(October), 20–42.
- Berry, M. J., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. Willey.
- Blattberg, R. C., & Sen, S. K. (1974). Market segmentation using models of multidimensional purchasing behavior. *Journal of Marketing*, 38(4), 17–28.
- Bock, T., & Uncles, M. (2002). A taxonomy of differences between consumers for market segmentation. *International Journal of Research in Marketing*, 19(3), 215–224.
- Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, 2.
- Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), 1–27.
- Casteleyn, J., Mottart, A., & Rutten, K. (2009). How to use data from Facebook in your market research. *International Journal of Market Research* 51(4):439–447.
- Croft, M.J. (1994), *Market segmentation: A step-by-step guide to profitable new business*. Cengage Learning Emea.
- Dellaert, B. G., & Dabholkar, P. A. (2009). Increasing the attractiveness of mass customization: The role of complementary on-line services and range of options. *International Journal of Electronic Commerce*, 13(3), 43–70.
- DeSarbo, W. S., Grewal, R., & Scott, C. J. (2008). A clusterwise bilinear multidimensional scaling methodology for simultaneous segmentation and positioning analyses. *Journal of Marketing Research*, 45(3), 280–292.
- Dhalla, N. K., & Mahatoo, W. H. (1976). Expanding the scope of segmentation research: Segmentation Research Must Cover More of the Total Marketing Problem if it is to be Operational and Profitable. *Journal of Marketing* 40(2):34–41.
- Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: Lexicon versus machine learning. *Journal of Consumer Marketing*, 34(6), 480–488.
- Dibb, S., & Simkin, L. (1997). A program for implementing market segmentation. *Journal of Business & Industrial Marketing*, 12(1), 51–65.

- Dibb, S., & Simkin, L. (2001). Market segmentation: Diagnosing and treating the barriers. *Industrial Marketing Management*, 30(8), 609–625.
- Dolnicar, S., & Leisch, F. (2010). Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, 21(1), 83–101.
- Dolnicar, S., Grün, B., & Leisch, F. (2018). Market segmentation analysis. In *Market segmentation analysis* (pp. 11–22). Singapore: Springer.
- Frank, R. E., Massey, W. F., & Wind, Y. (1972). *Market segmentation*. Englewood Cliffs: Prentice Hall.
- Gilmore, J. H., & Pine, B. J. (1997). The four faces of mass customization. *Harvard Business Review*, 75(1), 91–102.
- Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3), 389–404.
- Grover, R., & Srinivasan, V. (1987). A simultaneous approach to market segmentation and market structuring. *Journal of Marketing Research*, 24(2), 139–153.
- Hassan, S. S., Craft, S., & Kortam, W. (2003). Understanding the new bases for global market segmentation. *Journal of Consumer Marketing*, 20(5), 446–462.
- Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics*, 108, 413–423.
- Hruschka, H. (1986). Market definition and segmentation using fuzzy clustering methods. *International Journal of Research in Marketing*, 3(2), 117–134.
- Jenkins, M., & McDonald, M. (1997). Market segmentation—Organizational archetypes and research agendas. *European Journal of Marketing*, 31(1), 17–32.
- Judge, T., Matthews, T., & Whittaker, S. (2012). *Comparing collaboration and individual personas for the design and evaluation of collaboration software*. Austin: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.
- Kamakura, W. A., & Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26(4), 379–390.
- Kamakura, W. A., & Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, 37(4), 490–498.
- Kotler, P. (1989). From mass marketing to mass customization. *Planning Review*.
- Kotler, P. (1997). *Marketing management, analysis planning and control*. Englewood Cliffs: Prentice Hall International.
- Kruglanski, A. W., & Szumowska, E. (2020). Habitual behavior is goal-driven. *Perspectives on Psychological Science*, 15(5), 1256–1271.
- Lastovicka, J. L., Murry, J. P., Jr., & Joachimsthaler, E. A. (1990). Evaluating the measurement validity of lifestyle typologies with qualitative measures and multiplicative factoring. *Journal of Marketing Research*, 27(1), 11–23.
- Lilien, G.L., Rangaswamy, A. (2004), *Marketing Engineering. Computer-Assisted Marketing Analysis and Planning*, Revised 2nd ed. Trafford Publishing, Victoria.
- Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115(May), 134–142.
- Mahajan, V., & Jain, A. K. (1978). An approach to normative segmentation. *Journal of Marketing Research*, 15(3), 338–345.
- McCann, J. M. (1974). Market segment response to the marketing decision variables. *Journal of Marketing Research*, 11(4), 399–412.
- Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4), 345–389.
- Myers, J.H. (1996). Segmentation and positioning for strategic marketing decisions. *American Marketing Association*.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543.
- Öztaysi, B., and Onar, S.C. (2013). User segmentation based on twitter data using fuzzy clustering. *Data Mining in Dynamic Social Networks and Fuzzy Systems*. IGI Global, 316–333.

- Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing*, 69(4), 167–176.
- Peppers, D., & Rogers, M. (1997). *Enterprise one to one: Tools for competing in the interactive age*. New York: Currency/Doubleday.
- Piller, F. T. (2004). Mass customization: Reflections on the state of the concept. *International Journal of Flexible Manufacturing Systems*, 16(4), 313–334.
- Pires, G., & Stanton, J. (2005). *Ethnic marketing—Accepting the challenge of cultural diversity*. London: Thomson Learning.
- Pires, G. D., Stanton, J., & Stanton, P. (2011). Revisiting the substantiality criterion: From ethnic marketing to market segmentation. *Journal of Business Research*, 64(9), 988–996.
- Postma, O. J., & Brokke, M. (2002). Personalization in practice: The proven effects of personalization. *Journal of Database Management*, 9(2), 137–142.
- Ries A., and Trout J. (1980). *Positioning: The Battle for your mind*. McGrawHill.
- Roberts, J.H., Kayande, U., and Stremersch, S. (2019). *From academic research to marketing practice: Exploring the marketing science value chain*. In: How to Get Published in the Best Marketing Journals. Edward Elgar Publishing.
- Sausen, K., Tomczak, T., & Herrmann, A. (2005). Development of a taxonomy of strategic market segmentation: A framework for bridging the implementation gap between normative segmentation and business practice. *Journal of Strategic Marketing*, 13(3), 151–173.
- Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing*, 21(1), 3–8.
- Steenkamp, J.-B. E. M., & Ter Hofstede, F. (2002). International market segmentation: Issues and perspectives. *International Journal of Research in Marketing*, 19(3), 185–213.
- Tsai, C.-Y., & Chiu, C.-C. (2004). A purchase-based market segmentation methodology. *Expert Systems with Applications*, 27(2), 265–276.
- Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 8, No. 1).
- Verhoef, P. C., Venkatesan, R., McAllister, L., Malthouse, E. C., Kraft, M., & Ganesan, S. (2010). CRM in data rich multichannel retailing environments: A review and future research directions. *Journal of Interactive Marketing*, 24(2), 124–137.
- von Hippel, E. (1998). Economics of product development by users: The impact of ‘sticky’ local information. *Management Science*, 44(5), 629–644.
- von Hippel, E. (2001). Perspective: User toolkits for innovation. *Journal of Product Innovation Management*, 18(4), 247–257.
- Wedel, M., & Kamakura, W. A. (2012). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Springer Science & Business Media, LLC.
- Weinstein, A. (1987). *Market Segmentation*. Chicago: Probus Publishing Company.
- Weinstein, A. (2004). *Handbook of market segmentation: Strategic targeting for business and technology firms*. Psychology Press.
- Wilkie, W. L., & Cohen, J. B. (1977). *An overview of market segmentation: Behavioral concepts and research approaches*. Cambridge, MA: Marketing Science Institute Working paper.
- Winger, R., & Edelman, D. (1989). Segment-of-one marketing. The Boston Consulting Group, (329).
- Yankelovic, D. (1964). New criteria for market segmentation. *Harvard Business Review*, 42(2), 83–90.
- Yankelovic, D., & Meer, R. (2006). Rediscovering market segmentation. *Harvard Business Review*, 84(2), 122–131.



Willingness to Pay

Wiebke Klingemann, Ju-Young Kim, and Kai Dominik Füller

Contents

Introduction	970
Conceptual Definitions of WTP	971
Methods for Measuring Willingness to Pay	972
Stated Preference Methods	973
Revealed Preference Methods	979
Summary of Methods for Measuring WTP	982
Drivers of WTP	982
Situational Factors	987
Individual Factors	988
Information-Related Factors	988
Market Research Application	989
Elicitation of Consumers' WTP	989
Application 1: Price Bundling	992
Application 2: Personalized Pricing	993
Application 3: Nonlinear Pricing	993
Conclusion	994
References	995

Abstract

Measuring accurate willingness to pay (WTP) is essential for designing pricing policies, particularly for pricing new products. Neglecting consumers' WTP may lead to unexploited surplus when prices are set too low or to low demand when prices are set too high. Additionally, information on consumers' WTP serves as

W. Klingemann (✉) · K. D. Füller
Karlsruhe Institute of Technology, Institute for Information Systems and Marketing – Services
Marketing, Karlsruhe, Germany
e-mail: wiebke.klingemann@kit.edu

J.-Y. Kim
Goethe University Frankfurt, Department of Marketing, Frankfurt, Germany
e-mail: ju-young.kim@kit.edu

valuable input to estimate sales and for use in optimization models, thus, to maximize profit. To date, various approaches to measure WTP exist that differ regarding their elicitation approach (direct vs. experimental) and whether they rely on stated or revealed preferences (hypothetical vs. actual WTP). This chapter provides an overview of the most common methods for measuring WTP and further discusses determinants of WTP.

We further provide a practical illustration of WTP measurement. Therefore, we collected data on consumers' WTP for a hypothetical new product offer using two stated preference approaches (open-ended questions and dichotomous choice method following a sequential monadic approach) as well as one revealed preference approach (BDM mechanism). We compare the results of these different methods and discuss how to apply WTP measures in practice.

Keywords

Willingness to pay · Stated preference methods · Revealed preference methods · Hypothetical bias · Drivers of WTP

Introduction

Determining prices is a task that challenges all companies worldwide. As price is the most effective driver of profitability, superior to cost reductions and increases in sales quantity (Simon and Fassnacht 1982, pp. 1–24), optimal pricing can be considered one of the most crucial management decisions. Ideally, pricing decisions should account for production costs and competitors' prices, while also considering how much consumers are willing to spend for the product at maximum (Moorthy 1988). However, the latter, referred to as consumers' *reservation price* or consumers' *willingness to pay* (WTP) (Kalish and Nelson 1991, p. 328), is often neglected, resulting in prices that do not fully exploit consumer surplus: according to McKinsey, 80–90% of all poorly chosen prices are set too low (Marn et al. 2003). This may lead to substantial losses in revenue: for example, the car manufacturer Audi lost more than 200 million Euros because it sold its Q7 luxury SUV too cheap, thus running out of stock. Similarly, Asus, the Taiwanese electronics company, launched its mini-notebook “eee PC” in Germany at a price which was set so low that demand exceeded supply by 900% (Ramanujam and Tacke 2016, p. 22). But even if companies become aware of their undervaluation before running out of stock, they won't be able to easily adjust their prices. Consumers often react negatively to subsequent price increases, consequently leading to lower perceived value, satisfaction, and future purchase intentions (Calabuig et al. 2014). They may even boycott the seller (Sen et al. 2001).

In contrast, an overvaluation, i.e., when prices are set too high, may be adjusted more easily from a consumer's point of view, but it may often lead to market entry difficulties or even product failure. For instance, the introduction of Apple's earlier devices, the handheld Newton (introduced in 1993) and the gaming machine Pippin (introduced in 1995), failed completely as they were perceived as too expensive (Greenberg 2008). Particularly companies that tend to “over-engineer,” i.e.,

equipping their products with features that consumers do not value, encounter the problem of overvaluation. For example, Amazon took a 170-million-dollar write-down in 2014 because its high-end positioned “Fire Phone” was equipped with too expensive features that consumers were not willing to pay for (Ramanujam and Tacke 2016, pp. 16–17).

Without knowing consumers’ WTP, it is not only difficult to design a profitable product, it is also impossible to make educated decisions about how a new product should be introduced and offered, such as whether to create product bundles (Chung and Rao 2003), to partition prices (Hamilton and Srivastava 2008), or to estimate the effects of price promotions (Shaffer and Zhang 1995).

Due to the interdependence of price and demand, even small changes in price may significantly influence both the overall market share and profits (Winer 2005). Thus, knowing consumers’ true WTP can be considered the most important information for estimating sales (Jedidi and Jagpal 2009, pp. 37–38) and, ultimately, maximizing revenue.

As consumers’ true WTP is an unobservable construct, the challenge comprises a valid elicitation manner to find out its true value (Voelckner 2006, p. 137). Therefore, various methods for measuring consumers’ WTP have already been developed, with constant efforts to improve weaknesses associated with these measurement approaches, such as high complexity and costs of measurement, biases, lack of realism, or insufficient information specificity.

Therefore, this chapter provides an overview of the most common methods to measure WTP and gives some insights about how to counter these possible weaknesses associated with these methods. We further discuss the importance of context effects. Independent of the method, researchers as well as practitioners have to take into account that consumers’ WTP is usually not fixed but dependent on the respective context. Consumer preferences are not stable but are often newly formed during a choice situation, affected by various personal and contextual factors (e.g., Slovic 1995; Bettman et al. 1998; Hoeffler and Ariely 1999). Considering the importance of price to be a profit driver, it is therefore essential to know what circumstances influence consumers’ WTP and to understand when and why it changes. In addition to an overview of common methods to measure WTP, we discuss situational, individual, and information-related factors driving consumers’ WTP.

To illustrate the theory, we present a simple example and compare the results of three elicitation methods. We measure consumers’ WTP for a hypothetical new product offer using two stated preference approaches (open-ended questions and dichotomous choice method following a sequential monadic approach) as well as one revealed preference approach (BDM mechanism). At last, we further discuss how to apply elicited WTP in practice.

Conceptual Definitions of WTP

When estimating WTP, the aim is to determine the maximum price a consumer would be prepared to pay, that is “the maximum sacrifice, in terms of [. . .] money, that one is willing to make to obtain a commodity” (Donaldson 1999, p. 551). This

means that WTP is not necessarily equivalent to consumers’ estimation of the value of a product. Consumers may have a WTP below what they believe a product to be worth if they cannot afford to pay more, that is, if their ability to pay is limited (Russell 1996). Therefore, measurements of WTP try to determine “the price at or below which a consumer will demand one unit of the good” (Varian 1992, p. 152). While this implies that a consumer will definitely make a purchase, other definitions state that WTP is the price “at which a consumer would no longer purchase” (Hauser and Urban 1986, p. 449) or “at which a consumer is indifferent between buying and not buying the product” (Jedidi and Zhang 2002, p. 1352).

These marginal technical differences in the definition of WTP illustrate the difficulty of generating a specific point estimate, raising the question whether such a price point exists. From an economic viewpoint, “WTP for a product is the amount of income that will compensate for the loss of utility obtained from the product” (Allenby et al. 2014, p. 430). As it is highly difficult for consumers to determine exactly how much utility they will derive from a product, more recent research suggests that “rather than specific WTP values for products, consumers probably have some range of acceptable values” (Ariely et al. 2003, p. 77). If the price falls below the lower bound of this range (“floor reservation price”), they will definitely buy; if it exceeds the upper bound of the range (“ceiling reservation price”), they will definitely not buy. Within this range, consumers’ response is not clearly predictable (Ariely et al. 2003; Wang et al. 2007). Under the premise that the WTP distribution is symmetric within the uncertainty interval, expected WTP is the midpoint of this range (Dost and Wilken 2012, p. 149).

Methods for Measuring Willingness to Pay

Methods for measuring WTP can be differentiated with regard to their elicitation approach (direct vs. experimental) and whether they rely on stated or revealed preferences (hypothetical vs. actual WTP) (c. Miller et al. 2011). With stated preference methods, participants’ answers are taken “as stated,” and their choices are only of hypothetical nature. Revealed preference methods, in contrast, lead to real consequences and actual purchases.

The following table provides an overview of common methods used to measure WTP (Table 1):

Table 1 Overview of common methods for measuring consumers’ WTP

	Stated	Revealed
Direct	Consumer surveys, e.g., – <i>Open-ended questions</i> – <i>Dichotomous choice method</i> – <i>Payment card method</i> Expert opinions	Auctions, e.g., – <i>Vickrey auctions</i> – <i>BDM mechanism</i> Market data
Experimental	Conjoint analysis Choice-based conjoint analysis	Lab experiments Field experiments

Direct methods estimate WTP by directly asking

- Consumers how much money they are willing to spend (i.e., open-ended questions or closed questions using dichotomous choice or payment card methods; or consumer auctions where consumers have to make bids pursuant to the Vickrey rule or the BDM mechanism)
- Experts about prices they believe achievable (i.e., expert judgments, management discussions)
- The market, by analyzing past (i.e., “natural”) market data that give insight about which prices are accepted

Experimental methods, in contrast, actively create or manipulate choice situations that are affected by price, deducting WTP from participants’ behavior, without explicitly asking about the price itself.

These can be

- Hypothetical choice scenarios where participants have to evaluate options or choose among options (i.e., conjoint analysis, choice-based conjoint analysis)
- Real, but artificial choice situations where participants have to make actual purchase decisions (i.e., lab experiments)
- Real, natural choice situations, where participants are not aware that their purchase decisions are part of a pricing experiment and that prices are manipulated for experimental purposes (i.e., field experiments)

In the following, we will explain each method in detail, having a closer look on the most popular ones. We then discuss advantages and disadvantages associated with each method and conclude with an overview of method validity and feasibility.

This section concludes with a guide suggesting when to apply which method.

Stated Preference Methods

Direct Stated Preference Methods

Direct stated preference methods can be divided into *consumer surveys* and *expert opinions*. They are usually the fastest and simplest ways to measure consumers’ WTP.

The *open-ended questions method* asks consumers directly how much they would be willing to pay for a certain good or service. This is most probably the easiest method to use. A special variant of the open-ended questions is the *van Westendorp method*, a consumer survey that also measures price perception and price sensitivity (Müller 2009). The van Westendorp method generates a pricing corridor based on four questions about what price consumers would consider too cheap vs. attractively cheap as well as expensive but acceptable versus too expensive (van Westendorp 1976). The van Westendorp method thereby generates a price sensitivity meter (PSM). Figure 1 depicts an example on how to interpret the results of such a survey.

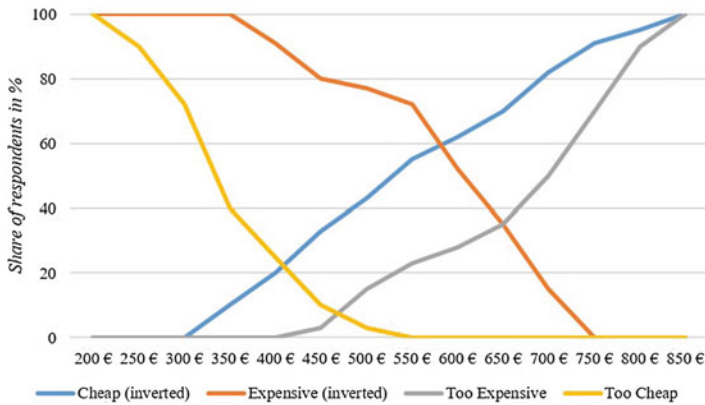


Fig. 1 Example for a price sensitivity meter

The intersection of “too cheap” and “cheap” (inverted) and the intersection of “too expensive” and “expensive” (inverted) mark the lower ends (LE) and upper ends (UE) of the price corridor for optimal pricing. In the example, the product should be priced between 410€ and 650€.

The van Westendorp method allows respondents to indicate a price range instead of one price point, an approach supported by recent research. Wang et al. (2007) argue that consumers are often not completely sure about their own preferences and the exact performance of a product, which means that forcing them to state their WTP as one absolute point does not adequately reflect reality. They therefore propose to ask respondents at which price they would definitely buy/be indifferent between buying and not buying/buy with a very low probability, thereby generating a price range. Dost and Wilken (2012) support this approach by arguing that range-based methods allow consumers to indicate a price range that reflects their decision-making uncertainty, whereas point-based methods ignore this uncertainty and force consumers to make a guess about their WTP that might or might not be true.

However, open-ended questions methods, both point-based methods and price range methods, require high cognitive effort from consumers, who are used to evaluating given prices instead of generating prices on their own (Chernev 2003). Therefore, *dichotomous choice methods* use closed, polar questions, asking respondents whether they would buy a product at a specific price. Dichotomous choice methods are easy to implement; however, the consumer’s response only gives information about whether the price is acceptable or not but lacks information on precise WTP. Thus, researchers may work with either a large sample size (*monadic approach*) or with a sequential approach (*sequential monadic approach*). Applying the *monadic approach*, participants are divided into different groups, with each group being confronted with a different price. The amount of price points should be limited, as each requires a new group of participants (Lyon 2002, p. 10). An alternative approach that does not require such a high sample size is the *sequential monadic approach*, sometimes also referred to as *Gabor Granger method*.

Participants who answer that they would not buy the product at the given price will be asked the same question again with a lower price (with the exception that the rejection is based on the fact that the price was too cheap). In contrast, participants accepting the initial price presented, will be asked whether they would buy the product at a higher price, until the maximum WTP is determined (Gabor and Granger 1966).

An undesired effect of this approach is that the price attribute becomes very dominant in contrast to the other attributes of the good or service. This method also suffers from a starting point bias, meaning that the first price serves as an anchor and impacts participants' evaluation of the subsequent prices (Herriges and Shogren 1996).

The *payment card method* avoids this bias by presenting participants with an array of prices, asking them to indicate which of the presented prices reflects their WTP (Fig. 2). However, here researchers must pay attention to select intervals that actually cover participants' differences in WTP, and to choose a decent number of price points manageable for participants (Rowe et al. 1996).

Direct stated methods are the most common method used for contingent valuation, i.e., for measuring the value of goods that cannot be sold, such as public or environmental goods (e.g., air quality) (Mitchell 2013). In contingent valuation, survey participants are asked to imagine how much they would be willing to pay to improve the status quo or to prevent it from deteriorating (e.g., to prevent a reduction of air quality), for instance, through taxes or donations that help to improve or preserve the current status (see Boyle 2017 for a detailed description of how to conduct a contingent valuation study).

While contingent valuation surveys need a sample of participants that represent the respective interest group, companies conducting WTP surveys can also rely on *expert opinions*, thus questioning internal market experts instead of consumers. These experts can be marketing managers or sales people who work closely with consumers and can therefore provide relevant insights for pricing decisions. A disadvantage of working with market experts is that their judgments might be biased. Particularly, when their compensation depends on the number of sales, they may

Please indicate how much you would be willing to annually pay to reduce current night-time noise in your neighborhood by 30%:

<input type="checkbox"/> 0 €	<input type="checkbox"/> 150 €
<input type="checkbox"/> 20 €	<input type="checkbox"/> 200 €
<input type="checkbox"/> 40 €	<input type="checkbox"/> 250 €
<input type="checkbox"/> 75 €	<input type="checkbox"/> 300 €
<input type="checkbox"/> 100 €	<input type="checkbox"/> more than 300 €

Please, indicate your maximum amount: _____ €

Fig. 2 Example for a contingent valuation survey using the payment card method

have an incentive to state a lower price than they believe achievable (Hanna and Dodge 1995, p. 70).

Consumer surveys involve a similar incentive problem: as a direct question predominantly shifts attention toward the price, consumers are tempted to act prospectively and state a price lower than their true WTP in order to save money (Lyon 2002, p. 9).

Qualitative research has therefore made efforts to limit strategic answering by improving the formulation of direct questions. For instance, instead of asking what price a respondent would be willing to pay, the researcher could alternatively ask the respondent to state a price that he/she considers fair for both seller and buyer or to ask the respondent to guess at which price the product would sell (see Henderson (2002) for a more detailed description of possible qualitative questions).

Another, although more complex way to eliminate the problems of consumers' strategic understatements of their WTP is to use *experimental stated preference methods*.

Experimental Stated Preference Methods

Experimental stated preference methods, also known as *conjoint analyses*, employ choice scenarios where respondents have to compare and evaluate different options. These methods are decompositional, which means that the respondents' ratings of the presented products, or their choice decisions, are used to deduct WTP for the overall product and its different features ("top-down approach"). Despite being more complex and time-consuming than direct surveys, the holistic approach of these analyses is closer to consumers' actual product evaluation behavior and puts less emphasis on the price attribute. It thereby overcomes most of the previously mentioned problems of direct WTP measurements, such as strategic or biased answering behavior or respondents' difficulties in accurately stating their WTP (Green and Srinivasan 1990).

The underlying idea of conjoint analyses is that consumers' WTP for a product depends on the utility they derive from it, with the overall utility being the sum of the utilities they derive from the different product attributes.

Traditional *conjoint analysis* (CA) (Luce and Tukey 1964) determines these utilities by asking respondents to rank (or rate) different products that are presented through multiple relevant product attributes and product attribute values (e.g., Green and Rao 1971; Green and Srinivasan 1978; see Gustafsson (2007) for a detailed description of different forms and applications of conjoint analysis).

Choice-based conjoint analysis (CBC) (Louviere and Woodworth 1983) does not demand respondents to rank different products but rather to choose among them, including the option to choose nothing (Fig. 3). Therefore, the main advantage of CBC is its similarity to real choice situations (compare Louviere et al. (2000) for a detailed analysis of both approaches).

CA relies on the assumption that respondents derive more utility from the product the higher they rank it. CBC infers that the utility of an option that has been chosen is positive – otherwise the no-choice option would have been selected – and higher

3 Please choose the Bundle you would most likely consider buying.

Shipping Time	2 Business Days	2 Business Days	1 Business Day	
Shipping Cost	Free of Charge at a Minimum Order Value of 50€	Free of Charge	Shipping Coast charged	
Monthly Bill	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	I would not consider to buy any of these bundles.
Webinars	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Special online account with access to exclusive content	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	
Price per Month	9€	39€	29€	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	

Fig. 3 Example for a choice-based conjoint task for an industrial service bundle

than the utilities of each of the rejected options (Halme and Kallio 2011). Using statistical estimation methods, such as monotonic analysis of variance or multiple regression (c. Green and Srinivasan 1990, p. 5), an additive utility function can then be generated (see Eq. 1).

Equation 1 Exemplary additive utility function of a consumer h for a product i .

$$\hat{U}_{h,i} = \hat{\beta}_{h,0} + \hat{\beta}_{h,j=Price} * x_{i,j=Price} + \hat{\beta}_{h,j=Attribute_1,m=Value_1} * x_{i,j=Attribute_1,m=Value_1} + \dots + \hat{\beta}_{h,j=Attribute_n,m=Value_n} * x_{i,j=Attribute_n,m=Value_n}$$

Equation 1 shows an exemplary basic formula of a utility function for a consumer h and a product i . It starts with a constant ($\beta_{h,0}$) that represents the basic utility that consumer h derives from product i . The function includes the attribute “price,” which is coded with a vector model, meaning that the utility increases/decreases linearly with the attribute value, i.e., the price level. The price parameter ($\beta_{h,j = Price}$) is usually negative, meaning the overall utility of a product decreases with increasing price and vice versa (there are rare exceptions, for instance in the area of luxury goods). For example, if $\beta_{h,j = Price}$ is estimated as -1.5 and the price equals 10 monetary units, the utility of the product decreases by 15 utility points. Apart from the price attribute, all further relevant attributes must be considered in the function. Nominal attributes are coded with a part-worth utility model, which means that a parameter is estimated for each specific attribute value, with the specific values being incorporated into the function via dummy variables. By offsetting the utility of an attribute with the price that results in a utility of zero, WTP can be calculated for each product attribute and consequently, the overall product if all relevant attributes are considered (Kohli and Mahajan 1991). More precisely, WTP for a specific attribute value is calculated by dividing its part-worth utility through the nominal value of the price parameter of the utility function (compare Eq. 2).

Equation 2 Calculation of WTP using utility parameters

$$WTP_{h,j=\text{Attribute}_x,m=\text{Value}_x} = \frac{\hat{\beta}_{h,j=\text{Attribute}_x,m=\text{Value}_x}}{|\hat{\beta}_{h,j=\text{Price}}|}$$

For example, if attribute 1 was color, with value 1 = “red” and value 2 = “blue,” a parameter of 3 (6) for value 1 (value 2) would indicate that the utility increases by 3 utility points if the product is red, and by 6 utility points if the product is blue. This means consumer h would be willing to pay 2 monetary units if product i was red ($\rightarrow 3: |-1,5| = 2$) and 4 monetary units if it was blue ($\rightarrow 6: |-1,5| = 4$). Consequently, consumer h would be willing to pay 2 monetary units more if product i was blue instead of red ($\rightarrow (6-3): |-1,5| = 2$).

For a more comprehensive description of how to conduct conjoint experiments, please refer to the chapter “[Choice-Based Conjoint Analysis](#)” of this handbook.

Although (choice-based) conjoint analysis relies on several assumptions that are often flawed in reality (e.g., assuming a compensatory relationship between different attributes; compare Srinivasan 1988), it has proven to be a reliable tool for estimating sales and is widely used in both research and practice (Green et al. 2001). However, like all stated preference methods, it suffers from *hypothetical bias*.

Hypothetical Bias in Stated Preference Methods

Stated preference methods suffer from hypothetical bias, which occurs when values are collected in a hypothetical context (Harrison and Rutström 2013). In hypothetical contexts, consumers tend to neglect thinking about what they could alternatively do with the money that is required to buy a product, leading to an exaggerated stated willingness to purchase (Dhar and Gorlin 2013, p. 533). Comparing hypothetical and real purchases, participants more often indicate that they would buy a product in hypothetical contexts (Miller et al. 2011). They further tend to report a significantly higher WTP when they do not have to pay the stated amount in reality (Neill et al. 1994; Voelckner 2006). Consequently, stated purchase intention and related WTP can differ substantially from real behavior (Kalwani and Silk 1982; Morrison 1979).

As stated preference methods are sometimes indispensable, for example to elicit WTP for new products that are not available yet, research has tried to address hypothetical bias in stated preference methods altering qualitative and quantitative aspects.

One simple suggestion is to make respondents more aware of their hypothetical answering behavior. Researchers may inform participants about hypothetical bias and request them to think about what they would really do, i.e., use *cheap talk*. However, results of this approach are mixed, with some studies demonstrating a reduction of hypothetical bias (e.g., Cummings and Taylor 1999) that others could not confirm (e.g., Blumenschein et al. 2008). Another approach is to remind participants of their budget constraints, making them aware that they could also use their money for something else. Effects of this approach are not consistent either, with

some research reporting improved results (e.g., Frederick et al. 2009) whereas other research found no difference (e.g., Loomis et al. 1994).

Another alternative to reduce hypothetical bias is the *certainty approach*. The certainty approach only considers data from respondents who are highly sure about their answer. These responses were shown to be closer to real behavior (Blumenschein et al. 2008). Similarly, Hofstetter et al. (2013) note that not all consumers are equally suitable for WTP estimations. Their research on WTP for innovative products shows that individual personal characteristics of respondents regarding abilities and motivation have a positive influence on the validity of their WTP estimates. However, to exclude respondents whose answers are possibly less viable requires a high number of participants.

Quantitative efforts in research may also reduce hypothetical biases. Lieven and Lennerts (2013) present a survey method where participants have to make trade-offs between cash and vouchers that are earmarked for a specific (hypothetical) product/service. Regardless of its face value, the subjective value of the voucher can naturally never be higher than participants' WTP for the respective product. This method enables researchers to use participants' choice of cash/voucher combinations to draw inferences about their WTP. Ding et al. (2005) have developed an approach that makes conjoint analyses incentive-compatible, leading participants to reveal their actual WTP. Their research suggests to combine conjoint analysis with the Becker–DeGroot–Marschak (BDM) mechanism (Becker et al. 1964; see below for a more detailed description), a revealed preference method (Ding 2007) or to tell participants that they will have the chance of winning their favorite product that will be determined through their answers in the analysis (Dong et al. 2010).

Another option to circumvent the problem of hypothetical bias occurring with stated preference methods is to use *revealed preference methods* instead.

Revealed Preference Methods

Direct Revealed Preference Methods

Direct revealed preference methods are *auctions* or the analysis of *market data*.

Auctions are used in the field (e.g., traditional auctions or online auctions such as eBay) and in the lab as a direct method to elicit WTP, which is revealed through participants' bids and bidding behavior. In the following, we will discuss the most common auction mechanisms.

In *English auctions*, the highest bid wins the auction and determines the purchase price. Competing bidders can overbid each other openly as often as they want what may lead to a systematic overestimation of WTP: they may bid prices that are higher than that of the average consumer, thereby distorting the overall data (Barrot et al. 2010). The phenomenon of auction fever exacerbates this problem. A bidder facing auction fever may be affected by various factors, such as rivalry, social facilitation, time pressure, or the joy of winning, thus leading to overbidding (Ku et al. 2005). However, depending on competition intensity, bidders may also bid below their true WTP.

First-price auctions do not suffer from the problem of auction fever, as they allow only one sealed bid, with the highest bidder obtaining the auctioned good at the price of his or her bid. However, bidders again have an incentive to bid below their true WTP, hoping to still win the auction and to obtain the good at a price lower than their WTP (Hoffman et al. 1993).

In *Dutch auctions*, the auctioneer sets a high starting price and decreases the price by a predetermined increment until one bidder is willing to buy the product at this price. Again, bidders do not have an incentive to bid their true WTP: knowing that they can acquire the good at a lower price if no one else purchases it, bidders may refrain from buying, even if the price matches their WTP. Due to higher transaction costs – with bidders having to wait for the price to decrease – the resulting prices may be higher than with first-price auctions (Carare and Rothkopf 2005) but still not necessarily equal to bidders' actual WTP.

Name-Your-Own-Price (NYOP), also known as *reverse pricing*, can be considered a special kind of auction: consumers do not compete with each other but simply name the price that they are willing to pay, which can be accepted or refused by the seller depending on a threshold value. When multiple bidding is possible, WTP can be derived from the consumers' bidding behavior, assuming that the threshold price is uniformly distributed on a minimum /maximum price interval and frictional costs are constant. As consumers will maximize their expected surplus (i.e., $(WTP - \text{price paid}) \times \text{probability of bid acceptance} - \text{frictional costs of submitting bids}$), it is thus possible to calculate WTP from the number and values of submitted bids (Spann et al. 2004). However, NYOP is also described as "online haggling" (Terwiesch et al. 2005), with buyers trying to make a bargain instead of revealing their actual WTP.

The above described auctions are not incentive-compatible, as bidders may withhold their true WTP. Thus, research has suggested the use of *second-price sealed-bid auctions*, also known as *Vickrey auctions* (Vickrey 1961) and the *Becker–DeGroot–Marschak (BDM) mechanism* (Becker et al. 1964) as incentive-compatible methods.

The mechanism of Vickrey auctions is considered the least prone to biases (Noussair et al. 2004). In the Vickrey auction, each bidder submits a sealed bid. Also here, the highest bidder wins the auction, but the purchase price is determined by the bid of the second-highest bidder (Vickrey 1961). This way, participants do not have any incentive to withhold their true WTP, as they will never pay more than what is necessary to win the auction. If they submit a lower bid than their actual WTP, they may risk losing the auction and the possibility to purchase the product. Bidding above their true WTP is not an optimal strategy either, because participants may win the auction but at a price higher than their WTP. Thus, Vickrey auctions are robust against many kinds of strategic behavior (Barrot et al. 2010). However, there are also limitations associated with this method. Some researchers criticize that bidders assuming that their WTP is too low to win the auction do not have an incentive to bid sincerely (Shogren et al. 2001). They may either submit a lower bid due to lack of motivation or a higher bid in order to increase the price the winning competitor has to pay, leading to distorted WTP results. Further limitations refer to operational problems with the implementation of auctions in general (Wertenbroch and Skiera 2002)

and to the fact that auctions typically do not meet realistic decision processes in retailing (Hoffman et al. 1993).

The BDM mechanism (developed by Becker et al. 1964) determines the price of a product through a random draw. First participants submit a sealed bid then someone (e.g., the auctioneer) draws a “market” price. Bidders with bids equal to or higher than the drawn price are obliged to buy the product, whereas those who submitted bids below the “market” price are not allowed to buy it. The mechanism is incentive-compatible, as bidding anything else than one’s actual WTP is not an optimal strategy. Bidding more than one’s WTP does not affect competing bidders but may result in a purchase obligation at a price above one’s WTP. Bidding below the WTP bears the risk of not being allowed to purchase the good. As the price that needs to be paid is determined by a random draw, there is no incentive to deviate from bidding one’s true WTP in the hope of impacting the final price. While this mechanism is robust against biases, it suffers from the limitations associated with auctions: the BDM mechanism is difficult to implement in practice and also does not reflect most regular purchase situations, as consumers typically do not have to bid for a limited resource (Wertenbroch and Skiera 2002). Instead, they are influenced by reference prices, deciding whether or not to purchase the product at a given price.

This problem occurs with most direct methods: the generated answers do not fully represent reality. The only direct approach that circumvents this problem is the use of real instead of generated data, that is, the analysis of *market data*. These can be a company’s own sales figures, panel data provided by market research agencies, or store scanner data from participating stores (Breidert 2006, p. 39). Based on actual purchase data, market data are incentive-compatible and have high external validity. However, this approach bears obvious weaknesses: first, market data are historical data and are therefore unavailable for new products that have not been put on the market yet. Second, market data only provide information on how many units were sold at a given price, but not on how many people would have paid a higher price, or refrained from purchasing a product due to its price. As this represents a serious constraint to measuring WTP, some researchers have suggested that the analysis of market data should be combined with stated preference methods to generate a more complete picture (Ben-Akiva et al. 1994). An additional severe problem is that market data are confounded by noise, such as promotions or competitor activities, thus, they do not allow for a systematic, clear variation of prices.

One way to investigate actual purchase behavior contingent on price variations is to conduct price *experiments*.

Experimental Revealed Preference Methods

Experimental revealed preference methods estimate WTP using *laboratory experiments* or *field experiments*.

When conducting price experiments, researchers create purchase scenarios that include purchase obligations. Experiments can be conducted in a laboratory, offering the relevant products for sale to the participants, or in the field, selling

the products in a real store. When full purchase obligation is not feasible in lab experiments, a helpful solution may be to include a purchase obligation for only a fraction of the participants (e.g., 10%) who are determined by lot (Voelckner 2006).

Assuming that consumers buy as long as the price does not exceed their WTP, the experimenter systematically varies the price and measures the resulting purchase rates and quantities. In laboratory experiments, participants are randomly divided into groups that are confronted with the same purchase situation, but with a variation in price between the groups. In field experiments, prices are usually varied in predefined time intervals. Analyzing the units sold across the different time intervals and prices, the average WTP can be estimated.

Laboratory experiments do not suffer from hypothetical bias; however, participants' WTP may be censored by field opportunities. Participants who know the actual price of a product offered in the laboratory may use this value as a reference and refuse to pay more than it (Harrison et al. 2004). Furthermore, laboratory experiments also suffer from experimental bias: participants may behave not the same way as they would in real purchase settings, knowing that they are participating in an experiment.

In contrast, field experiments have the highest external validity when it comes to measuring true WTP, as they measure WTP under reality conditions. A limitation is, however, that they usually do not allow for the collection of additional controls (e.g., age, income, varying context factors). Field experiments also require high organizational and logistic efforts.

Summary of Methods for Measuring WTP

In the following, we compare the presented methods along main criteria that impact their external validity and feasibility (Table 2).

We then shortly summarize all previously discussed preference measurement methods, presenting examples on how and when to use the methods and provide advantages and disadvantages associated with each kind of approach (Table 3). The tables shall serve as a guidance for academics and practitioners who need to decide which approach is most suitable in which case.

Drivers of WTP

WTP is difficult to measure not only for methodological reasons, but also due to the instability of the value itself. Individual WTP is not constant, but is highly context-dependent and subject to various influencing factors. The following figure provides an overview of important situational, individual, and information-related factors affecting WTP (Fig. 4).

Table 2 Overview of important factors regarding external validity and feasibility of different WTP measurement approaches

Measurement approach	Validity of results					Feasibility of approach			Complexity/ Costliness
	Hypothetical bias	Strategic answering bias	Experimental bias	Similarity to real (choice) situations	Consideration of multiple (competitor) products	Applicable for products under development			
Consumer surveys	yes	yes	yes	low-moderate	no	yes	yes	low	
Expert opinions	/	yes	/	/	yes	yes	yes	low	
Vickrey auctions	no	no	yes	low	no	no	no	moderate	
BDM mechanism	no	no	yes	very low	no	no	no	moderate	
Market data	no	no	no	equal	yes	no	no	moderate-high	
Conjoint analysis	yes	no	yes	low	yes	yes	yes	moderate-high	
CBC	yes	no	yes	high	yes	yes	yes	moderate-high	
Lab experiments	no	no	yes	high	yes	no	no	high	
Field experiments	no	no	no	equal	yes	yes	no	high	

Table 3 Common methods for measuring WTP in detail: recommended application, examples, and approach-specific advantages and disadvantages

Direct stated preference methods													
Recommended application	<ul style="list-style-type: none"> - Easy/fast/comparably cheap measurement of WTP - As a starting point, to get a first impression about achievable prices for new products - For contingent valuation 												
	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 25%;"></th> <th style="width: 45%;">Advantage of specific kind of DSP method</th> <th style="width: 30%;">Disadvantage of specific kind of DSP method</th> </tr> </thead> <tbody> <tr> <td>Open-ended questions</td> <td> <p>Example (WTP for a salad)</p> <p>How much would you pay for this salad?</p> </td> <td> <p>No starting point bias</p> <p>Consumers are not used to formulate WTP without reference values</p> </td> </tr> <tr> <td>Van Westendorp method</td> <td> <ol style="list-style-type: none"> 1. At which price would you think that the salad is expensive, but still worth buying? 2. At which price would you think the salad is cheap, without doubting its quality? 3. At which price would you consider the salad to be too expensive? 4. At which price would you consider the salad to be too cheap, in the sense that you would doubt its quality? <p><i>Monadic approach</i></p> </td> <td> <p>Generates a pricing corridor that reflects consumer uncertainty</p> <p>May result in a very wide price range</p> </td> </tr> <tr> <td>Dichotomous choice method</td> <td> <p>Would you pay 4 € for this salad? (group 1) ... 4.50 € for this salad? (group 2)/... n € for this salad? (group n)</p> <p><i>Sequential monadic approach</i></p> <p>Would you pay 4 € for this salad? (each participant)</p> <p>→ If answer is yes, increase price: ... 4.50 € for this salad?</p> <p>→ If answer is no, decrease price: ... 3.50 € for this salad?</p> <p>→ Continue until the accepted price cannot be increased anymore.</p> </td> <td> <p>Requires a large amount of participants</p> <p>Starting point bias</p> <p>Puts strong focus on the price attribute</p> </td> </tr> </tbody> </table>		Advantage of specific kind of DSP method	Disadvantage of specific kind of DSP method	Open-ended questions	<p>Example (WTP for a salad)</p> <p>How much would you pay for this salad?</p>	<p>No starting point bias</p> <p>Consumers are not used to formulate WTP without reference values</p>	Van Westendorp method	<ol style="list-style-type: none"> 1. At which price would you think that the salad is expensive, but still worth buying? 2. At which price would you think the salad is cheap, without doubting its quality? 3. At which price would you consider the salad to be too expensive? 4. At which price would you consider the salad to be too cheap, in the sense that you would doubt its quality? <p><i>Monadic approach</i></p>	<p>Generates a pricing corridor that reflects consumer uncertainty</p> <p>May result in a very wide price range</p>	Dichotomous choice method	<p>Would you pay 4 € for this salad? (group 1) ... 4.50 € for this salad? (group 2)/... n € for this salad? (group n)</p> <p><i>Sequential monadic approach</i></p> <p>Would you pay 4 € for this salad? (each participant)</p> <p>→ If answer is yes, increase price: ... 4.50 € for this salad?</p> <p>→ If answer is no, decrease price: ... 3.50 € for this salad?</p> <p>→ Continue until the accepted price cannot be increased anymore.</p>	<p>Requires a large amount of participants</p> <p>Starting point bias</p> <p>Puts strong focus on the price attribute</p>
	Advantage of specific kind of DSP method	Disadvantage of specific kind of DSP method											
Open-ended questions	<p>Example (WTP for a salad)</p> <p>How much would you pay for this salad?</p>	<p>No starting point bias</p> <p>Consumers are not used to formulate WTP without reference values</p>											
Van Westendorp method	<ol style="list-style-type: none"> 1. At which price would you think that the salad is expensive, but still worth buying? 2. At which price would you think the salad is cheap, without doubting its quality? 3. At which price would you consider the salad to be too expensive? 4. At which price would you consider the salad to be too cheap, in the sense that you would doubt its quality? <p><i>Monadic approach</i></p>	<p>Generates a pricing corridor that reflects consumer uncertainty</p> <p>May result in a very wide price range</p>											
Dichotomous choice method	<p>Would you pay 4 € for this salad? (group 1) ... 4.50 € for this salad? (group 2)/... n € for this salad? (group n)</p> <p><i>Sequential monadic approach</i></p> <p>Would you pay 4 € for this salad? (each participant)</p> <p>→ If answer is yes, increase price: ... 4.50 € for this salad?</p> <p>→ If answer is no, decrease price: ... 3.50 € for this salad?</p> <p>→ Continue until the accepted price cannot be increased anymore.</p>	<p>Requires a large amount of participants</p> <p>Starting point bias</p> <p>Puts strong focus on the price attribute</p>											

Payment card approach	Which of these prices is the highest you would pay for this salad? 2 €/2.50 €/3 €/3.50 €/4 €/4.50 €/5 €	Answering is comparably easy for respondents No starting-point bias	Respondents' answers are impacted/restricted by the provided range
Expert opinions	In your opinion, which price will our customers pay for this product?	Easy and fast to conduct Expert perspective	Quality of results depends on expertise
Experimental stated preference methods			
Recommended application	<ul style="list-style-type: none"> - If revealed preference methods are not feasible (e.g., unavailable new products; very expensive products) - If multiple (competitor) products shall be considered - To determine not only overall WTP, but also WTP for specific product attributes 		
	Example (WTP for a salad)	Advantage of specific kind of ESP method	Disadvantage of specific kind of ESP method
Conjoint analysis	<p>Please rank the following options according to your preference:</p> <ul style="list-style-type: none"> - Salad 1: Green salad, tomatoes, onions, 500 grams, organic, 2.99€ - Salad 2: Corn salad, tomatoes, no onions, 400 grams, organic, 3.49€ - [...]] - Salad n: 	Less repetitive for participants than choice-based conjoint analysis	Ranking task does not necessarily reflect actual choice behavior, might be difficult for participants
Choice-based conjoint analysis	<p>Which of these options would you choose:</p> <ul style="list-style-type: none"> - Salad 1: Green salad, tomatoes, onions, 500 grams, organic, 2.99€ - Salad 2: Corn salad, tomatoes, no onions, 400 grams, organic, 3.49€ - None of the two 	Mimics actual choice behavior	Number of choice scenarios must be limited to avoid overburdening participants
Direct revealed preference methods			
Recommended application	<p>Auctions</p> <ul style="list-style-type: none"> - Survey method without hypothetical bias/strategic answering bias - Especially suitable for unique products/products of limited availability <p>Market data</p> <ul style="list-style-type: none"> - When data about past purchases is available 		

(continued)

Table 3 (continued)

Direct stated preference methods			
	Example for questions (WTP for a salad)	Advantage of specific kind of DRP method	Disadvantage of specific kind of DRP method
Vickrey auction	Please state the maximum amount of money you would be willing to pay for this salad. In case you are the highest bidder, you must purchase the salad at the price of the second highest bidder. Otherwise, you are not allowed to purchase the salad	Can be put directly into practice for selling unique products	Consumers with low WTP might place a bid that does not correspond to their WTP (lack of motivation/bid with the purpose of raising prices for competitors)
BDM mechanism	Please state the maximum amount of money you would be willing to pay for this salad. We will then determine the price of the salad through a random draw. In case your bid is equal or above this price, you must purchase the salad at the price determined by the random draw. Otherwise, you are not allowed to purchase the salad	Bids are not impacted by beliefs about the behavior of other bidders	Consumers must believe in the randomness of the draw
Market data	Data from purchase panels (e.g., Nielsen), own shop, etc.	Real data/Secondary data (= data exists already)	Historical data To deduct WTP from accepted prices, variation is necessary
Experimental revealed preference methods			
Recommended application	<ul style="list-style-type: none"> - If reduction of biases is more important than costs - If competitor products shall be considered and the product is market-ready 		
	Example for questions	Advantage of specific kind of ERP method	Disadvantage of specific kind of ERP method
Lab experiments	Do you want to purchase one of these salads at the prices stated? If yes, we will sell you the salad at the price stated	Control variables (age, gender, income, etc.) can be collected	Experimental bias
Field experiments	Salad is offered in a regular store environment (e.g., supermarket, cafeteria)	High external validity	Collection of control variables often not possible Adequate test store necessary

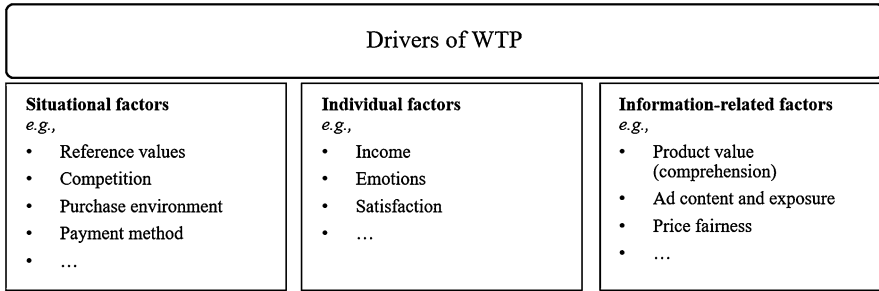


Fig. 4 Drivers of WTP

Situational Factors

Consumers usually form their WTP at the point of purchase. Therefore, WTP is highly susceptible to situational anchors, that is, prices or other numerical cues that are provided or remembered at the moment of purchase. These can be completely unrelated stimuli, e.g., prices of completely other products (Nunes and Boatwright 2004) or random figures such as someone’s social security number (Ariely et al. 2003). However, the effect is strongest when the respective anchor has a direct relation to the source of uncertainty (Simonson and Drolet 2004), thereby providing a reference for a possible price. These reference prices are very important, because without them, consumers may have difficulties translating product value into monetary units and articulate their WTP (Chernev 2003). Reference prices serve as standards when assessing product prices (Monroe 1973) and can be either remembered from past purchases or provided, e.g., by displaying recommended retail prices or the prices of competitor products (for a detailed review of literature on reference prices see Mazumdar et al. 2005). Without the possibility of comparing a product with other alternatives, consumers have difficulties evaluating it and often misjudge its value and their related WTP (Hsee 1998; Sevдалis and Harvey 2006).

Besides from serving as comparative values, the presence of competitor products lowers price tolerance, as switching to another attractive alternative is easy (Anderson 1996, p. 271). This means that market thickness generally lowers WTP (Chan et al. 2007); for example, the maturation of private labels has lowered consumers’ WTP for branded products (Steenkamp et al. 2010).

Further, purchase environment and place of purchase play a role, as the same consumer may have different WTP values for the same product, depending on the consumption location and occasion. For example, WTP for beverages increases strongly when they are purchased in a restaurant or discotheque instead of a supermarket. WTP might also depend on the circumstances, e.g., a social or nonsocial setting (Wakefield and Inman 2003), or the surroundings, e.g., store atmospherics (Borges et al. 2013; Fiore et al. 2000). Even subtle cues such as colors may impact WTP: for instance, experiments by Bagchi and Cheema (2013) show that red heightens aggressiveness, leading to higher WTP in auctions, but lower WTP in negotiations.

Furthermore, WTP depends on both where and how consumers purchase products: for instance, consumers' WTP may simply increase by paying with a credit card instead of paying with cash, sometimes to a very large extent (up to 100%) (Prelec and Simester 2001).

Individual Factors

WTP may also vary due to individual factors, such as consumer-specific attributes and emotions.

Consumers' price sensitivity is generally positively related to income (Degeratu et al. 2000, p. 64). Therefore, consumers' tolerance for higher prices and the implied higher WTP increases with rising income, whereas the price search decreases (Urbany et al. 1996). As there are high between-subject differences with regard to income, researchers should control for this variable when WTP is measured. As income is usually relatively stable over a certain amount of time, it is possible to account for different WTP across consumers by using price discrimination (e.g., to offer lower prices for students).

In contrast, emotions vary within consumers, with consumers' WTP changing from one moment to the next. Positive affect (i.e., positive feelings about owning a good) was found to increase WTP (Peters et al. 2003), while positive mood did not seem to have any effect (Capra et al. 2010). Negative mood, however, may impact consumers' WTP: if consumers experience a situation that makes them feel disgusted, they afterward display a reduced WTP, as the experienced disgust has triggered an "avoid taking anything in" goal. Sadness, however, increases WTP because sadness provokes the urge to change one's situation (Lerner et al. 2004).

However, there are also more stable types of emotions, such as general satisfaction with a product or seller. Anderson (1996) shows that consumer satisfaction is likely to decrease price elasticity, which means that satisfied consumers have a higher price tolerance before switching to competitor products. Similarly, Homburg et al. (2005) demonstrate a positive impact of satisfaction on WTP. The underlying relationship can be depicted as an inverse s-form, as disappointment (elation) leads to a strong decrease (increase) in WTP, whereas a mediocre level of satisfaction does not impact WTP.

Information-Related Factors

WTP may also vary depending on the information status. Ajzen and Driver (1992) explain that the WTP that is based on heuristics (i.e., rules of thumb based on few information) deviates from the WTP that is formed after thoroughly considering the value of a product.

Therefore, Smith and Nagle (2002) stress the importance of value comprehension and argue that consumers do not know their WTP until they are fully informed about the value of a product. Thus, due to high costs for information search and evaluation,

consumers often underestimate their WTP because they are not fully aware of all the benefits of a product.

Rao and Sieben (1992) support the importance of knowledge. By measuring product knowledge, they show that low-knowledge subjects have a significantly lower WTP than medium- or high-knowledge subjects. Similarly, consumers with higher involvement, i.e., knowledgeable subjects willing to search for information, show a lower price consciousness, and thus, a higher tolerance toward higher prices (Lichtenstein et al. 1988). However, these findings are context-dependent: Chan et al. (2007) show that experience and extensive search lower WTP in the context of online auctions. Here, expertise leads bidders to a more realistic estimation of achievable prices, thus making them more immune to overbidding.

Advertising also plays a decisive role in influencing consumers' WTP. Advertisements communicate information, such as a product's unique selling proposition, advantages, and value, thereby shaping consumers' product perception. Depending on the message delivered, WTP may differ; for example, Kaul and Wittink (1995) show that price advertising increases price sensitivity and therefore lowers WTP, whereas a rise in non-price advertising increases price tolerance. Kalra and Goodstein (1998) refine these results by demonstrating that advertising a minor brand using value positioning reduces WTP, whereas comparisons with premium brands or advertising a unique brand attribute increases it.

In addition to the content of an ad, the ad execution, such as its quality (Hampel et al. 2012) or the chosen medium (Li and Meshkova 2013), may further impact consumers' WTP, as well as the timing of the ad delivery. For example, ads that interrupt current consumer activities may significantly reduce consumers' WTP for the advertised product (Acquisti and Spiekermann 2011).

Finally, understanding the benefits of a product is not necessarily the only information consumers seek when determining their WTP. Consumers may be also concerned with price fairness: their WTP might decrease if they believe prices to be unfair toward consumers. In contrast, their WTP may increase if they feel that the seller does not make an adequate profit (Kahneman et al. 1986).

Market Research Application

In this section, we illustrate how to deal with the information on WTP in practice using some simple examples. Therefore, we collected data on consumers' WTP applying direct stated and direct revealed preference methods, in particular the open-ended question method, the dichotomous choice method, and the BDM mechanism. We then discuss three different application areas.

Elicitation of Consumers' WTP

We collected the data on WTP at a large German university campus. The focal product was described as a fresh, locally sourced salad that was to be offered in three

variants (Grilled Veggie, Avocado Superfood, and Caesar Chicken) in the students' cafeteria. The salad was supposed to be sold pre-packaged in standard off-the-shelf sizes, enabling students to not only eat directly at the cafeteria, but also to take the salad with them. We provided further information on the size of the product and the brand (*dean&david*, a restaurant chain specialized in healthy and locally sourced food).

We approached students asking them whether they would be willing to take part in a short study regarding a hypothetical new product offer at the cafeteria. We then explained the new offer as described above and showed pictures of the salads.

Depending on the method we either asked the students to state how much they would be willing to pay for such a salad in the cafeteria (open-ended questions method), to state if they accept or decline a given price (dichotomous choice method), or to make a bid pursuant to the rules of the BDM mechanism, which we explained to them in detail.

Measurement 1: Open-Ended Questions

Using this method, participants directly stated their WTP for the focal product. In sum, 35 students participated and stated prices ranging from 2€ to 5€ for the salad.

Measurement 2: Dichotomous Choice Method

In contrast to measurement 1, participants did not have to state a price, but to agree or to disagree on a given price. To obtain more precise information on their WTP, we further applied a sequential approach. Thus, depending on whether the participants agreed (disagreed) on the given price, they were then asked whether they would be willing to buy the product at a higher (lower) price. The average market price of comparable products in the market (comparable products in university cafeterias: between 3.00€ and 6.00€) was chosen as a starting price, here 4.50€. If participants rejected (accepted) the price, we lowered (increased) the price by 0.50€ until participants changed their mind. We then increased (decreased) the price by another 0.25€. Our chosen ending points equal the margins of the market prices (see Fig. 5 for the price pathways used).

In sum, we interviewed 33 students and collected prices between 3.00€ and 5.00€.

Measurement 3: BDM Mechanism

We conducted the BDM mechanism with 30 students of two marketing classes in the respective lecture rooms. At the end of the classes, we asked the students whether they would be willing to stay to participate in a small experiment that required the use of money.

We needed to provide further information on the procedure, as the mechanism is more complex than the previous methods. Besides giving the students information on the focal product, we explained them that they could actually buy the salads, or more specifically a voucher for the respective salads that could be redeemed at a nearby *dean&david* store. As the offer did not yet exist in reality, we operationalized the purchase by using vouchers instead.

The participants were told that no purchase price had been determined yet, but that one student may draw a price from an envelope containing different price tags. We

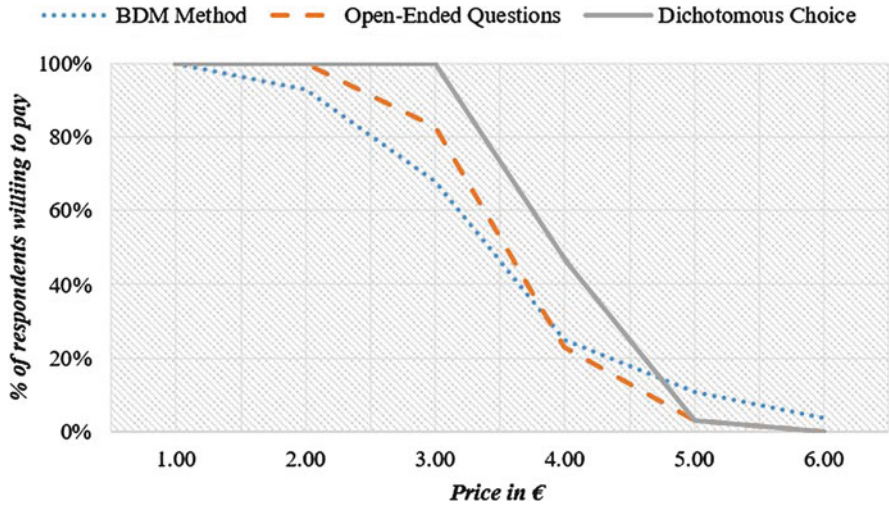


Fig. 5 Tree diagram of price pathways used

pointed out the consequences of bids above or below the price drawn. We then asked students to write down their WTP on an empty paper sheet that had been handed out in the beginning. After a student had drawn the threshold price, all participants disclosed their bids by handing out their sheets. We collected all bids and then sold vouchers to all students whose stated WTP exceeded the threshold price.

Results of the Measurements

A total of 98 students participated in our WTP measurements. Figure 6 shows the demand functions dependent on the elicitation method.

As we can see from Fig. 6, highest WTP are stated in the Dichotomous Choice group with mean WTP that equals 3.80€. In contrast, lowest WTP were elicited in the BDM group (mean WTP = 3.26€). Differences between lowest and highest group are significant ($p < 0.05$, using Satterthwaite approximation). These findings support previous research that respondents in hypothetical settings tend to overstate their WTP compared to the actual cash and incentive-compatible setting of the BDM mechanism, where we observe the lowest WTP. The average WTP is also higher in the Open-Ended Questions group (mean WTP = 3.39€) compared to the BDM group, however, the difference is not significant.

It is also salient that the demand functions are particularly steep between 3.00€ and 4.00€. While 83% of respondents in the Open-Ended Questions group were willing to pay 3.00€ for the salad, only 23% of respondents were willing to pay 4.00 €. More than half of the respondents who were willing to pay 3.00€ were not willing to pay 33% more. Thus, there seems to be a price threshold at 4.00€, which has to be considered when determining prices. Table 4 illustrates the decrease between these price points and the corresponding price elasticities of demand.

In the following we will show how managers can apply this information in practice based on some simple examples.

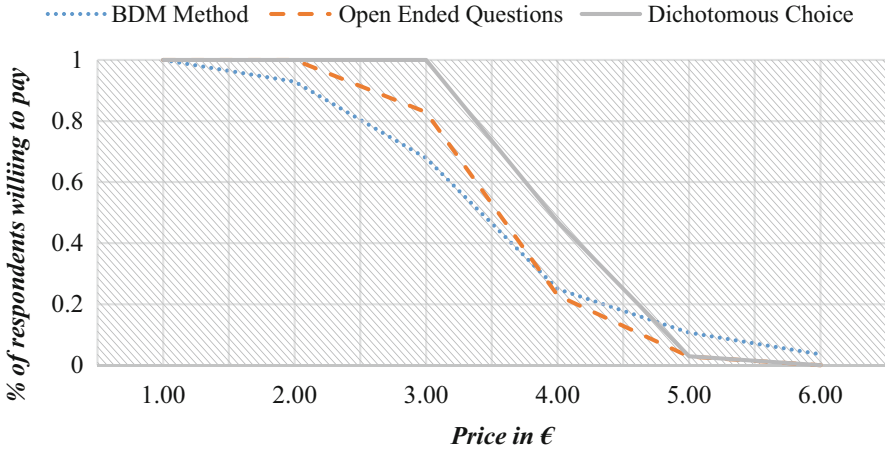


Fig. 6 Comparison of responses depending on method

Table 4 Decrease in WTP

	% of respondents willing to pay		Elasticity
	3.00 €	4.00 €	
BDM	68%	25%	-1.89
Open-ended questions	83%	23%	-2.17
Dichotomous choice	97%	47%	-1.54

Application 1: Price Bundling

Now let us consider that the university cafeteria is indeed interested in expanding the menu by offering the three variants of fresh salads bundled (with a new organic soft drink) and/or unbundled. Suppose that the university cafeteria has additionally identified different segments, which are all of the same size (15,000 per segment) and has collected WTP data on the salad, the new organic soft drink, and the bundle of the two products. Table 5 shows average WTP for the two products in the three different segments.

Let us also assume for simplicity that marginal costs equal zero. Then the cafeteria has three pricing alternatives to choose from. They can apply a uniform pricing strategy, i.e., charge only one fixed price for the salad and for the soft drink, or they can use a pure bundling strategy, i.e., sell the two products in a bundle only. The third alternative is to apply a mixed bundling strategy and to offer the products both separately and as a bundle.

Now let us first suppose that the cafeteria applies the uniform pricing strategy. Then the optimal price for the salad is 3.90€ and for the soft drink is 2.00€, thus

Table 5 Average WTP for unbundled and bundled products

	Average WTP in €		
	Salad	Soft drink	Bundle
Segment 1	4.50	0.50	5.00
Segment 2	3.90	2.00	5.90
Segment 3	1.70	2.50	4.20

resulting in a total revenue of 177,000€. In the second case, the pure bundling strategy, the optimal price for the bundle is 4.20€, resulting in a total revenue of 189,000€. In the last case, the mixed bundling strategy, optimal price for the salad is 4.50€, 2.50€ for the soft drink and 5.90€ for the bundle. This strategy results in a revenue of 193,500€. Consequently, the optimal strategy for the cafeteria is to apply the mixed bundling strategy.

Please note that this a simplified representation of a price bundling case. In practice, managers usually have to deal with more segments and products, and have to consider additional costs, thus making the analysis far more complex.

Application 2: Personalized Pricing

Suppose the cafeteria has additionally collected WTP data for the salad from their university staff as well as external cafeteria visitors who are neither students nor are affiliated with the university. Table 6 shows the elicited WTP and the segment size.

Assume that marginal costs are zero. The cafeteria can now either apply a uniform pricing strategy or charge personalized prices. Using the uniform pricing strategy, the optimal price is 3.50€, resulting in a revenue of 199,500€. However, applying personalized pricing, thus discriminating between the three segments, results in a revenue of 210,700€.

Since personalized pricing is commonly used in this type of service sector, it is easily feasible in this context.

Application 3: Nonlinear Pricing

Now suppose that the student cafeteria on another campus has observed that some students consume more than one salad and thus is interested to give price discounts for a higher amount of purchases, i.e., the second salad is cheaper than the first salad when purchasing two. Assume the following WTP measurements of three different consumer segments which are all of the same size ($n = 3000$) for different amounts of purchases (see Table 7).

We again assume that marginal costs are zero. Using a uniform pricing strategy without discounting additional purchases, the optimal price would be 2.40€ leading to a total revenue of 50,400€. Considering the changes in WTP with a higher amount

Table 6 WTP for different segments

	Students	Staff	External
Average WTP in €	3.50	4.40	4.60
Segment size	45,000	10,000	2,000

Table 7 Average WTP for different purchase amounts

	Average WTP in €		
	First salad	Second salad	Third salad
Segment 1	4.00	3.50	3.00
Segment 2	3.50	3.10	2.40
Segment 3	2.50	2.00	1.70

of purchases, the cafeteria can also determine the prices sequentially, an approach referred to as “price-point” method (see Dolan and Simon 1996). Following the sequential approach, the optimal price for one salad is 2.50€, the optimal price for the second salad is 2.00€, and for the third salad is 1.70€, resulting in a revenue of 55,800€ (22,500€ + 18,000€ + 15,300€). In this case, the cafeteria yields a higher revenue when using the price-point method compared to the uniform pricing strategy. Using the price-point method, the cafeteria can sell a higher amount of products: all three segments will purchase three salads resulting in 27,000 units sold, whereas in the uniform pricing strategy the cafeteria will only sell 21,000 units.

Also note here that this example is a simplified representation with limited hypothetical data and without considering costs. Managers seeking to measure WTP for product bundles or for different amounts of purchases can, for example, use self-stated data or choice-based data where the design of the offer (bundled/unbundled) or the amount of products offered is varied.

Conclusion

Accurately measuring consumers’ WTP is of great importance for pricing decisions and predicting sales. Our chapter gives an overview of common methods for measuring WTP that are widely used in both theory and practice. We discuss the advantages and limitations associated with each method. Stated preference approaches (direct and indirect survey approaches) are advantageous in terms of feasibility but suffer from hypothetical bias and strategic behavior, whereas revealed preference methods may overcome the hypothetical bias, but usually involve more effort (financial and organizational).

We further point to the possible factors influencing WTP and differentiate situational factors, individual factors, and information-related factors. Depending on the situation, on individual consumer-specific attributes and emotions, and on the information status, consumers’ WTP may vary. Firms being aware of the drivers of WTP can take advantage of this information and react accordingly.

In the remainder of our chapter, we present a practical application of consumers' WTP measurement for a potential new product offer. We measure WTP via two direct stated preference approaches (open-ended question and dichotomous choice method employing the sequential monadic approach) and one direct revealed preference approach (BDM mechanism). We compare the three different approaches and discuss application areas of WTP using three examples.

References

- Acquisti, A., & Spiekermann, S. (2011). Do interruptions pay off? Effects of interruptive ads on consumers' willingness to pay. *Journal of Interactive Marketing, 25*, 226–240. <https://doi.org/10.1016/j.intmar.2011.04.003>.
- Ajzen, I., & Driver, B. L. (1992). Contingent value measurement: On the nature and meaning of willingness to pay. *Journal of Consumer Psychology, 1*, 297–316. [https://doi.org/10.1016/S1057-7408\(08\)80057-5](https://doi.org/10.1016/S1057-7408(08)80057-5).
- Allenby, G. M., Brazell, J. D., Howell, J. R., & Rossi, P. E. (2014). Economic valuation of product features. *Quantitative Marketing and Economics, 12*, 421–456.
- Anderson, E. W. (1996). Customer satisfaction and price tolerance. *Marketing Letters, 7*, 265–274. <https://doi.org/10.1007/BF00435742>.
- Ariely, D., Loewenstein, G., & Prelec, D. (2003). “Coherent arbitrariness”: Stable demand curves without stable preferences. *The Quarterly Journal of Economics, 118*, 73–106. <https://doi.org/10.1162/00335530360535153>.
- Bagchi, R., & Cheema, A. (2013). The effect of red background color on willingness-to-pay: The moderating role of selling mechanism. *Journal of Consumer Research, 39*, 947–960. <https://doi.org/10.1086/666466>.
- Barrot, C., Albers, S., Skiera, B., & Schäfers, B. (2010). Vickrey vs. eBay: Why second-price sealed-bid auctions Lead to more realistic price-demand functions. *International Journal of Electronic Commerce, 14*, 7–38.
- Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science, 9*, 226–232.
- Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., et al. (1994). Combining revealed and stated preferences data. *Marketing Letters, 5*, 335–349.
- Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of Consumer Research, 25*, 187–217. <https://doi.org/10.1086/209535>.
- Blumenschein, K., Blomquist, G. C., Johannesson, M., Horn, N., & Freeman, P. (2008). Eliciting willingness to pay without bias: Evidence from a field experiment. *The Economic Journal, 118*, 114–137. <https://doi.org/10.1111/j.1468-0297.2007.02106.x>.
- Borges, A., Babin, B. J., & Spielmann, N. (2013). Gender orientation and retail atmosphere: Effects on value perception. *International Journal of Retail & Distribution Management, 41*, 498–511.
- Boyle, K. J. (2017). Contingent valuation in practice. In *A primer on nonmarket valuation* (pp. 83–131). Dordrecht: Springer.
- Breidert, C. (2006). *Estimation of willingness-to-pay: Theory, measurement, application*. Wiesbaden: DUV Deutscher Universitäts-Verlag.
- Calabuig, F., Núñez-Pomar, J., Prado-Gascó, V., & Añó, V. (2014). Effect of price increases on future intentions of sport consumers. *Journal of Business Research, 67*, 729–733.
- Capra, C. M., Lanier, K. F., & Meer, S. (2010). The effects of induced mood on bidding in random nth-Price auctions. *Journal of Economic Behavior & Organization, 75*, 223–234. <https://doi.org/10.1016/j.jebo.2010.04.002>.
- Carare, O., & Rothkopf, M. (2005). Slow dutch auctions. *Management Science, 51*, 365–373. <https://doi.org/10.1287/mnsc.1040.0328>.

- Chan, T. Y., Kadiyali, V., & Park, Y.-H. (2007). Willingness to pay and competition in online auctions. *Journal of Marketing Research*, 44, 324–333. <https://doi.org/10.1509/jmkr.44.2.324>.
- Chernev, A. (2003). Reverse pricing and online price elicitation strategies in consumer choice. *Journal of Consumer Psychology*, 13, 51–62.
- Chung, J., & Rao, V. R. (2003). A general choice model for bundles with multiple-category products: Application to market segmentation and optimal pricing for bundles. *Journal of Marketing Research*, 40, 115–130.
- Cummings, R. G., & Taylor, L. O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *The American Economic Review*, 89, 649–665.
- Degeratu, A. M., Rangaswamy, A., & Wu, J. (2000). Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *International Journal of Research in Marketing*, 17, 55–78. [https://doi.org/10.1016/S0167-8116\(00\)00005-7](https://doi.org/10.1016/S0167-8116(00)00005-7).
- Dhar, R., & Gorlin, M. (2013). A dual-system framework to understand preference construction processes in choice. *Journal of Consumer Psychology*, 23, 528–542.
- Ding, M. (2007). An incentive-aligned mechanism for conjoint analysis. *Journal of Marketing Research*, 44, 214–223.
- Ding, M., Grewal, R., & Liechty, J. (2005). Incentive-aligned conjoint analysis. *Journal of Marketing Research*, 42, 67–82. <https://doi.org/10.1509/jmkr.42.1.67.56890>.
- Dolan, R. J., & Simon, H. (1996). *Power pricing*. New York: The Free Press.
- Donaldson, C. (1999). Valuing the benefits of publicly-provided health care: Does ‘ability to pay’ preclude the use of ‘willingness to pay’? *Social Science & Medicine*, 49, 551–563. [https://doi.org/10.1016/S0277-9536\(99\)00173-2](https://doi.org/10.1016/S0277-9536(99)00173-2).
- Dong, S., Ding, M., & Huber, J. (2010). A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing*, 27, 25–32.
- Dost, F., & Wilken, R. (2012). Measuring willingness to pay as a range, revisited: When should we care? *International Journal of Research in Marketing*, 29, 148–166. <https://doi.org/10.1016/j.ijresmar.2011.09.003>.
- Fiore, A. M., Yah, X., & Yoh, E. (2000). Effects of a product display and environmental fragrancing on approach responses and pleasurable experiences. *Psychology and Marketing*, 17, 27–54.
- Frederick, S., Novemsky, N., Wang, J., Dhar, R., & Nowlis, S. (2009). Opportunity cost neglect. *Journal of Consumer Research*, 36, 553–561.
- Gabor, A., & Granger, C. W. J. (1966). Price as an Indicator of quality: Report on an enquiry. *Economica*, 33, 43. <https://doi.org/10.2307/2552272>.
- Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, 8, 355–363.
- Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5, 103–123.
- Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54, 3–19. <https://doi.org/10.2307/1251756>.
- Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31, 56–73. <https://doi.org/10.1287/inte.31.3s.56.9676>.
- Greenberg, A. (2008). When apple failed. https://www.forbes.com/2008/10/29/apple-product-flops-tech-personal-cx_ag_1030apple, Accessed 18 June 2018
- Gustafsson, A. (Ed.). (2007). *Conjoint measurement: Methods and applications* (4th ed.). Berlin u. a.: Springer.
- Halme, M., & Kallio, M. (2011). Estimation methods for choice-based conjoint analysis of consumer preferences. *European Journal of Operational Research*, 214, 160–167. <https://doi.org/10.1016/j.ejor.2011.03.049>.
- Hamilton, R. W., & Srivastava, J. (2008). When 2+2 is not the same as 1+3: Variations in price sensitivity across components of partitioned prices. *Journal of Marketing Research*, 45, 450–461.

- Hampel, S., Heinrich, D., & Campbell, C. (2012). Is an advertisement worth the paper it's printed on? *Journal of Advertising Research*, 52, 118–127. <https://doi.org/10.2501/JAR-52-1-118-127>.
- Hanna, N., & Dodge, R. (1995). *Pricing – Policies and procedures*. London: Macmillan.
- Harrison, G. W., & Rutström, E. E. (2013). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott & V. L. Smith (Eds.), *Handbook of experimental economics results* (Vol. 1, pp. 752–767). Burlington: Elsevier.
- Harrison, G. W., Harstad, R. M., & Rutström, E. E. (2004). Experimental methods and elicitation of values. *Experimental Economics*, 7, 123–140.
- Hauser, J. R., & Urban, G. L. (1986). The value priority hypotheses for consumer budget plans. *Journal of Consumer Research*, 12, 446–462.
- Henderson, N. R. (2002). Avoiding the pricing trap in qualitative interviews. *Marketing Research*, 14, 38–39.
- Herriges, J. A., & Shogren, J. F. (1996). Starting point Bias in dichotomous choice valuation with follow-up questioning. *Journal of Environmental Economics and Management*, 30, 112–131. <https://doi.org/10.1006/jeem.1996.0008>.
- Hoeffler, S., & Ariely, D. (1999). Constructing stable preferences: A look into dimensions of experience and their impact on preference stability. *Journal of Consumer Psychology*, 8, 113–139. https://doi.org/10.1207/s15327663jcp0802_01.
- Hoffman, E., Menckhaus, D. J., Chakravarti, D., Field, R. A., & Whipple, G. D. (1993). Using laboratory experimental auctions in marketing research: A case study of new packaging for fresh beef. *Marketing Science*, 12, 318–338.
- Hofstetter, R., Miller, K. M., Krohmer, H., & Zhang, Z. J. (2013). How do consumer characteristics affect the bias in measuring willingness to pay for innovative products? *Journal of Product Innovation Management*, 30, 1042–1053. <https://doi.org/10.1111/jpim.12040>.
- Homburg, C., Koschate, N., & Hoyer, W. D. (2005). Do satisfied customers really pay more?: A study of the relationship between customer satisfaction and willingness to pay. *Journal of Marketing*, 69, 84–96. <https://doi.org/10.1509/jmkg.69.2.84.60760>.
- Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, 11, 107–121.
- Jedidi, K., & Jagpal, S. (2009). Willingness to pay: Measurement and managerial implications. In *Handbook of pricing research in marketing* (pp. 37–60).
- Jedidi, K., & Zhang, Z. J. (2002). Augmenting conjoint analysis to estimate consumer reservation price. *Management Science*, 48, 1350–1368.
- Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, 76, 728–741.
- Kalish, S., & Nelson, P. (1991). A comparison of ranking, rating and reservation price measurement in conjoint analysis. *Marketing Letters*, 2, 327–335. <https://doi.org/10.1007/BF00664219>.
- Kalra, A., & Goodstein, R. C. (1998). The impact of advertising positioning strategies on consumer Price sensitivity. *Journal of Marketing Research*, 35, 210. <https://doi.org/10.2307/3151849>.
- Kalwani, M. U., & Silk, A. J. (1982). On the reliability and predictive validity of purchase intention measures. *Marketing Science*, 1, 243–286.
- Kaul, A., & Wittink, D. R. (1995). Empirical generalizations about the impact of advertising on Price sensitivity and price. *Marketing Science*, 14, G151–G160. <https://doi.org/10.1287/mksc.14.3.G151>.
- Kohli, R., & Mahajan, V. (1991). A reservation-price model for optimal pricing of multiattribute products in conjoint analysis. *Journal of Marketing Research*, 28, 347–354.
- Ku, G., Malhotra, D., & Murnighan, J. K. (2005). Towards a competitive arousal model of decision-making: A study of auction fever in live and internet auctions. *Organizational Behavior and Human Decision Processes*, 96, 89–103. <https://doi.org/10.1016/j.obhdp.2004.10.001>.
- Lerner, J. S., Small, D. A., & Loewenstein, G. (2004). Heart strings and purse strings carryover effects of emotions on economic decisions. *Psychological Science*, 15, 337–341.

- Li, T., & Meshkova, Z. (2013). Examining the impact of rich media on consumer willingness to pay in online stores. *Electronic Commerce Research and Applications*, 12, 449–461. <https://doi.org/10.1016/j.elerap.2013.07.001>.
- Lichtenstein, D. R., Bloch, P. H., & Black, W. C. (1988). Correlates of price acceptability. *Journal of Consumer Research*, 15, 243–252. <https://doi.org/10.1086/209161>.
- Lieven, T., & Lennerts, S. (2013). Measuring willingness to pay by means of the trade-off between free available cash and specific-purpose vouchers. *Business Research*, 6, 154–171. <https://doi.org/10.1007/BF03342747>.
- Loomis, J., Gonzalez-Caban, A., & Gregory, R. (1994). Do reminders of substitutes and budget constraints influence contingent valuation estimates? *Land Economics*, 70, 499. <https://doi.org/10.2307/3146643>.
- Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, 20, 350–367.
- Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and applications*. Cambridge: Cambridge University Press.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- Lyon, D. W. (2002). The price is right (or is it)? Accurate pricing starts with asking the right questions. *Marketing Research*, 14, 8–13.
- Marn, M. V., Roegner, E. V., & Zawada, C. C. (2003). Pricing new products. *McKinsey Quarterly*, 3, 40–49.
- Mazumdar, T., Raj, S. P., & Sinha, I. (2005). Reference price research: Review and propositions. *Journal of Marketing*, 69, 84–102. <https://doi.org/10.1509/jmkg.2005.69.4.84>.
- Miller, K. M., Hofstetter, R., Krohmer, H., & Zhang, Z. J. (2011). How should consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches. *Journal of Marketing Research*, 48, 172–184. <https://doi.org/10.1509/jmkr.48.1.172>.
- Mitchell, R. C. (2013). *Using surveys to value public goods: The contingent valuation method*. Hoboken: Taylor and Francis.
- Monroe, K. B. (1973). Buyers' subjective perceptions of price. *Journal of Marketing Research*, 10, 70. <https://doi.org/10.2307/3149411>.
- Moorthy, K. S. (1988). Product and price competition in a duopoly. *Marketing Science*, 7, 141–168. <https://doi.org/10.1287/mksc.7.2.141>.
- Morrison, D. G. (1979). Purchase intentions and purchase behavior. *The Journal of Marketing*, 43, 65–74.
- Müller, H. (2009). Empirische untersuchung zur messung der preiswahrnehmung mittels pricesensitivity-meter. *Marketing ZfP*, 31, 171–182.
- Neill, H. R., Cummings, R. G., Ganderton, P. T., Harrison, G. W., & McGuckin, T. (1994). Hypothetical surveys and real economic commitments. *Land Economics*, 70, 145–154.
- Noussair, C., Robin, S., & Ruffieux, B. (2004). Revealing consumers' willingness-to-pay: A comparison of the BDM mechanism and the vickrey auction. *Journal of Economic Psychology*, 25, 725–741. <https://doi.org/10.1016/j.joep.2003.06.004>.
- Nunes, J. C., & Boatwright, P. (2004). Incidental prices and their effect on willingness to pay. *Journal of Marketing Research*, 41, 457–466.
- Peters, E., Slovic, P., & Gregory, R. (2003). The role of affect in the WTA/WTP disparity. *Journal of Behavioral Decision Making*, 16, 309–330. <https://doi.org/10.1002/bdm.448>.
- Prelec, D., & Simester, D. (2001). Always leave home without it: A further investigation of the credit-card effect on willingness to pay. *Marketing Letters*, 12, 5–12.
- Ramanujam, M., & Tacke, G. (2016). *Monetizing innovation: How smart companies design the product around the Price* (1st ed.). Hoboken: Wiley.
- Rao, A. R., & Sieben, W. A. (1992). The effect of prior knowledge on Price acceptability and the type of information examined. *Journal of Consumer Research*, 19, 256. <https://doi.org/10.1086/209300>.

- Rowe, R. D., Schulze, W. D., & Breffle, W. S. (1996). A test for payment card biases. *Journal of Environmental Economics and Management*, 31, 178–185.
- Russell, S. (1996). Ability to pay for health care: Concepts and evidence. *Health Policy and Planning*, 11, 219–237.
- Sen, S., Gürhan-Canli, Z., & Morwitz, V. (2001). Withholding consumption: A social dilemma perspective on consumer boycotts. *Journal of Consumer Research*, 28, 399–417. <https://doi.org/10.1086/323729>.
- Sevdalis, N., & Harvey, N. (2006). Determinants of willingness to pay in separate and joint evaluations of options: Context matters. *Journal of Economic Psychology*, 27, 377–385. <https://doi.org/10.1016/j.joep.2005.07.001>.
- Shaffer, G., & Zhang, Z. J. (1995). Competitive coupon targeting. *Marketing Science*, 14, 395–416.
- Shogren, J. F., Margolis, M., Koo, C., & List, J. A. (2001). A random nth-price auction. *Journal of Economic Behavior & Organization*, 46, 409–421.
- Simon, H., & Fassnacht, M. (1982). *Preismanagement*. Wiesbaden: Springer.
- Simonson, I., & Drolet, A. (2004). Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of Consumer Research*, 31, 681–690. <https://doi.org/10.1086/425103>.
- Slovic, P. (1995). The construction of preference. *American Psychologist*, 50, 364–371.
- Smith, G. E., & Nagle, T. T. (2002). How much are customers willing to pay? *Marketing Research*, 14, 20.
- Spann, M., Skiera, B., & Schäfers, B. (2004). Measuring individual frictional costs and willingness-to-pay via name-your-own-price mechanisms. *Journal of Interactive Marketing*, 18(4), 22–36.
- Srinivasan, V. (1988). A conjunctive-compensatory approach to the self-explication of multi-attributed preferences. *Decision Sciences*, 19, 295–305.
- Steenkamp, J.-B. E. M., van Heerde, H. J., & Geyskens, I. (2010). What makes consumers willing to pay a price premium for national brands over private labels? *Journal of Marketing Research*, 47, 1011–1024. <https://doi.org/10.1509/jmkr.47.6.1011>.
- Terwiesch, C., Savin, S., & Hann, I.-H. (2005). Online haggling at a name-your-own-Price retailer: Theory and application. *Management Science*, 51, 339–351. <https://doi.org/10.1287/mnsc.1040.0337>.
- Urbany, J. E., Dickson, P. R., & Kalapurakal, R. (1996). Price search in the retail grocery market. *Journal of Marketing*, 60, 91. <https://doi.org/10.2307/1251933>.
- van Westendorp, P. H. (1976). NSS price sensitivity meter (PSM) – a new approach to study consumer-perception of prices. In *Proceedings of the ESOMAR congress*.
- Varian, H. R. (1992). *Microeconomic analysis*. New York: Norton.
- Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16, 8–37.
- Voelckner, F. (2006). An empirical comparison of methods for measuring consumers' willingness to pay. *Marketing Letters*, 17, 137–149.
- Wakefield, K. L., & Inman, J. J. (2003). Situational price sensitivity: The role of consumption occasion, social context and income. *Journal of Retailing*, 79, 199–212. <https://doi.org/10.1016/j.jretai.2003.09.004>.
- Wang, T., Venkatesh, R., & Chatterjee, R. (2007). Reservation price as a range: An incentive-compatible measurement approach. *Journal of Marketing Research*, 44, 200–213. <https://doi.org/10.1509/jmkr.44.2.200>.
- Wertenbroch, K., & Skiera, B. (2002). Measuring consumers' willingness to pay at the point of purchase. *Journal of Marketing Research*, 39, 228–241.
- Winer, R. S. (2005). *Pricing*. Cambridge, MA: Marketing Science Institute.



Modeling Customer Lifetime Value, Retention, and Churn

Herbert Castéran, Lars Meyer-Waarden, and Werner Reinartz

Contents

Introduction	1002
A Taxonomy of Customer Lifetime Value Measurement Models	1003
Retention Models for CLV Measurement	1006
Migration Models for CLV Measurement	1010
Continuous Mixed Models: The Family of NBD Models to Measure CLV.....	1012
An Application of Stochastic Pareto/NBD and BG/NBD Models for Customer Base Management	1016
Data and Methodology	1016
Estimation	1018
Results	1023
Parameter Estimations	1023
Purchase Prediction Validity	1025
Conclusion	1029
References	1030

Abstract

Customers represent the most important assets of a firm. Customer lifetime value (CLV) allows assessing their current and future value in a customer base. The customer relationship management strategy and marketing resource allocation are

H. Castéran (✉)
Humanis Institute, EM Strasbourg Business School, Strasbourg, France
e-mail: herbert.casteran@em-strasbourg.eu

L. Meyer-Waarden
School of Management, CRM CNRS University Toulouse 1 Capitole, IAE Toulouse, Toulouse, France
e-mail: lars.meyer-waarden@iae-toulouse.fr

W. Reinartz
University of Cologne, Köln, Germany
e-mail: werner.reinartz@uni-koeln.de

based on this metric. Managers therefore need to predict the retention but also the purchase behavior of their customers.

This chapter is a systematic review of the most common CLV, retention, and churn modeling approaches for customer-base analysis and gives practical recommendations for their applications. These comprise both the classes of deterministic and stochastic approaches and deal with both, contractual and noncontractual settings. Across those situations, the most common and most important approaches are then systematically structured, described, and evaluated. To this end, a review of the CLV, retention, as well as churn models and a taxonomy are done with their assumptions and weaknesses. Next, an empirical application of the stochastic “standard” Pareto/NBD, and the BG/NBD models, as well as an explanatory Pareto/NBD model with covariates to grocery retailing store loyalty program scanner data, is done. The models show their ability to reproduce the interindividual variations as well as forecasting validity.

Keywords

Customer lifetime value · Customer churn · Customer retention · NBD/Pareto model · BG/NBD model

Introduction

Customers represent assets and the cost of acquiring them relates to the cash flow they are expected to generate over time (Bolton et al. 2004; Tarasi et al. 2011). Customer retention and churn as well as customer lifetime value (CLV), retention, and churn measurement have become a powerful customer valuation metric (Gladly et al. 2015; Gupta et al. 2004, 2006; Kumar and Reinartz 2006).

Customer retention refers to the ability of a company or product to retain its customers over some specified period. Defection or churn is the number of customers moving out of a cohort in a firm’s database over a specific period of time. CLV is the value of individual customers, based on their past, present, and projected future cash flows (Gupta et al. 2004). To model CLV, it is important to measure customer retention and churn rates. CLV is an important concept on which the customer relationship management strategy; marketing resource allocation (to profitable customers), such as promotions; and the assessment of the marketing efficiency are based on Schulze et al. (2012). The CLV paradigm recognizes customers as the primary source of both current and future cash flows. According to this framework, the firm tries to maximize the net present value of both current and future customers (customer equity, Hogan et al. 2002), which represents a good proxy for the firm’s value (Borle et al. 2008; Gupta et al. 2004), as well as an effective segmentation tool. Thus, CLV models offer a powerful means to maximize the return on marketing investments and guide allocations of the marketing budget (Blattberg and Deighton 1996; Reinartz et al. 2005).

A CLV model has prototypically three parameters: (1) margin (purchase baskets minus the costs including retention expenditure), (2) retention probability or lifetime

duration, and (3) purchase frequency (Kumar 2007). One way of increasing CLV is to undertake marketing initiatives to reduce churn or the defection rate (and therefore increase the retention rate) of customers – which will have the impact of increase in the customer lifetime periods. Putting it another way, CLV analyses involve distinguishing active customers from defectors and then predicting their lifetime and future levels of transactions according to their observed past purchase behavior. Developing a valid measurement framework that adequately describes the process of birth, purchase activity, and defection is thus a crucial albeit not a trivial task, particularly due to the randomness of individual purchasing behavior and customer heterogeneity (Jain and Singh 2002; Reinartz and Kumar 2000). Whereas the analysis may be easier for contractual, “lost for good” relationships (e.g., subscription markets in which the inactivity date is known), it becomes particularly difficult for noncontractual relationships in which customers do not notify the firm when they disappear (Dwyer 1989; Jackson 1985); in this scenario, identifying active and inactive customers in the database at any given time requires systematic investigation (Schmittlein and Peterson 1994).

The objective of this chapter is to provide a systematic review of the most common retention and churn modeling approaches to model CLV. These comprise both the classes of deterministic and stochastic approaches and deal with both, always-a-share and lost-for-good situations. Across those situations, the most common and most important approaches are then systematically structured, described, and evaluated. To this end, first the retention models, their assumptions, and weaknesses are reviewed and thus a taxonomy is provided. After having presented the taxonomy of CLV, churn, and retention measurement models, this article shows in “[A Taxonomy of Customer Lifetime Value Measurement Models](#)” a practical application by using some of the presented stochastic models (Pareto/NBD, explanatory Pareto/NBD, BG/NBD) to model a customer base and the impact of a retail grocery loyalty program on customer churn, retention, and activity. The goal is to show how to implement, use, and interpret these sophisticated models by applying them on firms’ frequently used grocery store loyalty program databases and panel data.

This article then concludes with a discussion, some limitations, and recommendations for future research directions.

A Taxonomy of Customer Lifetime Value Measurement Models

In practice, to choose an adequate CLV measurement model, one has to understand whether or not customer defection is observable. One thus has to differentiate between two types of market (Jackson 1985; Dwyer 1989), namely, contractual (*Lost-for-Good*) and non-contractual (*always-a-share*) markets.

In the first type of market, the customer enters into a contractual relationship with a firm (e.g., phone or insurance services, magazine subscriptions, etc.) and is consequently faced with a tangible cost of change. Defection is observable and occurs when consumers end their relationship with the firm. In this scenario, the seller can identify defection as soon as it occurs. This means it is easy to predict defection for modeling purposes and one has to adopt a simple retention model

(Ayache et al. 2006). The notion of *lost-for-good* merges in practice with the contractual situation since it considers that absence of transaction means the customer has become inactive. In the contractual approach, retention is in fact the most important aspect. Generally, it goes hand in hand with a more or less constant flow of income. The models are usually simple with a clear predominance of survival models.

In markets where the customer has no contractual relationship (typically consumer goods), the cost of switching is low and a buyer can simultaneously purchase from different suppliers (*always-a-share*). The supplier has no way of knowing if the customer has defected. The model therefore focuses on churn probability, customer migration, and customer “life span” (Berger and Nasr 1998). The longer the period of inactivity, the more likely it is that the customer has churned. Migration models more specifically cover this scenario.

Fader and Hardie (2009) add an additional distinction depending on whether the purchase occurs at a specific moment (discrete time) or whether it can occur at any time (continuous). This distinction mainly has technical consequences, which can also be computed more or less approximately in a relatively direct way by taking more or less extended periods of time into consideration. Fader and Hardie (2009) themselves admit that this distinction is less meaningful. However, the contractual/noncontractual distinction is conceptually and methodologically fundamental.

The vast majority of markets concern noncontractual markets (Allenby et al. 1999). Many researchers and business practitioners have attempted to develop forecasting systems in this context. Contributions fall into two main categories: purely descriptive approaches (deterministic) and stochastic approaches. Deterministic approaches are primarily based on calculations of actuarial values, reflecting financial flows without the inclusion of random factors or explanatory variables (e.g., expected individual cash flow models as applied by Berger and Nasr 1998). However, they fail to take interindividual heterogeneity into account. Calciu and Salerno (2002) highlighted the relations between these different attempts.

The following table provides an overall view of the models according to the nature of their affiliation with the company and the methodology (deterministic/stochastic) used. Some contributions may be found in two different scenarios, in that they include a comparison of several cases.

Other aspects of model characterization are also included: level of aggregation, inclusion of the competition, return on investment, and the capacity to optimize resource allocation. The nature of the model and the level of aggregation help to determine the model’s sophistication and precision. Taking the competition into account is likely to affect the results of the models in that the long-term perspective is more complex when the competitive context is explicitly included. Finally, the capacity to determine return on investment or to optimize the distribution of marketing investment affects the model’s operational nature.

Table 1 suggests several trends. The first is the increasing focus on stochastic models as compared to the deterministic models. Since 2005, eight new stochastic models have been presented against only two in the deterministic context. As already stated, probabilistic models are significantly more efficient than deterministic models. This tendency thus seems logical and desirable.

Table 1 Models of customer retention-churn modeling (Adapted from Villanueva and Hanssens 2007)

Authors	Level of analysis	Competition present	Return on investment	Allocation of resources
Deterministic models				
No application				
Rust et al. (2004)	Company	Yes	Yes	Yes
Blattberg et al. (2001)	Segment	No	Yes	No
Application to contractual cases				
Keane and Wang (1995)	Regions	No	No	No
Blattberg and Deighton (1996)	Company	No	Yes	No
Dwyer (1997)	Segment	No	No	No
Ryals (2005)	Individual	No	No	No
Wiesel et al. (2008)	Company	No	No	No
Application to noncontractual cases				
Dwyer (1997)	Segment	No	No	No
Berger and Nasr (1998)	Individual	No	No	No
Stauss and Friege (1999)	Individual	No	No	No
Berger and Nasr (1998)	Company	No	No	No
Gupta et al. (2002)	Company	No	No	No
Gupta and Lehman (2003)	Company	No	No	No
Stochastic models				
Application to contractual cases				
Bitran and Mondschein (1996)	Segment	No	No	No
Thomas et al. (2004)	Individual	No	Yes	No
Lewis (2005)	Individual	No	Yes	No
Villanueva et al. (2008)	Company	No	No	No
Application to noncontractual cases				
Schmittlein et al. (1987)	Individual	No	No	No
Reinartz and Kumar (2000)	Consumer	No	Yes	No
Pfeifer and Carraway (2000)	Segment	No	Yes	No
Libai et al. (2002)	Segment	No	Yes	Yes
Rust et al. (2004)	Company	Yes	Yes	Yes
Venkatesan and Kumar (2004)	Individual	No	Yes	Yes
Fader et al. (2005a)	Individual	No	No	No

(continued)

Table 1 (continued)

Authors	Level of analysis	Competition present	Return on investment	Allocation of resources
Reinartz et al. (2005)	Company	Yes	Yes	No
Villanueva et al. (2008)	Segment	No	Non	No
Simester et al. (2006)	Individual	No	Yes	No
Lewis (2006)	Individual	No	Yes	No
Castéran et al. (2007a, b)	Individual	No	No	No

The second underlying trend involves the increasing disaggregation of the models. From wholly aggregated models, one shifts to an analysis by company, then to one by segment, and, finally and increasingly often, to one by individual. While informational limitations may explain the inclusion of a company level, nothing, on the other hand, justifies grounding a marketing analysis on wholly aggregated models.

Two aspects have been relatively neglected to date, namely, inclusion of the competition and the way managers interpret models for resource allocation. Lack of information is frequently used to explain this shortcoming, but it nonetheless remains detrimental. This is especially true of the failure to include the competition insofar as its absence may substantially impact on conclusions and managerial implications (cf. Fudenberg and Tirole 2000). While the examination of optimized resource allocation remains fundamental, its absence does not, on the other hand, imply an analysis bias.

We present these approaches in more detail in a dual customer relations and methodology framework.

Retention Models for CLV Measurement

These models are divided between deterministic and probabilistic models. To determine CLV, customer retention and churn have to be modeled. Customer retention refers to the ability of a company or product to retain its customers over some specified period. It is measured in the following way (Gupta et al. 2004).

$$\text{Retention rate} = \frac{n \text{ customers in cohort buying in } (t)}{n \text{ customers in cohort buying in } (t-1)} \times 100 \quad (1)$$

The period t can refer to specific durations: months or years are the most frequently used. Customer defection or churn is the number of customers moving out of a cohort in a firm's database over a specific period of time. It is measured in the following way (Gupta et al. 2004):

$$\text{Churn rate} = 1 - \text{Retention rate} \quad (2)$$

Deterministic Models

Berger and Nasr (1998) provide the following general formula for the customer lifetime value (CLV):

$$\text{CLV} = \sum_{t=1}^n \pi(t) \frac{\rho^t}{(1+d)^t} \quad (3)$$

with $\pi(t)$ profit generated in period t , ρ the rate of retention, and d the discount rate. If one considers profit stability over time for an annual net gain h , then CLV is formulated as

$$\text{CLV} = h \frac{\rho^t}{(1+d)^t} \quad (4)$$

We have a monetary component h and an expected number of transactions (or products, *discounted expected transactions*). This expression has the advantage of being extremely simple: one just has to estimate the retention rate to obtain the CLV. On the other hand, this approach assumes that the retention rate is stable over time.

However, this assumption fails to take into account the customer base composed of different segments, over and above all considerations of variation in the retention rate at individual level. Imagine that a same cohort of customers is composed of p homogeneous segments, each with an annual retention rate assumed to be constant from 1 year to the next for purpose of simplicity, with ρ_i for each segment i . One also can reason in discrete time for greater simplicity, but the situation can easily be extrapolated to continuous time. Let us assume that by nature segment 1 has the highest retention rate. The average retention rate, for example, in the first year is equal to

$$\bar{r} = \frac{\sum_{i=1}^p n_i \rho_i}{\sum_{k=1}^p n_k} \quad (5)$$

with n_i the size of segment i . Traditionally, portfolio value is calculated on the basis of this average rate.

However, because of the retention dynamic, the probability of belonging to segment 1 will converge toward 1, and, at the same time, the average retention rate will also converge toward the retention rate of segment 1. In effect, according to Bayes' theorem, one gets the probability of customer c belonging to segment 1 active after t years, formulated as

$$\begin{aligned}
 P(c \in S_1 | \text{active after } t \text{ years}) &= \frac{P(c \in S_1)P(\text{active after } t \text{ years} | c \in S_1)}{P(\text{active after } t \text{ years})} \\
 &= \frac{p_1 \rho_1^t}{\sum_{i=1}^n p_i \rho_i^t} = \frac{1}{1 + \frac{p_2 \rho_2^t}{p_1 \rho_1^t} + \dots + \frac{p_n \rho_n^t}{p_1 \rho_1^t}} \tag{6} \\
 &= \frac{1}{1 + \frac{p_2}{p_1} \left(\frac{\rho_2}{\rho_1}\right)^t + \dots + \frac{p_n}{p_1} \left(\frac{\rho_n}{\rho_1}\right)^t}
 \end{aligned}$$

However, since, by definition, $r_1 \geq r_i, \forall i \neq 1$ then $\forall i \neq 1, \lim_{t \rightarrow +\infty} \left(\frac{\rho_i}{\rho_1}\right)^t = 0$.

So the more time that passes (t becomes large), the higher the probability of belonging to segment 1, leaning toward a limit of 1. The average retention rate for a cohort thus converges toward the retention rate of segment 1. Variation in the retention rate is linked to the heterogeneous nature of the population. The use of an aggregate rate is not adapted for assessing the CLV. It can however be used by companies as a proxy for business health. Nowadays, adopting a stable retention rate represents a very particular case and is often inadequate.

Probabilistic Models

There are two types of probabilistic models: parametric and semi-parametric.

Parametric Models

In terms of parametric models, more elaborate models than the deterministic ones have been developed in the contractual framework. Thus, Fader and Hardie (2007b) used a survival function to obtain an expression such as (7)

$$E(\text{CLV}) = h \frac{S(t)}{(1+d)^t} \tag{7}$$

considering time as discrete. The link with the preceding form is obvious apart from the fact that $S(t)$ is the survival or retention function on date t and one can no longer speak about CLV but of expectancy of CLV. The authors assume that life span is given by a geometric distribution. The customer remains as such from one period to another with a probability $1-p$. In this context, $S(t) = (1-p)^t$. Interindividual heterogeneity in terms of probability p is given by a beta distribution (with values between 0 and 1). One thus obtains the shifted beta-geometric model (sBG).

Naturally, other expressions of survival are possible, notably with the inclusion of explanatory variables and the shift to continuous time. Schweidel et al. (2008) thus included explanatory variables while retaining a formulation with latent traits in continuous time. They developed the formula

$$S(t) = \int S[t | \theta_i, X(t)] g(\theta_i) . d\theta_i \tag{8}$$

with $X(t)$ as all of the explanatory variables for t and θ_i a set of individual latent traits. $g(\theta_i)$ represents the distribution of θ_i . This formulation ensures the harmonious integration of latent traits and explanatory variables, giving us a mixed effects model with fixed and random components.

$g(\theta_i)$ is the distribution that can be used to measure interindividual heterogeneity. It generally involves a gamma distribution for reasons of flexibility and compatibility with most survival distributions. It is expressed as follows:

$$g(\theta_i | r, \alpha) = \frac{\alpha^r \theta_i^{r-1} e^{-\alpha\theta_i}}{\Gamma(r)} \tag{9}$$

One can express the survival function in the form of the hazard function. The hazard function measures the instantaneous risk of mortality.

$$S [t | \theta_i, X(t)] = e^{-\sum_{v=1}^t \int_{v-1}^v h[u | \theta_i, X(t)] du} \tag{10}$$

If one concentrates on the stochastic dimension, the basic hazard function h_0 can adopt the Weibull distribution:

$$h_0(t | \theta_i, c) = c\theta_i t^{c-1} \tag{11}$$

This formulation takes into account risk that evolves over time. Variation in the retention rate depends as much on heterogeneity (interindividual variations) as on intrinsic individual variations. If $c = 1$, one then shifts to the exponential-gamma (EG) model. Note that in continuous time, this model is the equivalent of the sBG model (Fader et al. 2003).

Semi-Parametric Models

The most famous representative of semi-parametric models is the Cox model, often called the proportional hazard model. It models a life span considered as a random variable with a probability density $f(t)$ and a distribution function $F(t)$. The survival function is expressed as

$$S(t) = P(T \geq t) = 1 - F(t) \tag{12}$$

This function is of course monotonically decreasing.
The hazard function is written as

$$h(t) = \lim_{dt \rightarrow 0} \frac{P[(t \leq T < t + dt)(T \geq t)]}{dt} = \frac{f(t)}{S(t)} \tag{13}$$

Instead of taking the hazard function into consideration in a parametric way as in the preceding point, one estimates it following the Kaplan-Meier procedure. The cumulated hazard function is expressed as

$$H(t) = \int_0^t h(u) \cdot du = -\ln[S(t)] \quad (14)$$

The addition of explanatory variables in the form of an X matrix allows us to adopt a semi-parametric formulation:

$$h(t|X) = h_0(t)e^{X\beta} \quad (15)$$

$h_0(t)$ only depends on time. With Eq. 15, based on expression (11), the survival function becomes

$$S(t|X) = [S_0(t)]^{e^{X\beta}} \quad (16)$$

with the same formulation logic as the hazard function.

Migration Models for CLV Measurement

These models represent a generalization of retention models. The absence of transactions at any given moment does not mean that the customer has become inactive. This is typically the case in a noncontractual situation, in which customer inactivity cannot be observed.

The main idea is that customers go through different stages in their relationship with the brand with specific characteristics governing each stage. One therefore needs to describe the characteristics of these stages as well as the conditions for the transition from one stage to another.

Deterministic Models

Heuristics are frequently used to identify the situation of a customer in a deterministic context. The best known of these is the RFM segmentation (recency, frequency, and monetary value). Recency is the determinant factor to assess whether or not a customer is active. Customers are segmented on the basis of more or less valid thresholds. Traditionally, one distinguishes three levels per criterion R, F, and M, representing 27 segments. The more recently a customer has made a purchase, the greater his or her purchasing frequency, and the higher the average basket, the greater his or her supposed potential. This apparently logical hypothesis is, as noted earlier, qualified by observation of the behaviors of these different segments (e.g., Fader et al. 2005b).

At managerial level, a customer is traditionally considered as inactive beyond a certain length of time without arbitrarily fixed purchases. This method has been presented to us many times by firms that adopt a customer relation management approach. Schweidel et al. (2008) also noted its predominance in professional practice to determine whether or not a customer is active. Likewise, forecasts of future sales are made through a simple extrapolation of past sales.

Probabilistic Models

Two forms of approaches coexist. The first is in the form of Markov processes and the second in the form of combinations of models.

Markov Processes

In terms of migration models, the most widely used method is certainly that of Markov chains, also called Hidden Markov Models. Popularized by Pfeifer and Carraway in 2000, it has been the object of numerous extensions through the integration of sociodemographic or RFM variables. A customer is assumed to be in a certain relational situation with respect to the company, defined in advance. Naturally, these stages are never observed but remain latent which explains the term “Hidden Markov Models.” One can subsequently calculate the probability of transition from one state to another. Thus, Pfeifer and Carraway (2000) identified five levels of customer relations, from the most recent customers to buyers that bought such a long time ago they are considered as “non” or former customers. The transition pattern can be expressed graphically as follows (Fig. 1):

In this framework, there is perfect sequentiality. At stage 5, customers are considered as definitively lost with no chance of reactivation. This hypothesis can easily be changed. These models may be likened to latent class models except that adherence to a segment in the framework of hidden Markov models is dynamic and follows a Markov process.

Adapting Kumar (2007), a customer’s CLV may be expressed in the following way:

$$CLV = \sum_{t=0}^T \frac{MM_t P_t}{(1 + d)^t} \tag{17}$$

with MM_t the matrix of probability of transition from one state to another at t , d the rate of loss, and P_t the value generated by the customer on date t . Over time, the probability matrices merge with one another. Thus, if one starts from the probability of initial MM_0 transitions, one gets $t = 1$ $MM_1 = MM_0 \times MM_0 = (MM_0)^2$ and so $MM_t = (MM_0)^{t + 1}$.

A specific application is that of Rust et al. (2004) with a brand change matrix. Combined with a logit model, this application demonstrates the flexibility and the potential of the Markov approach.

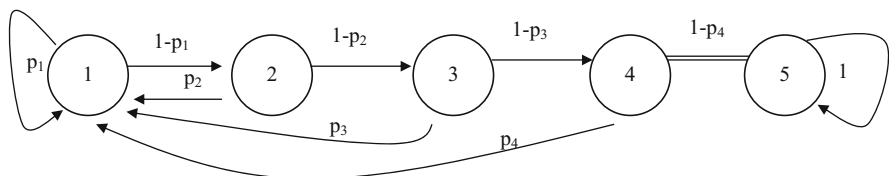


Fig. 1 Transition from one stage to another according to a Markov process

Combinations of Models

This approach was largely developed by Schweidel and Fader (2009). It considers that a stronger customer relationship with a brand (measured by the number of repeat purchases) is expressed through greater purchasing stability and so more stable interpurchase times. In short, it is the transcription of the transition from new customer to existing customer. There are thus two interpurchase periods that follow one another, the first characterized by an exponential distribution and the second by an Erlang 2 distribution. This distribution is a specific case of gamma distributions with a shape parameter equal to 2. The density is thus written as $\lambda^2 x e^{-\lambda x}$.

The transition from one state to the other occurs after each purchase with probability p . One thus arrives at a transition which respects a geometric process.

All the parameters of the different models are assumed heterogeneous. The two parameters transcribing the rate of transactions (exponential distribution and Erlang 2) are themselves gamma distributed, and probability is distributed according to a beta law. This line of research is interesting for several reasons: it takes into account explanatory variables, can be generalized to a larger number of situations and other distributions, etc.

Continuous Mixed Models: The Family of NBD Models to Measure CLV

This term is rarely used but allows us to describe the underlying nature of these models. It involves estimating the different processes simultaneously: consumption, attrition, etc. To this end, each process is assumed to correspond to a specific law. Customer heterogeneity is also expressed by a distribution. All the consumption characteristics are considered to be governed by latent traits. Explanatory variables may be integrated, depending on the degree of sophistication of the different models. Fundamentally though, the introduction of explanatory variables is not in accordance with the philosophy of these models based on stochastic determinants.

The whole palette of statistical tools is used here.

The Pareto/NBD and BG/NBD Model

In this context, continuous mixed models are considered as one of the most promising lines of research, especially the negative binomial formalization (NBD) model. The Poisson distribution of data y is combined with a gamma distribution of purchasing frequency. This approach, developed by Ehrenberg (1959), has been extended by taking into account the inactivity factor: the Pareto/NBD model (Schmittlein et al. 1987; Morrison and Schmittlein 1988; Schmittlein and Peterson 1994; Abe 2009; Jerath et al. 2011) or betaPareto/NBD model-geometric/NBD model (BG/NBD by Fader et al. 2005a). Thus, consumer behavior is represented by a continuous representation that, in theory, takes all of the individual specificities into account.

However, continuous mixed models imply a total parametric specification (generally Poisson with a specific frequency parameter distribution) that is by nature

restrictive and very often not very well adapted given the fundamental hypotheses of these distributions. Semi or nonparametric generalizations are naturally possible. However, their introduction requires a highly complex mathematical conceptualization process. In the same way, the introduction of explanatory variables is also possible but always at the price of a demanding mathematical formulation (Castéran et al. 2007a). Consequently, the operational and managerial scope of these models appears to be greatly reduced.

Finite mixed-effect models have been used for many years. The first principles were laid down by Newcomb (1886) and Pearson (1894). Finite mixed-effect models provide a specific case of latent class models (Baltagi 2003). They postulate the existence of latent classes within the population under study and a specific link between explained and explanatory variables within each of these classes. In this way, they underpin the existence of segments with specific behavioral patterns; the marketing implications are clearly apparent.

However, applications in a specifically marketing framework were initiated relatively late, mainly by Wedel et al. (1993). They provide a segmentation of the population beyond traditional behavioral segmentation. While apparently offering less detailed analysis than a continuous approach, segmentation does provide a clear interpretation of the results obtained as well as directly accessible managerial and operational implications. These implications are reinforced by the presence of explanatory variables. Each segment may be studied according to its own behavioral characteristics, which are explained by a set of variables. These explanatory variables help to determine the most effective marketing actions at the level of each segment. Finite mixed-effect models thus appear to be a promising alternative to continuous mixed models.

Nonetheless, to our best knowledge, the comparative efficiency of these models has only been demonstrated one time by Castéran et al. (2008). This comparison in terms of predictive validity between the finite mixed models and models of the NBD family (NBD simple, Pareto/NBD, BG/NBD) is worth exploring further.

The Explanatory Pareto/NBD and BG/NBD Model

The fact that all of these models are purely stochastic implies that they only have limited managerial potential. It is therefore important to reconcile the predictive validity of these purely stochastic models with an interpretative dimension resulting from the presence of explanatory variables. The introduction of explanatory variables within the Pareto/NBD model is a promising approach (Castéran et al. 2007b). This is done by the introduction of the explanatory variables in the gamma-gamma model by breaking down the variability of the scale parameter into two elements by distinguishing two components of parameter λ (purchasing frequency) and by using a regression with explanatory variables as well as a parameter λ_0 :

$$\lambda = \lambda_0 e^{X_i \beta} \quad (18)$$

with β the vector of coefficients and X_i the individual characteristics and marketing actions. Parameter λ_0 is distributed according to a gamma law of

parameters r (form) and α (scale). This parameter captures the residual heterogeneity not taken into account by the explanatory variables. Its density is expressed as

$$f(\lambda_0 | r, \alpha) = \frac{\alpha^r}{\Gamma(r)} \lambda_0^{r-1} e^{-\alpha \lambda_0} \quad (19)$$

with $\lambda_0 > 0$, $r > 0$ et $\alpha > 0$.

Then one has to adopt the same process for inactivity with regard to parameter μ . One notes the matrix of personal characteristics and marketing actions as X_2 (this can partially or entirely correspond to X_I). Inactivity thus becomes

$$\mu = \mu_0 e^{X_2 \gamma} \quad (20)$$

with parameter μ_0 following a gamma distribution of parameters s (form) and δ (scale),

$$f(\mu_0 | s, \delta) = \frac{\delta^s}{\Gamma(s)} \mu_0^{s-1} e^{-\delta \mu_0} \quad (21)$$

with $\mu_0 > 0$, $s > 0$ et $\delta > 0$.

In addition to the explanatory variables X (composed of X_I and X_2), one needs three additional elements: number of purchases y made during the period $[0, T]$, recency of the last purchase t_y (date of last purchase), and length of the period of estimation T . H is the combination of all three variables, $H = (y, t_y, T)$, and Θ the vector of all the coefficients, $\Theta = (r, \alpha, s, \beta, \gamma)$.

The limitation is due to the fact that one only deals with variables without a dynamic perspective, as they are constant over time.

Finally, Fader and Hardie (2007a) developed a general expression to introduce invariant explanatory variables over time within Pareto/NBD and BG/NBD models. The inclusion of these variables is conducted in a less complex way than the approach of Castéran et al. (2007b).

The Fig. 2 presents an overview of the CLV models.

The fundamental distinction is due to the nature of the relations between the customer and the company: is the customer's inactivity observed (contractual relations) or not? The second criterion comes from the type of model adopted: whole population, segment, or individual.

Finally, the last parameter is the distinction between continuous and discrete purchasing opportunities. However, this distinction is less crucial than the others insofar as certain discrete cases may be considered as continuous cases, while continuous cases can always be "discrete."

Casteran et al. (2007b) did not distinguish between variants with or without explanatory variables in this process. The presence of explanatory variables within purely stochastic formulations presents a methodological as well as a conceptual improvement.

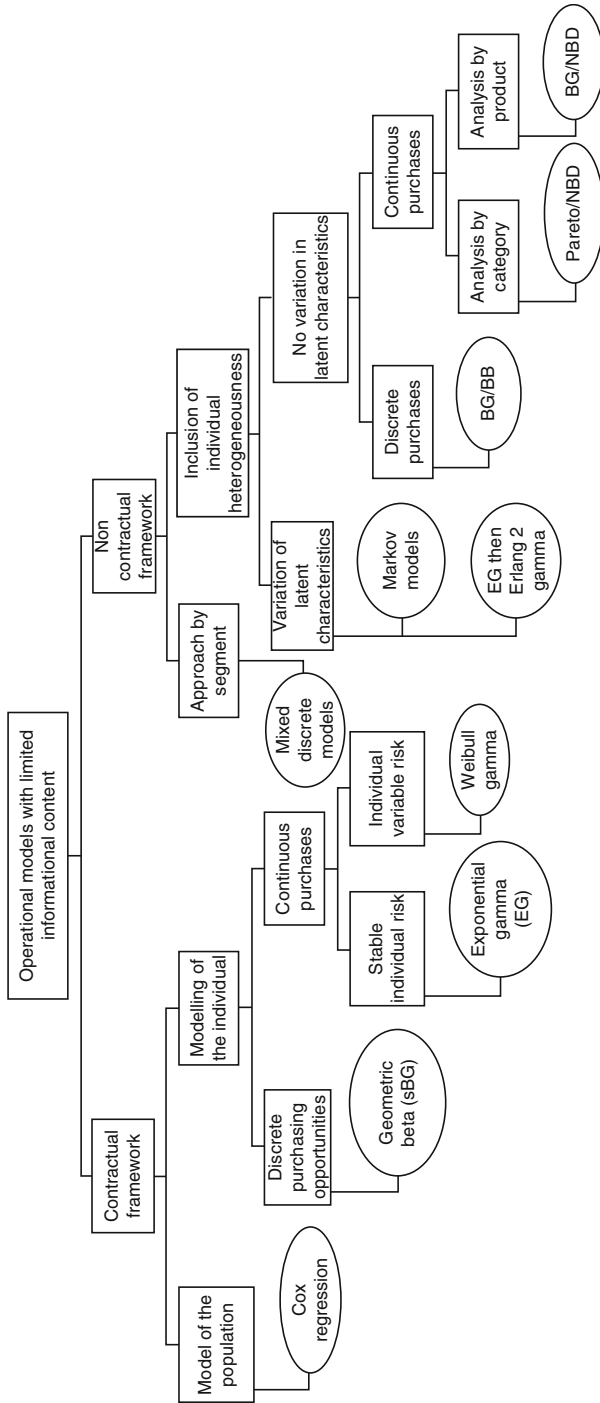


Fig. 2 Overview of the CLV measurement models

An Application of Stochastic Pareto/NBD and BG/NBD Models for Customer Base Management

After having presented the taxonomy of CLV, churn, and retention measurement models, this chapter shows a practical application by using some of the presented stochastic models (Pareto/NBD, BG/NBD) to model a customer base and the impact of a retail grocery loyalty program on customer churn, retention, and activity. The goal is to show how to implement, use, and interpret these sophisticated models by Customer base management applying them on firms' frequently used grocery store loyalty program databases and panel data.

Data and Methodology

The data for the practical application of customer base analysis come from a store loyalty program (LP) database of a large French grocery retailer (surface area 9,000 m²).

An LP consists of integrated and ruled systems of marketing actions (based on collection and redemption rules) that aim to encourage repeat purchases and increase the cost of switching as well as retention and subsequently CLV by providing short- and long-term incentives (Meyer-Waarden 2007; Blattberg et al. 2008; Bijmolt et al. 2010) and enhance "true loyalty," that is, increase behavioral (e.g., cross purchases, repeat purchases, mean basket size) and attitudinal (relationship building through positives attitudes, trust, attachment; Morgan and Hunt 1994) loyalty.

Loyalty programs (LP) are vastly popular – 90% of Europeans and 90% of US shoppers own at least one loyalty card. In 2010, the number of LP memberships in the United States exceeded 2.1 billion memberships, growing by 16% from the previous year despite the worldwide recession (Hlavinka and Sullivan 2011). For example, research estimates that the UK pharmacy chain Boots invested 30 million British pounds in the launch of its Advantage Card LP (Temporal and Trott 2001), and the U.K. retailer Tesco has spent an estimated 60 million pounds to operate its Clubcard LP (Bijmolt et al. 2010).

The store's loyalty program, launched in 1994, is free and provides price discounts, points exchangeable for gifts, and purchase vouchers on a varying set of items. The value of points won increases linearly according to the amount customers spend. Cardholder purchases account for 70% of store revenues. In the analysis, cardholder information is used, identifiable on an individual basis, which includes the majority of customers. The data set contains also information about competing loyalty card memberships and household characteristics (e.g., age, revenue). Scanner data include the transaction details, such as date of purchase and amount paid. Because people often shop on a weekly basis, the daily purchases are aggregated by individuals into a weekly frequency. The transaction data pertain to 5,000 households over a period of 156 weeks (week 2/1998 to week 2/2001).

The LP data is matched with BehaviorScan panel test market data from Angers, France. Scanning technology provides exhaustive recording of the purchasing

behavior (95% of fast-moving consumer goods sales in the area) of the panelist households, which are representative of the national population.

The implementation conditions of the models with LP data relate to two aspects: the consideration of new customers that have realized their first purchase, with certainty, and the duration of the estimation period. The date of first purchase in a noncontractual framework may be considered known, whether because of the nature of the data (e.g., Fader et al. 2005b) or the data processing method (Batislam et al. 2007). The Pareto/NBD, BG/NBD, and PDO models forecast all purchases by combining the number of customers at a given date with the unconditional expectancy of the number of purchases, according to the customer maturity level (Fader et al. 2005a). As stated higher, the first purchase is easy to identify in certain settings, when the entire customer history is available. Nevertheless, it is difficult, in most noncontractual relationships, to determine an exact date of the actual first purchase. (An exception are mail order business or e-commerce retail settings, where the shipping address of the customer is known and the purchase can be identified.) In grocery retailing, data from a loyalty program (LP) are left censored, and the first purchase cannot be characterized with certainty, because customers probably purchase before they enroll in a loyalty scheme. Therefore, it is unclear what types of customers are observed: “real” new customers, previously existing customers who have lately adopted the LP or customers who had lapsed or have low usage patterns. Batislam et al. (2007) and Jerath et al. (2011) both resort to drastic truncations to achieve a sample that consists entirely of new customers, who made no purchases in the first 13 months of their observations, whom they logically argue are genuine new customers. This approach means the loss of substantial information and raises questions regarding sample representativeness. For example, consumers who make their first purchases at different times, later or earlier, behave differently in their purchase frequency and loyalty in business settings (Schmittlein and Peterson 1994). The best customers often self-select as early adopters (Meyer-Waarden and Benavent 2009; Rogers 2003), so truncating samples could exclude insights into some of the firm’s best customers and earliest adopters. Another option to solve the left-censored data issue is to treat the first customer purchase observed in the LP database as the customer’s first actual buying act. This method creates the risk of combining different cohorts though, with different actual dates of first purchase; its consequences for models’ predictive validity have never been examined.

Both the Pareto/NBD and the BG/NBD models require tracking customer transactions, starting with their initial purchases, which raises the possibility of left-censored data, because one does not know when people became aware of the outlet’s loyalty program and if the first purchases recorded after enrollment are really their first transactions. In other words, households may have bought before adopting the loyalty scheme. Because the store does not have information about the initial purchases of cardholders, this methodological problem treated by left-filtering the panel customer records with transactions before October 14, 1998, which guarantees that the customers in the analysis are newcomers with known initial purchase times data (see Batislam et al. 2007). Thus, 5,000 households of a cohort of 997 new households is extracted that made their first purchases within the same 3-month

period (October–January) and realized a total of 6,005 transactions. The panel data is left-filtered and aggregated on a weekly basis, and the final observation window covers 78 weeks, from October 14, 1998, to April 13, 2000. The estimation period is restrained to 26 weeks and the calibration period to 52 weeks (hold-out sample) to establish the predictive validity of the models.

Finally, we get a matrix with one row for each customer, and at least three columns:

1. Customer's frequency, y : number of repeated transactions made in the estimation period
2. Customer's recency, t_y : the time of their last transaction
3. Customer's total time observed in the estimation period, T

Other columns can be added, one for each explanatory variable. The explanatory variables have to be quantitative or dummy variables.

Estimation

Parameter estimation of the Pareto/NBD model is more complex (Fader et al. 2005a; Reinartz and Kumar 2003; Fader and Hardie 2013); in particular, the maximum likelihood estimation (MLE) approach to estimating key parameters used requires a numerical search algorithm that must evaluate the Gauss hypergeometric function (Schmittlein and Peterson 1994).

For the estimation, free statistical software *R* 3.3.1 (R Development Core Team 2010) is used. Our advice is to adopt a two-step approach. Firstly, we estimate the purely stochastic model parameters (Pareto/NBD or BG/NBD) without explanatory variables. Secondly, we incorporate the explanatory variables and launch a new estimation process based on the results of the first estimation.

Estimation of the Purely Stochastic Parameters

The easiest way for estimating simple Pareto/NBD and BG/NBD models is now to use a dedicated package BYTD (<https://CRAN.R-project.org/package=BYTD>) “Buy 'Til You Die Models” (Dziurzynski et al. 2014). The estimation of the parameters is made through, respectively, the functions *pnbd.EstimateParameters* (Pareto/NBD) and *bgnbd.EstimateParameters* (NG/NBD).

There are two major estimation issues. First, the procedures are time-consuming, especially with regard to the initial values. Second, depending on these values, non-convergence might be faced, which makes it even more difficult to find an operational set of initial values. Many tries can be required in order to assure a clear convergence. The best choice, even if convergence occurs, is to relaunch with several starting points in order to compare the results and the value of the log-likelihood.

A good starting point is to consider that the average purchase rate is the ratio r/α . It is not sufficient to determine the exact starting values of the parameters, but it can

be useful that the initial values of r and α are chosen with respect to the average purchase rate of the dataset.

Since an optimal (in sense of log-likelihood maximization) set of parameters is found, the incorporation of explanatory variables can begin.

Estimation of the Parameters Including Explanatory Variables

The first step is to declare the explanatory variables. If *mydata* is the name of the dataset, the vector of the explanatory variables for the purchasing frequency X_1 could be the following:

```
X1=cbind(mydata$LoyCard, mydata$HhSize, mydata$NumbCard,
mydata$DisStore, mydata$SeniorManager, mydata$Unemployment,
mydata$Income_6, mydata$Income6_12, mydata$Income12_18,
mydata$Income18_24, mydata$Income24_30, mydata$A30_39,
mydata$A40_49, mydata$A50+)
```

Our variables describe:

1. The characteristics of the individuals: the household size (*HhSize*), their age (the *A...* dummy variables), their net wages (the *Income...* dummy variables), and the professional occupation (*SeniorManager*, *Unemployment*)
2. The relationship to the store: the distance in kilometers from the store (*DisStore*), the owning of loyalty cards (*LoyCard*), and the total number of loyalty cards owned by the household (*NumbCard*)

```
X2=X1 #Explanatory variable vector for inactivity part
```

At the beginning, the explanatory variables for purchasing frequency and inactivity can be the same. Further, during the selection process, the two sets will become different.

In order to incorporate qualitative variables, we divide them into dummy variables (e.g., income or customer age). To avoid overidentification, one modality of each variable shall be excluded from the estimation process. Whatever the modality is, the exclusion of one modality per qualitative variable is mandatory.

The set of initial values (b_0 in the example) is determined on the basis of the first estimation with *pnb.EstimateParameters* (Pareto/NBD). Those parameters are called *params* here.

```
b0<-c(params, rep(0, ncol(X1)+ncol(X2)))
```

The initial values for the explanatory variables are set to 0 in order to relaunch the estimation process at the same initial state as purely stochastic approaches. The reestimation process can now begin.

For the maximization of the log-likelihood function and estimation of the parameters, two functions are employed: *nlminb* and *optim*. The *nlminb* procedure is more flexible and presents fewer convergence problems. After estimating *nlminb*, the *optim* is used to compute the Hessian matrix for estimating the covariance matrix

(standard error of the coefficients). In the following example however, we directly use the *optim* function.

The estimation of the Pareto/NBD model – an effort whose difficulty is frequently cited as a usage limitation – is considerably facilitated by the *gsl* package, with the expression *hyperg_2F1*, that enables the estimation of a Gaussian hypergeometric function to increase external validity.

The final log-likelihood of the explanatory Pareto/NBD model can be written as

$$LL(\Theta_i|H_i, X_i) = \ln \Gamma(r + y) - \ln \Gamma(r) + r \ln \alpha + s \ln \delta + y \ln B + \ln[A_1A_2 + A_3A_0] \tag{22}$$

with

- (i) $B = e^{X_1\beta}$ and $G = e^{X_2\gamma}$
- (ii) Due to the presence of the Gaussian hypergeometric function and the form of the integrals, we must distinguish two cases: when $\alpha e^{X_2\gamma} \geq \delta e^{X_1\beta}$ and the opposite case. For each case, we note a different expression for A_0 :
If $\alpha G \geq \delta B$,

$$A_0 = \frac{\left(\frac{B}{G}\right)^{s+1}}{B} \times \left[\frac{{}_2F_2\left(s+1; r+s+y+1; \frac{\alpha - \delta \frac{B}{G}}{\alpha + t_y B}\right)}{(\alpha + t_y B)^{(r+s+y)}} - \frac{{}_2F_2\left(s+1; r+s+y+1; \frac{\alpha - \delta \frac{B}{G}}{\alpha + TB}\right)}{(\alpha + TB)^{(r+s+y)}} \right] \tag{23}$$

If $\alpha G \leq \delta B$,

$$A_0 = \frac{\left(\frac{G}{B}\right)^{r+y}}{G} \times \left[\frac{{}_2F_2\left(r+y; r+s+y+1; \frac{\delta - \alpha \frac{G}{B}}{\delta + t_y G}\right)}{(\delta + t_y G)^{(r+s+y)}} - \frac{{}_2F_2\left(r+y; r+s+y+1; \frac{\delta - \alpha \frac{G}{B}}{\delta + TG}\right)}{(\delta + TG)^{(r+s+y)}} \right] \tag{24}$$

- (iii) $A_1 = (TB + \alpha)^{-(r+y)}$ and $A_2 = (TG + \delta)^{-s}$
- (iv) $A_3 = \frac{Gs}{r+s+y}$

Since optimization algorithms classically perform minimization, we use the negative form of the log-likelihood function.

```

library(gsl)
LL_Paretoexp <-function(p) {
# Parameters vector
  r<-p[1]
  alpha<-p[2]
  s<-p[3]
  delta<-p[4]
# Number of covariates
  nX1=ncol(X1)          # for purchasing frequency
  nX2=length(p)-4-nX1  # for inactivity process
# Coefficients of explanatory variables
  b1=p[5:(4+nX1)]      # for purchasing frequency
  g1=p[(5+nX1):length(p)] # for inactivity process
# Regressions
  B<-exp(as.matrix(X1)%*%b1)
  G<-exp(as.matrix(X2)%*%g1)
#Meta-functions
  A1<-(B*T+alpha)^(-r-y)
  A2<-(G*T+delta)^(-s)
  A3<-G*s/(r+s+y)
# A0 expression
  coef1<-(B^s)/(G^(s+1))
  arg11<-hyperg_2F1(s+1, r+s+y, r+s+y+1, (alpha-delta*B/G)/
(alpha+t_y*B))/((alpha+t_y*B)^(r+s+y)) # t_y = t_y
  arg12<-hyperg_2F1(s+1, r+s+y, r+s+y+1, (alpha-delta*B/G)/
(alpha+T*B))/((alpha+T*B)^(r+s+y))
  coef2<-((G/B)^(r+y))/G
  arg21<-hyperg_2F1(r+y, r+s+y, r+s+y+1, (delta-alpha*G/B)/
(delta+t_y*G))/((delta+t_y*G)^(r+s+y))
  arg22<-hyperg_2F1(r+y, r+s+y, r+s+y+1, (delta-alpha*G/B)/
(delta+T*G))/((delta+T*G)^(r+s+y))
  A0<-ifelse(alpha*G>delta*B, coef1*(arg11-arg12), coef2*
(arg21-arg22))
# Log-likelihood function
  -sum(lgamma(r+y)-lgamma(r)+ r*log(alpha)+ s*log(delta)+ y*log
(B)+ log(A1*A2+A3*A0))
}

```

The lower bounds are 10^{-3} for the stochastic parameters:

```

min<-c(rep(1e-3,4), rep(-Inf,length(b0)-4))
max<-rep(Inf, length(b0))

```

The selection of the explanatory variables represents a significant challenge. On the basis of the Hessian matrix, small set of variables is kept, though traditionally, modeling purchase behavior is quite complex and the identification of relevant variables very difficult (Ehrenberg 1988).

```
optimal<-optim(b0, fn=LL_Paretoexp, method="L-BFGS-B",
control=list(trace=6, REPORT=1), hessian=TRUE, lower=min,
upper=max)
optimal # Result of the estimation process
```

Let us remind that we get the standard errors by taking the square root of the diagonal elements of the covariance matrix. The covariance matrix is the inverse of the Hessian matrix obtained through the minimization of the negative log-likelihood: *optimal*. All results are presented in a table with the coefficients (*Coeffs*), the standard errors (*StdError*), and the t-values (*t*). The null hypothesis that the coefficients are not significantly different from 0 (β_i or $\gamma_i = 0$) is rejected at a 5% significance level if $t \in]-\infty; -1.96] \cup [1.96; +\infty[$.

We compute also the Bayesian information criterion *BIC*, a common indicator, expressed as $-2LL + k \ln(N)$, where k is the number of parameters to be estimated ($length(b0)$), N indicates the number of individuals ($length(y)$), and LL is the maximum log-likelihood value for the model ($-optimal$value$).

We recommend a step-by-step descending selection process by removing one variable at each step:

1. We launch the estimation with the whole set of potential covariates.
2. We remove the covariate with the closest to 0 t-value while the t-value belongs to $] -1.96; 1.96[$.
3. We relaunch the estimation process with the new set of variables.
4. We check for the improvement of the BIC value.
5. While we have t-values comprised between -1.96 (excluded) and 1.96 (excluded) and while the BIC value is improving, we return to step 2.

This selection process can be quite slow but allows an appropriate selection of the covariates.

```
# Computation of the standard errors
inverse<-solve(optimal$hessian)
result<-cbind(optimal$par, sqrt(diag(inverse)), optimal$par/sqrt
(diag(inverse)))
colnames(result)<-c("Coeffs", "StdError", "t")
rownames(result)<-c("r", "alpha", "s", "delta", colnames(X1),
colnames(X2))
print(result)
# Computation of the Bayesian Information Criterion
BIC<-optimal$value*2+length(b0)*log(length(y))
BIC
```

Results

The descriptive results offer a comprehensive overview of the data sets from the grocery sector which is compared with different data used in previous investigations. In addition, the parameter estimation and comparison of the different models is demonstrated.

Of the 997 total customers in the 26-week cohort, 46.3% are zero repurchasers (Means = 1.69, SD = 3.59). The grocery data indicate that the median interpurchase times, even after excluding zero repurchase, is approximately 10.6 weeks, which is low compared with the other applications of the Pareto/NBD model, for which the median interpurchase time is 7 months (office supplies; Schmittlein and Peterson 1994), 17 weeks (catalogue sales; Reinartz and Kumar 2000), or 25 weeks (computer-related products; Reinartz and Kumar 2000). The grocery category features very short purchase cycles, because grocery items are not durable and require frequent replenishment. In addition, the number and heterogeneity of customers is higher in the grocery retail context. For example, the online CD customer base used by Fader et al. (2005a) includes a majority of customers (approximately 85%) who make zero (60%), one, or two repurchases. 46% of grocery retail customers are zero repurchasers, and customers with zero, one, or two repurchases constitute 80% of total grocery retail customers. In contrast, Batislam et al. (2007) find that approximately 40% of grocery retail customers are zero repurchasers, and customers with zero, one, or two repurchases make up around 65% of total grocery retail customers. Such high heterogeneity in grocery purchases decreases the precision of the models.

Parameter Estimations

In order to show a practical application, we interpret the estimated coefficients. They seem coherent for the Pareto/NBD model, with signs in the correct direction (see Table 2).

Purchase frequency is positively influenced by a LP, which is coherent with existing literature (Meyer-Waarden 2007; Leenheer et al. 2007; Liu 2007). The professional occupation of the household members has a direct impact on the purchase activity (frequency) and retention, though a professional situation has the same positive impact on purchase frequency as does lower income. Furthermore,

Table 2 Regression coefficients of the Pareto/NBD model

Frequency regression (β)		Inactivity regression (γ)	
Loyalty card of the shop	+1.01	# of loyalty cards	-0.17
Senior manager	+0.30	Low wage (dummy)	+0.10
Low wage (dummy)	+0.29	Unemployment (dummy)	+1.10
Over 50 (dummy)	-0.73		

Notes: The insignificant coefficients ($p > 0.1$) are household size, profession (employee, worker), wages (1,000–2,000 €, > 2,000 €), and age (30–50 years, < 30 years)

people older than 50 years of age are less mobile than younger people and display lower purchase frequencies. Younger shoppers are more likely to engage in smaller, more frequent fill-in trips than are older ones, probably because the former buyers have more disposable time but less income, which drives them to buy in smaller quantities at higher frequencies (Kahn and Schmittlein 1989; Bell et al. 1998). Financial instability of households (i.e., low wages, unemployment) has a negative impact on inactivity. Grocery patronage behavior depends on the level of education and income, which increase the chances that the consumers uses a more rational purchase process and thus attaches less importance to marketing variables (e.g., store advertisement, promotions, loyalty program rewards). Generally, the more education people possess, the less sensitive they are to a store's promotions or other marketing actions, and the less loyal they are, which means their defection probability is higher and retention is lower (Narasimhan 1984). Less educated households with lower incomes tend to remain loyal, because they experience more influence from store marketing variables. According to an alternative but not incompatible explanation, they also probably have higher switching costs related to mobility constraints (money, transports), which increases the utility of the closest and most familiar store.

Multiple LP memberships relate positively to inactivity, which is coherent with the results of Meyer-Waarden (2007) and may indicate a learning effect with regard to the use of loyalty schemes. Disloyal, opportunistic buyers who regularly shop in several stores and are members of different loyalty schemes (on average, European households possess three grocery retailing loyalty cards; ACNielsen 2005) are more experienced and have smaller switching costs. These purchasers join LP more readily and quickly (Meyer-Waarden 2007; Leenheer et al. 2007).

Table 3 provides the results of the gamma and beta distributions. The parameters for frequency do not vary significantly, despite the introduction of explanatory variables. However, the parameters for the inactivity or dropout rates vary strongly; the drastic growth of δ probably relates to the explanatory variables.

Table 3 Coefficients of the gamma/beta distributions

	Basic NBD	BG/NBD	Standard Pareto/NBD	Explanatory Pareto/NBD
Index of homogeneity in purchase rate: r	0.50	0.43	0.57	0.66
α	5.72	3.94	5.60	6.91
Average purchase rate: r/α	0.09	0.10	0.10	0.10
a		0.22		
b		1.14		
Average inactivity probability: $a/(a + b)$		0.16		
Index of homogeneity in inactivity rate: s			0.63	1.56
δ			30.16	107.55
Average inactivity rate: s/δ			0.02	0.01

The parameters r (which can be seen as purchase rates) and s (which can be seen as churn rates) increase in the explanatory model. Both provide an index of homogeneity (Schmittlein et al. 1987), and their increase denotes more significant homogeneity across customers in the explanatory model. For the explanatory formulation, gamma functions capture residual heterogeneity, not all the heterogeneity, as in the case of purely stochastic formulations.

Purchase Prediction Validity

Empirical analysis carried out for both the 26- and 52-week observation periods for the cohort relies on a popular criterion for adjustment, the Bayesian information criterion (BIC), whose values are based on a log scale. The expression is written as follows: $BIC = -2LL + k \ln(n)$, where k the number of parameters and n the sample size.

The adjustment differences between the BG/NBD approach and the explanatory Pareto/NBD model are not very important, and the BIC is very close for both (see Table 4). If one considers the mean absolute percent error (MAPE) as an empirical criterion, the explanatory Pareto/NBD model has slightly worse results than either the standard or the BG/NBD model (15.5% vs. 12% and 10.5%; the basic NBD achieves the worst results at 38.1%). This result makes sense. According to Fader et al. (2005a), the BG/NBD forecasts are better when purchase frequency is high, as in the grocery retailing context, because of the differences among the model structures. Under the Pareto/NBD model, dropout occurs at any time – even before a customer has made a first purchase. However, under the BG/NBD, a customer cannot become inactive before making his or her first purchase. If buying rates are fairly high, BG/NBD and Pareto/NBD perform similarly well. However, in contexts in which purchase frequencies are low, the BG/NBD model suffers in comparison with the Pareto/NBD approach.

After having tested the robustness of the models, a more thorough investigation of their performance is completed. The accuracies of the different models are not similar (Fig. 3).

During the validation period, the BG/NBD model performs quite well, whereas the Pareto/NBD and explanatory Pareto/NBD formulations underestimate the weekly purchase frequency. The basic NBD model does not perform well at all. With the exception of the basic NBD model, the approaches converge to actual repeat purchases during the forecast period. Weekly sales rise during the first 14 weeks, due to new customers in the cohort and their repeat purchases. All models underestimate the peak in weekly actual purchases in the initial weeks, probably

Table 4 Log-likelihood and Bayesian information criterion

	Basic NBD	BG/NBD	Standard Pareto/NBD	Explanatory Pareto/NBD
Log-Likelihood	-4,954	-4,922	-4,935	-4,900
Bayesian information criterion	9,922	9,872	9,898	9,876

because they miss the increasing trend in repeat purchases due to promotions during the same period. Later in the observation period, all models (with the exception of the basic NBD) match the actual purchases.

The deviation of weekly estimates from actual purchases during the initial weeks leads to an underestimation of the cumulative repeat purchases in the initial weeks as well (see Fig. 4).

During the forecast period (52 weeks), the models underestimate actual purchases (Pareto/NBD model: -9% , explanatory Pareto/NBD model: -14% , BG/NBD

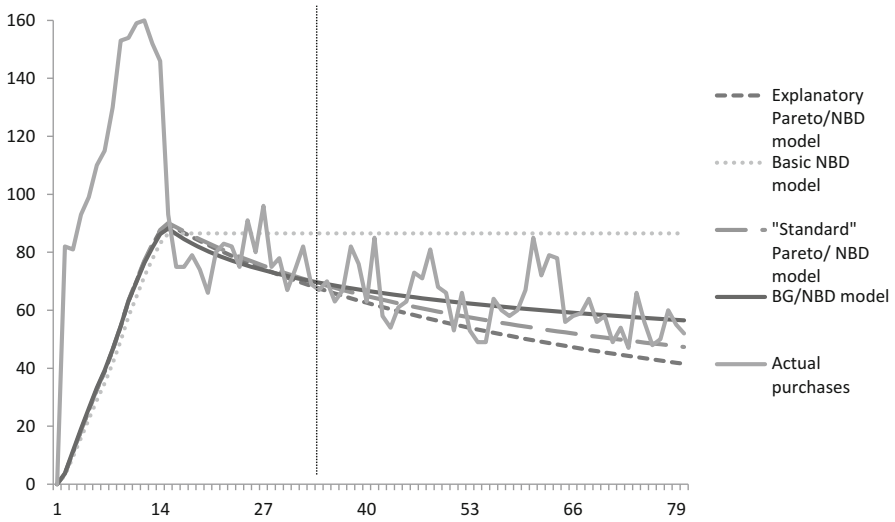


Fig. 3 Estimation of the weekly repeat purchases

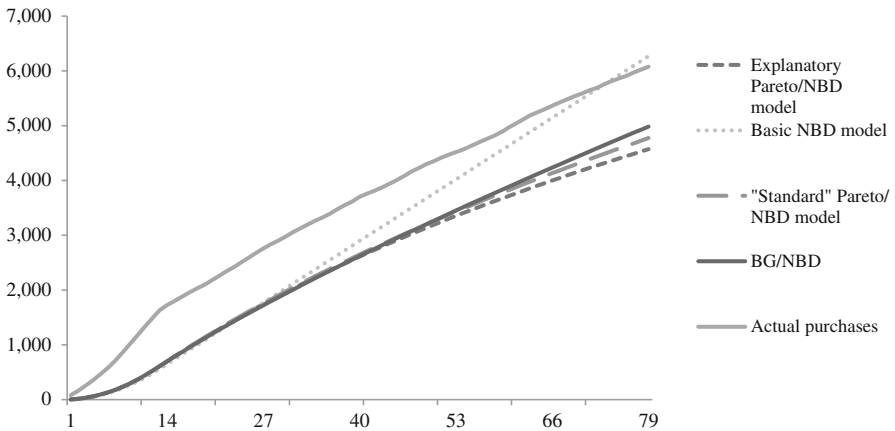


Fig. 4 Estimation of the cumulative repeat purchases

model: -2%). The fairly high purchase frequency rates in grocery shopping may explain the strong results derived from the BG/NBD model. In this case, the assumption that a customer is active until he or she makes a repeat purchase is not a problem. However, purchase frequency is not too high to affect the dropout time (“exhaustion” effect) of the BG/NBD model.

We measure individual-level performance according to the conditional expectations for the forecast period, depending on the number of repeat purchases in the observation period (Fig. 5). That is, for each value of x in the observation period, an average of the actual number of purchases in the forecast period is compared.

The forecasts of the BG/NBD and the standard Pareto/NBD models are very close and provide acceptable predictions of the expected number of transactions in the holdout period, consistent with the results of Fader et al. (2005b). The Pareto/NBD model offers slightly better predictions than the BG/NBD, but it is important to keep in mind that the number of heavy buyers is small. The explanatory model and the basic NBD model systematically overestimate the number of repeat purchases, especially for heavy customers.

Another way to assess the predictive validity of the models is to group customers on the basis of their recency and frequency characteristics. One can then compare the results with traditional recency/frequency (RF) segmentation analysis.

Each of the customers is assigned to a RF segment in the following manner. The terciles for recency and frequency (the customers who made no repeat purchases are coded as $R = F = 0$) have to be determined. High recency means a low number of days since the last purchase, i.e., a recent repurchaser. At the opposite, a low recency characterizes an exceptional repurchase. In Table 5, the size of each RF group is shown.

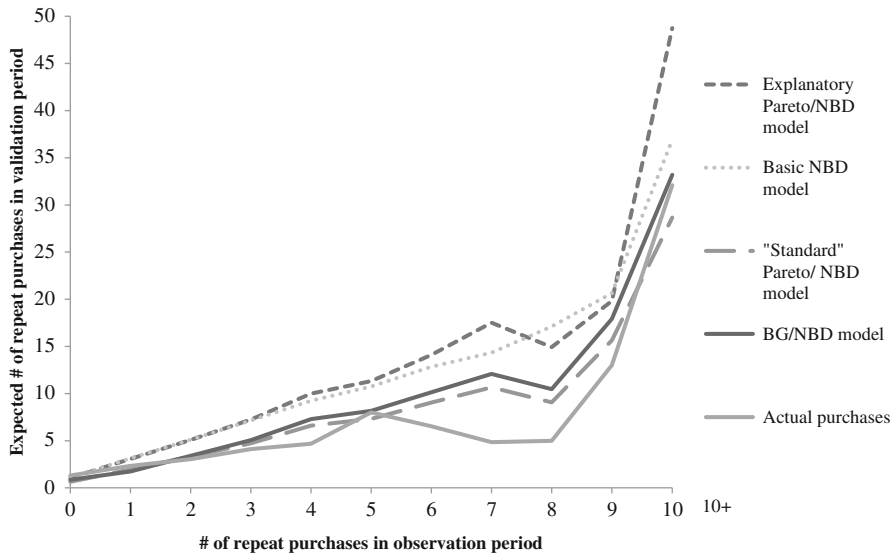


Fig. 5 Conditional expectations

Only 15% of the customers are frequent and recent repurchasers. On the other hand, the zero repurchasers during the estimation period represent almost half of the customers. In a traditional approach, managers in the retailing sector would assume that, after a half year of inactivity, a customer is inactive.

In fact, the average number of purchases made by those customers, during the following 52 weeks, is four times lower in average than the same number made by other segments. However, due to the size of this group, their contribution is really impressive: they represent 18.3% of the total of the purchases of the following 52 weeks, the second contribution of all the segments. This aspect is very interesting. It is taken into account by the models (even the contribution of the zero repurchaser is underestimated between 9.6% and 11.7% instead of 18.3%) (Fig. 6).

Table 5 Repartition of the customers between RF segmentation

Recency	Frequency of repeat purchases (estimation period: 26 weeks)	# of customers
No repeat purchase	0	46.3%
Low recency	1	3.0%
	2	0.5%
Total low recency		3.5%
Medium recency	1	14.1%
	2	5.6%
	3+	5.1%
Total medium recency		24.9%
High recency	1	5.2%
	2	4.9%
	3+	15.1%
Total high recency		25.3%

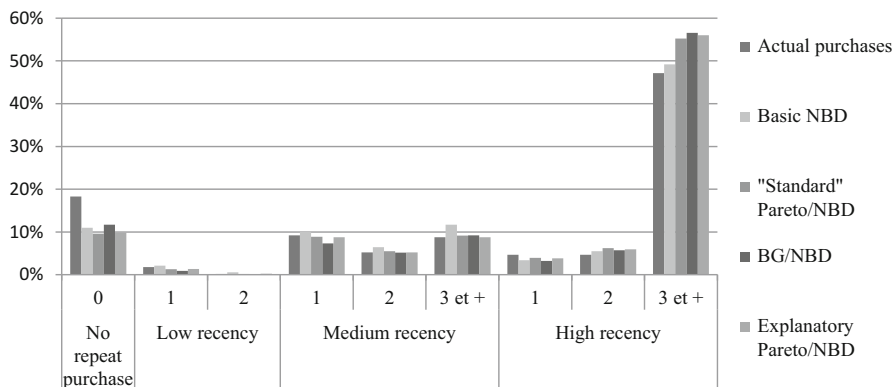


Fig. 6 Purchases by recency and frequency

Conclusion

We have seen an operational form of CLV, retention, and churn models, (namely, the Pareto/NBD, BG/NBD, and explanatory Pareto/NBD models) and their high degree of validity for customer base analysis and for forecasting a customer's future purchasing, conditional on his or her past buying behavior. The Pareto/NBD, BG/NBD, and explanatory Pareto/NBD models systematically outperform the basic NBD model, because it does not consider inactivity.

The Pareto/NBD and explanatory Pareto/NBD formulations underestimate weekly purchase frequency, whereas the BG/NBD model performs quite well. These results show that explanatory variables introduce more information and therefore generate a better forecast.

However, even if the predictive validity of the explanatory NBD/Pareto model is not necessarily better, its performance does not suffer in comparison with the Pareto/NBD and the BG/NBD models. Nevertheless, the advantages of the explanatory approach relate more to the opportunity to explain the impact of personal characteristics and the impact of marketing actions rather than the accuracy of the forecasts at an aggregate level. The ability of the explanatory Pareto/NBD model to predict future purchases is quite good. Even with a reduced set of explanatory variables, the explanatory Pareto/NBD model is as accurate as the standard formulation. Nevertheless, the results are not better than those of the BG/NBD approach. However, improvements are possible with other sets of variables (i.e., more marketing mix variables).

These CLV, retention, and churn models for customer base analysis can help managers understand why their marketing operations work, or do not work, and how and to which customer segments they should improve their efforts. The explanatory model approach represents a promising way to understand buyer behavior. The applications are broad, including segmentation, understanding customer life cycles, determining elasticities and elements that influence loyalty and purchase behavior, the possibility of analyzing marketing actions and personal characteristics, and a means to establish more valid customer CLV models to predict customer value.

Managers should be encouraged to use these models to determine their customer base analysis, CLV calculations, and resource allocations, using their often large longitudinal databases.

Further research should address underlying model assumptions that are unrealistic and not compatible with extant literature about purchasing behavior. For example, researchers could relax the Poisson distribution assumptions and perhaps use a Weibull distribution instead. The BG/NBD formulation suffers a major weakness because its underlying conditions (i.e., dropout rates independent of purchase frequencies) demand inactivity appear immediately after each repurchase act. This behavioral assumption is not compatible with purchasing behavior literature. In the same sense, the Pareto/NBD model supposes independence between purchase frequency and inactivity, which may be reasonable only in "always-a-share" markets (Reinartz and Kumar 2000). Some other authors also suppose a link between both variables (East et al. 2000).

Few current models explicitly incorporate competition, yet heightened competition can affect customer CLV in several ways – shortened expected lifetime, decreased prices, and increased acquisition costs. Panel data provide a promising source for some firms, and surveys can be very useful in capturing the effect of competition. Other empirical investigations should examine in which conditions (high/low purchase frequencies) and with which type of data (internal, panel) the different models perform best.

Compared with predicting purchase frequency and weekly repeat purchases, forecasts of individual purchases include more customer information and should provide higher accuracy in individual-level forecasts. However, it remains difficult to model individual purchase behavior, especially with regard to the highly heterogeneous purchase behavior encountered in grocery sales (Fader and Hardie 2013).

Finally, to allocate optimally, managers cannot simply measure CLV but instead must know how CLV reacts to changes in the marketing mix. Additional research should address this concern.

References

- Abe, M. (2009). 'counting your customers' one by one: A hierarchical Bayes extension to the Pareto/NBD model. *Marketing Science*, 28(3), 541–553.
- Allenby, G. M., Leone, R. P., & Jen, L. (1999). A dynamic model of purchase timing with application to direct marketing. *Journal of the American Statistical Association*, 94(446), 365–374.
- Ayache, A., Calciu, M., & Salerno, F. (2006). Stochastic approach to customer equity and lifetime value calculations with applications to customer retention models and some extensions. *35th Conference of EMAC*, Athènes.
- Baltagi, B. H. (2003). *A companion to theoretical econometrics*. Willinston: Blackwell.
- Batistlam, E., Denizel, M., & Filiztekin, A. (2007). Empirical validation and comparison of models for customer base analysis. *International Journal of Research in Marketing*, 24(3), 201–209.
- Bell, D. R., Ho, T. H., & Tang, C. S. (1998). Determining where to shop: Fixed and variable costs of shopping. *Journal of Marketing Research*, 35(3), 352–369.
- Berger P., & Nasr, N. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, Winter 98, 12(1), 17–30.
- Bijmolt, T., Dorotic, M., & Verhoef, P. (2010). Loyalty programs: Generalizations on their adoption, effectiveness and design. *Foundations & Trends in Marketing*, 5(4), 197–258.
- Bitran, G., & Mondschein, S. (1996). Mailing decisions in the catalog sales industry. *Management Science*, 42(9), 1364–1381.
- Blattberg, R. C., & Deighton, J. (1996). Manage marketing by the customer equity test. *Harvard Business Review*, 74, 136–144.
- Blattberg, R. C., Getz, G., & Thomas, J. S. (2001). *Customer equity: Building and managing relationships as valued assets*. Boston: Harvard Business School Press.
- Blattberg, R. C., Kim, B., & Neslin, S. A. (2008). *Database marketing. Analyzing and managing customers*. New York: Springer.
- Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The theoretical underpinnings of customer asset management: A framework and propositions for future research. *Journal of the Academy of Marketing Science*, 32(3), 271–292.

- Borle, S., Singh, S., & Jain, D. C. (2008). Customer lifetime value measurement. *Management Science*, 54(1), 100–112.
- Calciu, M., & Salerno, F. (2002). Customer value modelling: Synthesis and extension proposals. *Journal of Targeting, Measurement & Analysis for Marketing*, 11(2), 124–134.
- Castéran, H., Meyer-Waarden, L., & Benavent, C. (2007a). *Empirical evaluation of NBD models for the estimation of life time value in the retailing sector*. XXIIIrd International Annual Conference of the French Marketing Association, Aix-les-Bains.
- Castéran, H., Meyer-Waarden, L., & Benavent, C. (2007b). *Incorporation of covariates in the Pareto/NBD model: First formulations and comparison with others models in the retailing sector*. Third German French Austrian Conference, ESSEC, Paris.
- Castéran, H., Meyer-Waarden, L., & Benavent, C. (2008). *Application of latent class models to purchases in the retailing sector and comparison with the Pareto/NBD formulation*. 7ème Congrès Tendances du Marketing Paris Venise (Venise).
- Dwyer, F. R. (1997). Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing*, 11(4), 6–13.
- Dwyer, R. (1989). Customer lifetime valuation to support marketing decision making. *Journal of Direct Marketing*, 3(4), 8–15.
- Dziurzynski, L., Wadsworth, E., & McCarthy, D. (2014). BTYD: Implementing buy 'Til You Die Models. R package version 2.4. <https://CRAN.R-project.org/package=BTYD>
- East, R., Hammond, K., Harris, P., & Lomax, W. (2000). First-store loyalty and retention. *Journal of Marketing Management*, 16(4), 307–325.
- Ehrenberg, A. S. C. (1959). The pattern of consumer purchases. *Applied Statistics*, 8(1), 26–41.
- Ehrenberg, A. S. C. (1988). *Repeat buying: Facts, theory and applications*. London: C. Griffin & Co..
- Fader, P., & Hardie, B. (2007a). *Incorporating time-invariant covariates into the Pareto/NBD and BG/NBD models*. Working Paper, University of Pennsylvania and London Business School.
- Fader, P. S., & Hardie, B. G. S. (2007b). How to project customer retention. *Journal of Interactive Marketing*, 21, 76–90.
- Fader, P., & Hardie, B. (2009). Probability models for customer-base analysis. *Journal of Interactive Marketing*, 23(1), 61–69.
- Fader, P., & Hardie, B. (2013). Overcoming the BG/NBD model's #NUM! error problem. <http://brucehardie.com/notes/027/>. Accessed 23 Feb 2015.
- Fader, P. S., Hardie, B. G. S., & Zeithammer, R. (2003). Forecasting new product trial in a controlled test market environment. *Journal of Forecasting*, 22, 391–410.
- Fader, P., Hardie, B., & Lok, L. K. (2005a). Counting your customers the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24(2), 275–284.
- Fader, P., Hardie, B., & Lok, L. K. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4), 415–430.
- Fudenberg, D., & Tirole, J. (2000). Customer poaching and brand switching. *RAND Journal of Economics*, 31(4), 634–657.
- Glady, N., Lemmens, A., & Croux, C. (2015). Unveiling the relationship between the transaction timing, spending and dropout behavior of customers. *International Journal of Research in Marketing*, 32(1), 78–93.
- Gupta, S., & Lehmann, D. R. (2003). Customers as assets. *Journal of Interactive Marketing*, 17(1), 9–24.
- Gupta, S., Lehmann, D. R., & Stuart, J. A. (2002). Valuing customers. *Journal of Marketing Research*, XLI, 7–18.
- Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of Marketing Research*, 41(1), 7–18.
- Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., & Sriram, N. (2006). Modeling customer lifetime value. *Journal of Service Research*, 9(2), 139–155.
- Hlavinka, K., & Sullivan, J. (2011). The art and science of building customer value. *Colloquy*.

- Hogan, J. E., Lemon, K. N., & Rust, R. R. (2002). Customer equity management: Charting new directions for the future of marketing. *Journal of Service Research*, 5(1), 4–12.
- Jackson, B. B. (1985). Build customer relationship that last. *Harvard Business Review*, 63, 120–128.
- Jain, D., & Singh, S. (2002). Customer lifetime value research in marketing: A review and future directions. *Journal of Interactive Marketing*, 16(2), 34–46.
- Jerath, K., Fader, P., & Hardie, B. (2011). New perspectives on customer death using a generalization of the Pareto/NBD model. *Marketing Science*, 30(5), 866–880.
- Kahn, B. E., & Schmittlein, D. C. (1989). Shopping trip behavior: An empirical investigation. *Marketing Letters*, 1(4), 55–70.
- Keane, T. J., & Wang, P. (1995). Applications for the lifetime value model in modern newspaper publishing. *Journal of Direct Marketing*, 9(2), 59–66.
- Kumar, V. (2007). Customer lifetime value: The path to profitability. *Foundations & Trends in Marketing*, 2(1), 1–96.
- Kumar, V., & Reinartz, W. (2006). *Customer relationship management: A databased approach*. Hoboken: Wiley.
- Leenheer, J., Bijmolt, T. H. A., van Heerde, H. J., & Smidts, A. (2007). Do loyalty programs enhance behavioral loyalty? A market-wide analysis accounting for endogeneity. *International Journal of Research in Marketing*, 24(1), 31–47.
- Lewis, M. (2005). Incorporating strategic consumer behavior into customer valuation. *Journal of Marketing*, 69, 230–238.
- Lewis, M. (2006). Customer acquisition promotions and customer asset value. *Journal of Marketing Research*, 43(2), 195–203.
- Libai, B., Narayandas, D., & Humby, C. (2002). Toward an individual customer profitability model: A segment-based approach. *Journal of Service Research*, 5(1), 69–76.
- Liu, Y. (2007). The long-term impact of loyalty programs on consumer purchase behavior and loyalty. *Journal of Marketing*, 71(4), 19–35.
- Meyer-Waarden, L. (2007). The effects of loyalty programs on customer lifetime duration and share of wallet. *Journal of Retailing*, 83(2), 223–236.
- Meyer-Waarden, L., & Benavent, C. (2009). Retail loyalty program effects: Self-selection or purchase behavior change? *Journal of the Academy of Marketing Science*, 3(3), 345–358.
- Morgan, R., & Hunt, S. (1994). The commitment-trust theory of relationship marketing. *Journal of Marketing*, 58(3), 20–38. doi:10.2307/1252308.
- Morrison, D., & Schmittlein, D. (1988). Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *Journal of Business and Economic Statistics*, 6, 145–159.
- Narasimhan, C. (1984). A price discrimination theory of coupons. *Marketing Science*, Spring 84, 3(2), 128–147.
- Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8, 343–366.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions*, A185, 71–110.
- Pfeifer, P. E., & Carraway, R. L. (2000). Modeling customer relationships as Markov chains. *Journal of Interactive Marketing*, 14(2), 43–55.
- R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, 64(4), 17–35.
- Reinartz, W., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, 67(1), 77–99.
- Reinartz, W. J., Thomas, J. S., & Kumar, V. (2005). Balancing acquisition and retention resources to maximize customer profitability. *Journal of Marketing*, 69, 63–79.

- Rogers, E. M. (2003). *Diffusion of innovations*. New York: Free Press.
- Rust, R. T., Lemon, K. A., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing*, 68(1), 109–127.
- Ryals, L. (2005). Making customer relationship management work: The measurement and profitable management of customer relationships. *Journal of Marketing*, 69(4), 252–261.
- Schmittlein, D., & Peterson, R. (1994). Customer case analysis: An industrial purchase process application. *Marketing Science*, 13(1), 41–67.
- Schmittlein, D., Morrison, D., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science*, 33(1), 1–24.
- Schulze, C., Skiera, B., & Wiesel, T. (2012). Linking customer and financial metrics to shareholder value: The leverage effect in customer-based valuation. *Journal of Marketing*, 76(2), 17–32.
- Schweidel, D. A., & Fader, P. S. (2009). Dynamic change points revisited. An evolving process model of new product sales. *International Journal of Research in Marketing*, 26(2), 119–124.
- Schweidel, D. A., Fader, P. S., & Bradlow, E. T. (2008). Understanding service retention within and across cohorts using limited information. *Journal of Marketing*, 72, 82–94.
- Simester, D. I., Sun, P., & Tsitsiklis, J. N. (2006). Dynamic catalog mailing policies. *Management Science*, 52(5), 683–696.
- Stauss, B., & Friege, C. (1999). Regaining service customers: Costs and benefits of regain management. *Journal of Service Research*, 1(4), 347–361.
- Tarasi, C., Bolton, R., Hutt, M., & Walker, B. (2011). Balancing risk and return in a customer portfolio. *Journal of Marketing*, 75(3), 1–17.
- Temporal, P., & Trott, M. (2001). *Romancing the customer: Maximising brand value through powerful relationship management*. New York: Wiley.
- Thomas, J., Blattberg, R., & Fox, E. (2004). Recapturing lost customers. *Journal of Marketing Research*, 38(2), 31–45.
- Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, 68(4), 106–125.
- Villanueva, J., & Hanssens, D. (2007). Customer equity: Measurement, management and research opportunities. *Foundations and Trends® in Marketing*, 1(1), 1–95.
- Villanueva, J., Yoo, S., & Hanssens, D. M. (2008). The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. *Journal of Marketing Research*, 45(1), 48–59.
- Wedel, M., Desarbo, W. S., Bult, J. R., & Ramaswamy, V. (1993). A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8(4), 397–411.
- Wiesel, T., Skiera, B., & Villanueva, J. (2008). Customer equity: An integral part of financial reporting. *Journal of Marketing*, 72, 1–14.



Assessing the Financial Impact of Brand Equity with Short Time-Series Data

Natalie Mizik and Eugene Pavlov

Contents

Introduction	1036
Marketing Academics' Views on the Measurement of Brand Equity	1037
Customer Mindset Brand Equity	1039
Product Market-Based Brand Equity	1040
Financial Market-Based Brand Equity	1041
Assessing Long-Term Impact of Brand Equity	1043
Empirical Illustration	1046
Total Financial Impact of Brand Asset	1048
Heterogeneity in Brand Equity Impact	1049
Conclusion	1051
References	1053

Abstract

In this chapter, we describe an approach to estimating the total long-term impact of brand perceptions on financial performance. The approach relies on modeling the stock market reactions to changes in brand perceptions and allows estimating their total impact even with limited time-series data. We present an application of the method to the Y&R Brand Asset Valuator (BAV) data. The analyses show that, on average, the bulk of brand impact on financial performance is realized in the future and the contemporaneous effects reflect only a small portion of the total impact. The analyses, however, also show considerable heterogeneity across industries: while in some industries the whole impact of brand asset occurs in current period only (restaurants), in other industries it occurs in future periods only (high-tech). Further, some components of consumer perceptions have differential effects in different industries. Returns to brand building, and to

N. Mizik (✉) · E. Pavlov
Foster School of Business, University of Washington, Seattle, WA, USA
e-mail: nmizik@uw.edu; epavlov@uw.edu

marketing efforts in general, should not be evaluated based on contemporaneous outcomes, but should rather be evaluated over a long-time horizon.

Keywords

Brand equity · Customer mindset · Financial impact · Heterogeneity · Dynamic panel · Instrumental variables

Introduction

Aaker (2012, p. 7) defines brand equity as a “set of assets (and liabilities) linked to a brand’s name and symbol that adds to (or subtracts from) the value provided by a product or service to a firm and/or that firm’s customers.” Brand equity stems from the ability of a brand to create awareness and favorable image in consumer minds. It allows the branded product to accrue extra profit over an extended period of time compared to a nonbranded product with comparable physical attributes. The benefits of strong brand equity can be observed in greater sales, higher profitability, or greater market valuation of a firm. There is, however, no easy method for assessing financial impact of a brand, and there is no comprehensive and unambiguous approach to the measurement of brand equity.

One reason a standardized approach to brand equity measurement and assessment of its impact on financial performance is lacking is that brand is an abstract construct, a mental structure of values, perceptions, and attitudes that resides in consumer minds (Pavlov and Mizik 2017). The process of brand equity formation is inherently psychological and is very complex (Keller 1993). Much work is still needed to understand the mental structure that represents a brand and to achieve consensus within academic and practitioner community on the concept and general model of brand equity.

Another complication impeding the development of a standardized tool for measuring brand equity stems from the fact that brand equity is often a product of long-term marketing effort. It takes a long time to build brand equity because the effect of marketing effort on consumer perceptions, associations, and attitudes is not immediate, but rather can take a long time to materialize. That is, there is a high level of persistence and inertia to brand equity.

Please consider the case of Martha Stewart Living Omnimedia, Inc. Following the 2002 scandal involving Martha Stewart’s sale of ImClone stock – which prompted insider trading and perjury investigations by the SEC and FBI –both the Martha Stewart brand perceptions and the stock price of Martha Stewart Living Omnimedia plummeted (Fig. 1). The negative impact of brand damage on sales and profits, however, took several years to manifest itself (Fig. 2) and in the long run neither the brand perceptions nor firm performance ever fully recovered to the prescandal levels. The Martha Stewart case shows that contemporaneous accounting performance metrics (such as sales or operating income) can severely underestimate the full impact of a brand.

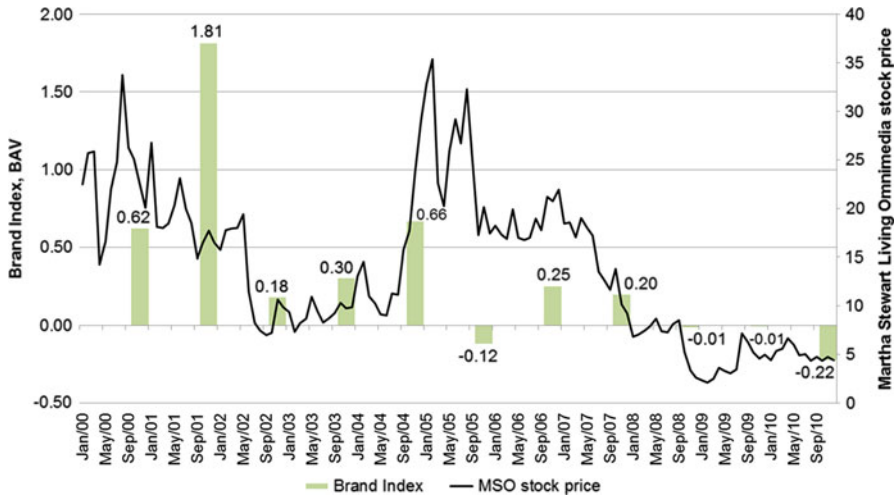


Fig. 1 Martha Stewart Living Omnimedia, Inc. stock price and Brand Perceptions of Martha Stewart brand (Brand Index) Brand Index is computed as a z-standardized equally weighted average of perceived brand Differentiation, Relevance, Esteem, Knowledge, and Energy over the sample of publicly traded monobrand firms in the BAV database in the 2000–2010 period

In this chapter, we describe an approach to estimating the total long-term impact of brand perceptions on financial performance. The approach relies on modeling the stock market reactions to changes in brand perceptions and closely follows Mizik (2014). Just as the case of Martha Stewart’s scandal suggests, the analyses show that the bulk of brand impact on financial performance is realized in the future and the contemporaneous effects significantly underestimate the total impact. The analyses also show considerable heterogeneity across industries: while in some industries the whole impact of brand assets occurs in the current period only (restaurants), in other industries it occurs in future periods only (high-tech). Further, different components of consumer perceptions have differential effects across industries.

Marketing Academics’ Views on the Measurement of Brand Equity

Academic researchers of brand equity have approached the construct from different viewpoints and proposed various metrics and methods for assessing brand equity. Keller and Lehmann (2003) suggest that brand equity can be measured at three different levels: customer mindset, product market, and financial market. Customer-mindset approach to measuring brand equity stems from the psychological value consumers attach to a branded product, and focuses on assessing two major constructs of brand awareness and brand image. This method primarily relies on consumer surveys. Product-market approach to measuring brand equity (e.g., Kamakura and Russel 1993; Ailawadi et al. 2003; Srinivasan et al. 2005) evaluates

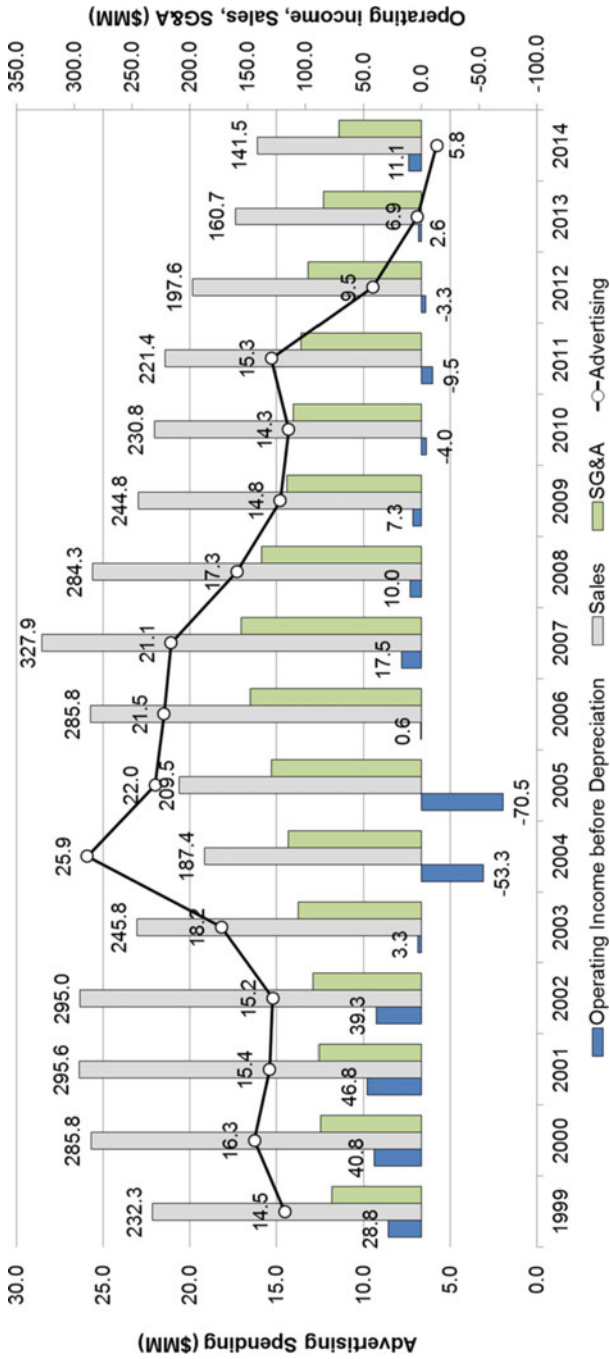


Fig. 2 Martha Stewart Living Omnimedia, Inc. operating performance indicators 1999–2014

incremental preference for a branded compared to a nonbranded product which manifests in incremental market share or price, revenue, or profit premium for the branded product. This approach combines survey-based methods with secondary data and makes use of conjoint analysis or purchase histories and scanner panel data. Financial-market-based approach (e.g., Simon and Sullivan 1993; Aaker and Jacobson 1994, 2001; Mizik and Jacobson 2009) aims to estimate incremental value (cash flows) arising from brand assets. Under this method, survey data is combined with secondary data from the stock market and accounting/financial statements to estimate the total value brand asset is expected to generate for its owner in the long run.

Customer Mindset Brand Equity

From a customer mindset perspective, positive brand equity exists when “consumer is familiar with the brand and holds some favorable, strong, and unique brand associations in memory” (Keller 1993, p. 2). Measuring brand equity from a consumer mindset perspective means dealing with familiarity and associations – constructs which are highly subjective and not directly observable. The abstract nature of consumer mindset-based brand equity gave rise to various proposals on what brand equity components are and what metrics are appropriate (e.g., Pappu et al. 2005; Lassar et al. 1995; Yoo and Donthu 2001). Most discussions of customer mindset-based brand equity center on constructs of brand familiarity and brand associations customers hold. Most of the proposed measurement approaches are, naturally, based on surveys and questionnaires.

Keller and Lehman (2003, p. 27), for example, recognize five components of customer mindset-based brand equity: awareness, associations, attitudes, attachment, and activity. Aaker (1996) advocates the idea of “Brand Equity Ten” – ten indicators contributing to brand asset value. Of the five subcategories (awareness, associations/differentiation, perceived quality/leadership, loyalty, market behavior), the first four are survey-based measures. Aaker (2012) notes that survey-based measures “can be expensive, inconvenient, time consuming, hard to implement/interpret.” Indeed, they are. With advent of the Web 2.0, however, it is becoming cheaper and easier to collect some customer mind-set branding data (Lee and Bradlow 2011; Netzer et al. 2012; Liu et al. 2017). These new approaches are often based on natural language processing techniques. They allow compiling relevant customer mindset-based metrics bypassing costly traditional surveying.

The various approaches to assess customer mindset-based brand equity can be classified into two broad categories: direct and indirect. Surveys fall into the category of “direct” measurement of customer mindset brand equity. Direct approaches also include studying customer response to marketing activities through experiments: the treatment group is exposed to marketing actions attributed to a particular brand, while control group is exposed to marketing actions attributed to a generic or unknown brand (Keller 1993, p. 13). “Blind” tests are examples of such experiments (Allison and Uhl 1964). Conjoint analysis is another direct method,

which allows to rank relative importance of product/service attributes/features based on customers' stated preferences and perceptions and to estimate price premium due to the brand (Rangaswamy et al. 1993). Cobb-Walgren et al. (1995, p.35), for example, conduct a conjoint analysis for the hotel industry and find that brand name is the fourth most important attribute after price, bed size, and availability of a pool. "Indirect" approach includes association tasks, interpretation of imagery, brand personality descriptors, etc. (Keller 1993, p. 12), and is far less suitable for translating into financial value.

Some brand perceptions and attitudes are collected through large-scale consumer surveys. Brand Asset Valuator by Young and Rubicam, EquiTrend by Harris Poll, Millward Brown's Brand Z are the main industry providers in this area. While brand attitude and perceptions data can be very valuable for brand management per se, they do not offer direct insight into the financial value of the brand. These perceptions and attitude data, however, can be used as a building block in the development of financial market-based measures of brand equity. We discuss these methods later in this chapter in more detail.

Product Market-Based Brand Equity

Under the product-market approach, brand equity is understood as the incremental value (e.g., in terms of market share, price, revenue, or profit) a branded product generates compared to its nonbranded analogue. For instance, in Park and Srinivasan (1994, p. 273), brand equity is conceptualized as an "incremental preference endowed by the brand to the product as perceived by an individual consumer." The authors suggest that brand equity stems from the difference between the overall product preference and objective preference, which is based on attribute-by-attribute evaluation. They use individual-level secondary data on actual in-store purchases to estimate this difference. Based on an empirical application of the model to the market of toothpaste and mouthwash, the authors found substantial effects of brand's equity on market share. For example, 12.2% of Colgate's 21.8% market share was attributed to brand equity and, compared to the store brand, Colgate-Palmolive was able to charge an estimated 37 cents more. The measures of brand equity obtained using this approach are relative: they do not reflect the absolute value of brand equity for a given brand. Rather, they provide an estimate of brand equity in comparison to another brand.

Srinivasan et al. (2005, p. 1433) define brand equity as an "incremental contribution per year obtained by the brand in comparison to the same product (or service) at the same price but without brand-building efforts (i.e., base product)." Using individual-level data the authors estimate incremental choice probability of a branded versus a base product for an individual customer. Three distinct sources of brand equity value were considered: brand awareness, nonattribute preference, and enhanced attribute perceptions (i.e., attribute perceptions compared to the base product's attribute perceptions). Using consumer data from the mobile phone market in South Korea, the authors obtain estimates of brand equity and its components for

four leading companies. Brand equity of Samsung is estimated to provide 34.8% of Samsung's 52.5% market share. Total value of brand equity obtained from incremental choice probabilities is estimated at \$127 MM for Samsung, \$69 MM for LG, \$32 MM for Motorola, and \$9 MM for Qualcomm (p. 1445). As for components of brand equity, awareness was found to have the strongest effect, followed by non-attribute preference and enhanced attribute perceptions.

Ailawadi et al. (2003) propose revenue premium – the difference in revenue between a branded good and a corresponding private label – as a measure of brand equity. The advantage of this measure is in simultaneously capturing the effects of brand equity on both price and volume. Studying a major grocery retailer in 1991–1996, the authors calculated yearly revenue premia for 111 brands and documented a median decrease in revenue premia of 11%. One of the challenges for the revenue premium calculation lies in identifying an appropriate benchmark brand or generic to compare prices and volumes with. That is, the revenue premium is a relative and not an absolute measure of brand equity as store brands are brands in their own right. Another challenge is the proper market definition for a particular brand. The revenue premium measure reflects the effect of competition. That is, the same brand might have a much greater revenue premium in the market where few competitors are present versus the market where there are many options for consumers to choose from. Other concerns with the revenue premium model have been noted in the literature and include its failure to account for the costs of brand management and maintenance and the lack of temporal dimension (only contemporaneous effects on sales are captured). For example, Srinivasan et al. (2005) suggest that profit premium would be a better measure of brand equity and advocate for assessing long-term brand-induced incremental profits.

Product-market models for estimating brand equity value provide important insights by leveraging secondary data (often, scanner panel data of actual purchase histories). They allow attributing observed differences in market shares and prices charged by producers in the same product category to psychological value consumers derive from choosing a particular brand. The limitations of product market-based models stem from (1) the relative nature of brand equity estimates (i.e., estimate of brand equity is defined in comparison to another brand) (2) subjectivity involved in the choice of an appropriate benchmark and/or a market definition, and (3) lack of temporal consideration for the brand effects. As the case of Martha Stewart illustrates, a large portion of the brand value might be missing in contemporaneous (same-year) product market performance metrics.

Financial Market-Based Brand Equity

The financial market-based approach to valuing brands views brands as assets capable of generating stream of profits over a long period of time. Financial market-based valuation of brand equity is “forward-looking” (Simon and Sullivan 1993, p. 32) in the sense that it reflects the sum of the discounted incremental future cash flows attributable to the brand. Under this view, the value of brand assets is a

portion of the company's stock market capitalization. As such, any changes in the brand assets will be reflected in the market valuation of the company.

Accounting Valuation of Brands

It is important to highlight the distinction between the market-based and accounting-based valuation of brands. Internally generated intangibles, including brands, are not included on the balance sheet. Exclusion of brands and other intangibles from the balance sheet gives rise to sizeable discrepancies between the book value and the market value of a company. For example, Sinclair and Keller (2014, p. 294) discuss the \$200 billion market capitalization of Procter & Gamble at the time when its net tangible assets are a negative \$18.7 billion.

Although internally generated brand assets are not on the balance sheets, brand assets acquired in business combinations (e.g., acquisitions) are recognized as assets and placed on the balance sheet of the acquirer (Cañibano et al. 2000; Austin 2007). Various valuation methods are used to arrive at the value of acquired brands. The most commonly used approaches employ earnings split (estimating the portion of the earnings attributable to the brand and projecting future earnings and discount factors) and relief from royalty (estimating the "royalty savings" from owning a brand based on a set of comparable brands, where royalty structure is known, and projecting future earnings) analyses. Both types of analyses involve significant subjective judgment in attributing earnings or selecting comparables and projecting future earnings and discount factors. Bahadir et al. (2008) report that the recognized value of brands in M&A transactions varies widely (it ranged from 1.16% to 49.7% of the transaction value) and comment that the value of a brand lies "in the eye of the beholder."

Financial Market Value-Relevance of Brands

Simon and Sullivan (1993, p. 29) define brand equity as "the incremental cash flows which accrue to branded products over and above the cash flows which would result from the sale of unbranded products." They criticize product market-based metrics of assessing brand equity such as price premium. Price premium method (1) does not account for brand's ability to reduce marketing costs in future periods and (2) it might be confounded with high-quality product attributes, resulting in biased estimates of brand equity value. Simon and Sullivan (1993) estimate that brand equity accounts for 19% of tangible asset value for the 638 firms in their sample.

Barth et al. (1998) analyzed stock market valuation of Interbrand's brand value measures. Controlling for fiscal year fixed effect, book value of equity per share, earnings per share from continuing operations, the authors found a significant association between stock price at the end of fiscal year and the Interbrand's brand value estimate. Madden et al. (2006) used World's Most Valuable brands (WMVB) ranking by Interbrand to compare performance of companies owning highly valuable brands to companies that do not. They construct a portfolio of 111 companies on the WMVB list and compare its performance against a benchmark portfolio comprised of all other companies in the CRSP database. The WMVB portfolio outperformed the benchmark and delivered higher returns with significantly smaller

risk, as measured by market Beta (the coefficient on market return in Fama-French model). A concern that has been raised with these analyses relates to the fact that Interbrand and other providers are not fully transparent on how brand values are calculated and that they use market capitalization of the firm as one of the inputs in their calculations of brand value (Hrustic 2012).

A few studies have lined customer mindset-based measures of brand equity directly to company stock performance. Aaker and Jacobson (1994) study the association between perceived product quality (EquiTrend) and stock returns. The authors find that perceived quality is significantly associated with stock market returns and has incremental explanatory power over ROI (profitability). Aaker and Jacobson (2001) apply a similar approach to a survey-based brand attitude measure (positive/neutral/negative) for a high-tech company (Techtel) and find a significant positive effect of brand attitude on abnormal stock return.

Mizik and Jacobson (2008) used Young & Rubicam Brand Asset Valuator (BAV) data to assess the financial value relevance of perceptual brand attributes. Based on eight waves of a large-scale annual customer survey, they examined five pillars of brand perceptions (differentiation, relevance, esteem, knowledge, and energy) to assess their incremental information content. Of the five pillars, relevance and energy were found to be significantly positively associated with abnormal stock returns and no contemporaneous effect of differentiation, esteem, or knowledge on stock returns has been found. However, the authors detected a significant effect of prior year change in differentiation on unanticipated changes in earnings, which suggested the existence of a market anomaly: past changes in brand differentiation predicting current abnormal returns. Additional analyses revealed a significant difference in mean abnormal stock returns for companies with brands which gained in perceived brand differentiation in the prior year versus companies with decreased brand differentiation in prior year and attributed this anomaly to the lack of transparency (private information) in brand strategy.

Most empirical studies employing financial market-based approach have focused on assessing the value relevance and incremental information content of various brand metrics (see Mizik and Jacobson (2009) for an exception and an illustration of comparables-based valuation approach to valuing brand assets). Mizik (2014) proposed extending the method to explicitly address their total *long-term* financial impact. We discuss the theoretical and empirical foundations of this approach and present an empirical illustration below.

Assessing Long-Term Impact of Brand Equity

Brands have both contemporaneous and delayed effects on firm profitability. Figure 3 depicts the dynamic framework of brand financial impact. γ_0 is the contemporaneous effect of brand asset on earnings. It reflects both the costs associated with developing the brand asset at time t and the realized incremental revenue which accrued due to the brand asset at time t . γ_0 can be positive or negative, depending on whether the costs or the incremental revenue effect dominates. λ_0 is the contemporaneous impact

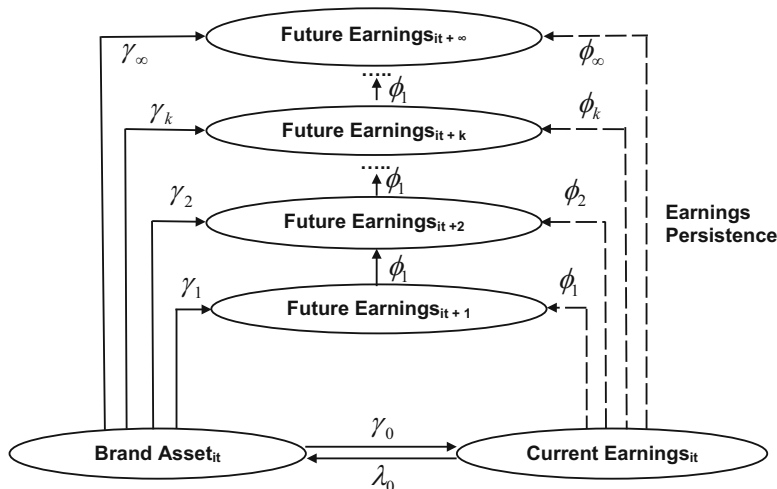


Fig. 3 Dynamic performance impact of brand assets

of earnings on brand asset. It too can be positive (if the firms increase brand asset building effort when profits increase) or negative (if the firms increase brand asset building when profitability falters).

γ_1 represents the impact of brand asset developed at time t on earnings in period $t+1$. Because the brand asset was developed in time period t (i.e., the development cost occurred in period t), γ_1 is nonnegative for value-generating brand assets (it is negative if the brand asset is value destroying, like poor reputation). γ_k represents the delayed impact of brand asset and is the *direct* impact of Brand Asset (t) on earnings in period $(t+k)$.

The total long-term impact of brand asset on profitability, however, exceeds the sum of its direct impacts described above. Because earnings persist (in Fig. 3, the dynamic coefficients ϕ_k indicate that current level of earnings depends on previous period earnings), a portion of contemporaneous impact of brand asset is carried over to future periods through earnings dynamics. A shock to earnings in period t is partially carried over to period $(t+1)$. The indirect effect of brand asset on earnings in period $t+1$ equals to $\gamma_0 * \phi_1$ and the total effect in period $t+1$ is equal to $\gamma_1 + \gamma_0 * \phi_1$. The total long-term impact of Brand Asset (t) on firm financial performance is the aggregate sum over all direct and indirect effects.

The estimation of the total impact of brand asset on profitability as depicted in Fig. 3 with standard distributed lag panel data models is typically not feasible because time series of branding data are often limited. However, under the assumption of efficient markets, the stock market-based approach can be implemented even with limited time-series data.

Under the hypothesis of financial markets efficiency (Fama 1970), the stock market value of a firm incorporates all information and rational expectations of a company's future financial performance. Unexpected changes in firm's brand assets

lead to changes in expected future cash flows and induce investors to recalculate company valuation. This change in firm valuation serves as an unbiased estimate of the total impact the change in brand asset is expected to generate in the long run.

Figure 4 summarizes the estimation framework. The interpretation of γ_0 and λ_0 coefficients remains the same as in Fig. 3. The future impact of brand asset (represented by coefficients γ_k , with $k > 0$, in Fig. 3) is now captured in the β_2 coefficient. The framework depicted in Fig. 4 generates the following two estimation equations:

$$StkRet_{it} = Eret_{it} + \beta_1 \Delta ROA_{it} + \beta_2 \Delta BrandAsset_{it} + \epsilon_{1it}, \tag{1}$$

$$\Delta ROA_{it} = \gamma_0 \Delta BrandAsset_{it} + \epsilon_{2it}, \text{ where} \tag{2}$$

$StkRet_{it}$ is the stock return for firm i at time t , $Eret_{it}$ is the expected return, ΔROA_{it} is the unanticipated change in size-adjusted earnings, $\Delta BrandAsset_{it}$ is the unanticipated change in the brand asset, ϵ_{1it} and ϵ_{2it} are i.i.d. normal error terms.

Several issues arise with the estimation of Eqs. 1 and 2. First, one needs to obtain the unanticipated components of ROA_{it} and $BrandAsset_{it}$ series. Second, γ_0 in Eq. 2 could be estimated consistently only under no simultaneity condition (i.e., no feedback from ΔROA_{it} to $\Delta BrandAsset_{it}$). That is, only if λ_0 is equal to zero.

Estimating Eqs. 1 and 2 generates both contemporaneous (γ_0) and long-term direct (β_2) impact of brand asset on financial performance. The indirect effect of $\Delta BrandAsset_{it}$ on future earnings occurs because the contemporaneous impact of $\Delta BrandAsset_{it}$ on ΔROA_{it} is transferred through the earnings-response coefficient β_1 .

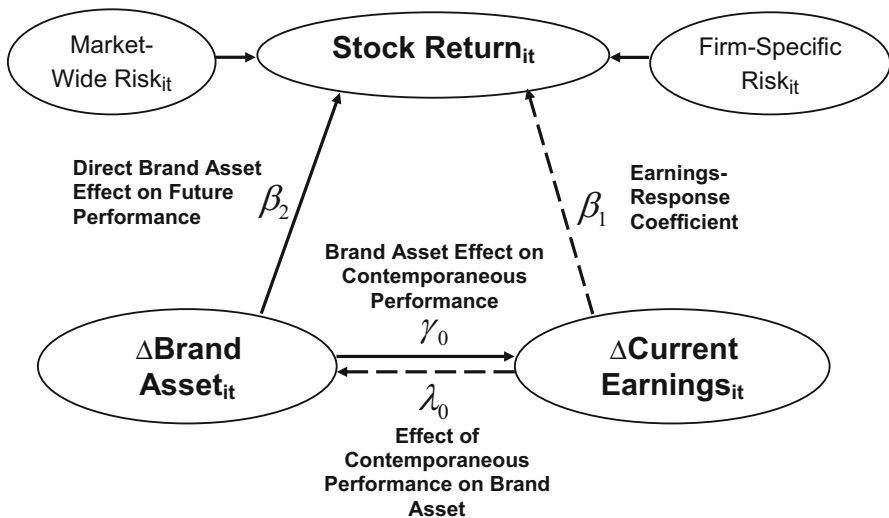


Fig. 4 Estimating framework for assessing the total financial impact of brand asset using the stock market-based approach

This indirect effect of brand asset on earnings equals to $\gamma_0 * \beta_1$. Hence, the total financial impact of brand asset on profitability is equal to the sum of direct future effect and the indirect effects: $\beta_2 + \gamma_0 * \beta_1$.

Empirical Illustration

The empirical application below draws on several data sources. Stock return data and the Fama-French-Carhart risk factors come from the Chicago's Center for Research in Security Prices (CRSP). The accounting data come from the quarterly Compustat database. Profitability (ROA) is measured as operating income before depreciation over assets. Brand perceptions data come from the 2000–2010 Y&R's BAV database. It includes 444 monobrand publicly traded firms. The measure of Brand Asset Index is a z-standardized index comprised of five brand perceptions of (1) differentiation, (2) relevance, (3) esteem, (4) knowledge, and (5) energy. The perceptual constructs used in constructing the Brand Asset Index and their measurement are discussed in detail in Mizik and Jacobson (2008).

Calculating Abnormal Stock Returns

Abnormal stock returns are calculated using Fama-French (1993) model augmented with the momentum factor (Carhart's 1997) and with risk characteristics of size and book-to-market (Daniel and Titman 1997), but findings are robust to alternative definitions of abnormal returns:

$$\begin{aligned} Ret_{it} - Ret_{riskfree,t} = & a_t + \beta_{mkt}(Ret_{mkt,t} - Ret_{riskfree,t}) + s(SMB_t) + h(HML_t) \\ & + m(MOM_t) + \eta_t Size_{it-1} + \nu_t BookMarket_{it-1} \\ & + u_{it}, \text{ where} \end{aligned} \quad (3)$$

Ret_{it} is stock return of firm i at period t , $Ret_{riskfree,t}$ is risk-free return at time t , $Ret_{mkt,t}$ is market return at time t , SMB_t is difference between large and small book-to-market ratio firms at period t , HML_t is difference between high and low capitalization firms at period t , MOM_t is Carhart's (1997) momentum at period t , $Size_{it-1}$ is firm-specific risk characteristic of size, calculated as log of lagged market value, $BookMarket_{it-1}$ is firm-specific risk characteristic of book-to-market, calculated as log of lagged book value over market value. Abnormal stock returns are the residuals from this regression, u_{it} .

Calculating Unanticipated Components of ROA and Brand Asset Index

Estimating Eqs. 1 and 2 requires computing a measure of unanticipated change in size-adjusted earnings or "earnings surprise" (ΔROA_{it}). Because ROA series exhibit significant persistence and seasonality, a four-quarter autoregressive fixed effects model is used to approximate the data-generating process:

$$ROA_{iq} = a_i + \sum_{k=1}^4 \phi_k ROA_{iq-k} + \sum_{q=1}^Q \delta_q \times Time_q + e_{iq} \quad (4)$$

a_i is the firm-specific indicator, $Time_q$ is the time period (year-quarter) indicator, and ϕ_k is the quarter k autoregressive coefficient. a_i and ϕ_k can be consistently estimated using Anderson-Hsiao (1982) or Arellano-Bond (1991) instrumental variables approach. First, the data is recomputed in terms of deviations from year-quarter-specific means (to cancel out time-period fixed effects δ_q). Then, first differences of series are taken to remove fixed effects a_i . Lagged values of ROA serve as instrumental variables for the first lag of ΔROA (ROA_{iq-2} and ROA_{iq-3} instrument for $[ROA_{iq-1} - ROA_{iq-2}]$).

Table 1 reports the results of this estimation. The 1st-order and the 4th-order lags of ROA have the highest autoregressive coefficients indicating the presence of strong seasonality. The measure of annual earnings surprise (ΔROA_{it}) is computed as a sum of prediction errors ($ROA_{iq} - \hat{ROA}_{iq}$) over the four quarters in a given year.

With short time series of branding data it is often impossible to use the same procedure to estimate unanticipated components of BAIindex. However, instead of modeling dynamics of the marketing metric directly, one can evaluate the stock market's beliefs about its dynamics. The market beliefs about the dynamic properties of brand equity metric can be assessed with the following model:

$$AbnormalStkRet_{it} = \beta_1 \Delta ROA_{it} + \beta_2^* (BAIndex_{it} - \beta_0 BAIindex_{(it-1)}) + \epsilon_{it}, \text{ where} \tag{5}$$

ϕ_0 is the persistence in brand index series.

Equation 5 could be rewritten as

$$AbnormalStkRet_{it} = \beta_1 \Delta ROA_{it} + \beta_2^* BAIindex_{it} + \beta_2^0 BAIindex_{it-1} + \epsilon_{it} \tag{6}$$

A finding of $\beta_2^* + \beta_2^0 = 0$ would suggest that ϕ_0 is equal to one and that $BAIndex_{it}$ contains a unit root. Because that is the case in our data sample, the surprise in brand asset index can be defined as the first difference of Brand Asset Index series: $\Delta BAIindex_{it} = BAIindex_{it} - BAIindex_{it-1}$. If no unit root in $BAIndex_{it}$ is detected, then the surprise in brand asset index can be computed as $\Delta BAIindex_{it} = BAIindex_{it} - \phi_0 BAIindex_{it-1}$, where ϕ_0 is the estimate from Eq. (5). In situations when sufficient time series of a marketing metric are available, the surprise can be

Table 1 Fourth-order autoregressive fixed-effects instrumental variable regression for ROA

	Estimate	SE	t-statistic
ϕ_1^a	.28131**	.02397	11.74
ϕ_2	.05048**	.01298	3.89
ϕ_3	.01645*	.00989	1.66
ϕ_4	.66795**	.00508	131.39
Number of obs	22,526		
F-statistic	6509.75		

* $p < .10$; ** $p < .01$

^aDenotes the use of IV estimation

computed following the approach described above for calculating ΔROA_{it} : [1] estimate an autoregressive panel model for the marketing metric, [2] calculate the residuals from the autoregressive model. These residuals then serve as ΔBAI_{it} .

Assessing the Presence of Simultaneity

Because we are working with short time series of BAI_{it}, we cannot address potential simultaneity bias in γ_0 directly (an appropriate instrument cannot be constructed due to short time series of BAI_{it}). But we can assess potential simultaneity between earnings (ROA) and BAI_{it} by estimating the following dynamic panel models because the ROA time series are sufficiently long such that an instrument for ROA can be constructed:

$$BAI_{it} = a_{BA,i} + \lambda_0 ROA_{it} + \lambda_1 ROA_{it-1} + \lambda_2 ROA_{it-2} + \delta_t Year_t + \kappa_{1,it} \quad (7)$$

$$BAI_{it} = a_{BA,i} + \lambda_0 ROA_{it} + \lambda_1 ROA_{it-1} + \lambda_2 ROA_{it-2} + \phi_1 BAI_{it-1} + \phi_2 BAI_{it-2} + \delta_t Year_t + \kappa_{2,it} \quad (8)$$

Here, λ_0 measures the contemporaneous impact of earnings on brand perceptions. A finding of $\lambda_0 = 0$ would suggest that no simultaneity is present (that is, earnings do not have a contemporaneous effect on brand perceptions). Estimating Eqs. 7 and 8, again, requires taking the first differences of the data to remove the fixed effects and using the instrumental variables approach. The estimated coefficient $\hat{\lambda}_0$ is small and insignificant in both models. Estimating model Eq. 7 generates $\hat{\lambda}_0 = 0.98$ (SE=9.90) and $\hat{\lambda}_0 = -15.55$ (SE=12.69) in model Eq. 8. As such, we find no evidence of feedback from contemporaneous earnings to Brand Asset Index.

If $\lambda_0 \neq 0$, OLS estimate of γ_0 in Eq. 2 will be biased and inconsistent due to presence of feedback loop effect. We refer the reader to Chap. 15 in Greene (2002) for the discussion on consistent estimation in simultaneous-equations models.

Total Financial Impact of Brand Asset

The total financial impact of Brand Asset Index can be assessed by estimating Eq. 2 and Eq. 9 below. Because no simultaneity is present in our sample, the total financial impact can also be estimated directly after substituting Eq. 2 into Eq. 9 to obtain estimating Eq. 10:

$$AbnRet_{it} = \beta_1 \Delta ROA_{it} + \beta_2 \Delta BAI_{it} + e_{1it} \quad (9)$$

$$AbnRet_{it} = \psi \Delta BAI_{it} + e_{2it}, \text{ where } \psi = \beta_2 + \gamma_0^* \beta_1 \quad (10)$$

The results of estimating Eqs. 9 and 10 are presented in Table 2. The information content in BAI_{it} is significantly and positively associated with abnormal stock return of a firm. It is positive and significant in both formulations – with and without including ROA information in the estimating equation. The estimates reported in Table 2 indicate that Brand Asset Index has a direct impact on stock returns of 0.068,

Table 2 Direct future and total performance impact of the BAIndex^a

	Model Eq. 8	Model Eq. 9
Unanticipated change in BAIndex	.068** .080 (.018)	.07568** .089 (.019)
Unanticipated change in ROA	3.18** .291 (.235)	
F-statistic	99.96	15.67
N	1956	1956

p* < .10; *p* < .01

^aStandardized regression coefficients are in *italics*, standard errors are in *parentheses*

which is incremental to ROA. The total impact of Brand Asset Index on stock return is 0.07568.

Based on the estimates reported in Table 2, one can ascertain the dynamics of Brand Asset Index impact on stock return and break down the total impact into its direct and indirect components. Difference in the estimates of BAIndex coefficient in Eqs. 9 and 10 is 0.07568–0.068=0.008. This difference is the indirect effect of Brand Asset Index on abnormal returns (which occurs through earnings persistence). That is, the immediate impact of Brand Asset Index on ROA is equal to 0.0025 (0.008/3.184). This suggests that only approximately 3.3% (0.0025/0.07568) of Brand Asset Index impact is realized in contemporaneous operating income, and the bulk of the impact occurs in future periods.

Heterogeneity in Brand Equity Impact

Sector-Specific Differences in the Impact of Brand Asset Index

One interesting research question remains: are the effects of Brand Asset uniform across different industrial sectors? Results of sector-specific analyses suggest that there exists a significant heterogeneity in the dynamics of the Brand Asset Index impact across different industrial sectors. Table 3 presents the results of estimating models Eq. 9 and 10 in six sectors: financial, distribution/retail, restaurants, computer/Internet, pharmaceuticals, and travel/entertainment.

We observe no significant findings for BAIndex for either finance or for travel/entertainment sectors. For distribution/retail sector, the estimates of BAIndex are largely in line with the aggregate findings reported in Table 2: the immediate impact of BAIndex on ROA is small. It can be calculated as (0.101–0.078)/4.125=0.006, which is about 5.5% of the total impact. That is, approximately 95% of the effect of Brand Asset is realized in the future. The pattern for the high-tech sectors (computer/Internet and pharmaceuticals) suggests that all of the BAIndex effect is realized in the future periods. This finding implies that a success or failure of brand-building efforts in these industries cannot be determined using contemporaneous product-market performance measures. Finally, in the restaurant sector, all of the effect of

Table 3 Differential impact of brand asset index by sector

	Finance	Distribution/ Retail	Computer/ Internet	Pharmaceutical	Restaurants	Travel/ Entertainment
<i>Panel A. Direct future impact of the BAIndex (equation model 9)</i>						
Unanticipated change in the BAIndex	0.036 (0.092)	0.078* (0.041)	0.121*** (0.045)	0.141*** (0.050)	0.083 (0.093)	0.116 (0.135)
Unanticipated change in ROA	5.607*** (1.792)	4.125*** (0.565)	2.236*** (0.471)	2.110*** (0.86)	5.968*** (1.103)	3.669 (2.335)
F-stat	5.15	29.79	14.43	6.60	17.06	1.37
N obs	172	379	420	61	106	91
<i>Panel B. Total performance impact of the BAIndex (equation model 10)</i>						
Unanticipated change in the BAIndex	0.066 (0.094)	0.101** (0.043)	0.113** (0.046)	0.144*** (0.060)	0.199* (0.102)	0.069 (0.132)
F-stat	0.49	5.5	6.06	6.58	3.82	0.27
N obs	172	379	420	61	106	91

* $p < .10$; ** $p < .05$; *** $p < .01$

Standard errors in parentheses

BAIndex (0.199) comes from its effect on contemporaneous earnings and no direct future effect exists. That is, if branding initiatives did not generate immediate (same-year) benefits, there is unlikely to be any benefit in the future periods either.

Component-Specific Differences in the Impact of Brand Perceptions

The insignificant effects of BAIndex for Finance and Travel/Entertainment sector might be suggesting that brand perceptions are not value-relevant in these sectors. Alternatively, the insignificant effects might be masking the heterogeneous impact of the individual perceptual components of Differentiation, Relevance, Esteem, Knowledge, and Energy in these sectors.

Table 4 reports disaggregate analyses for the individual components of Brand Asset Index in six different industrial sectors. Interestingly, three of the five perceptual components in Brand Asset Index – Differentiation, Relevance, and Energy – are significant in the Finance sector. Because they have opposite signs (increases in Differentiation have a negative effect while increases in Relevance and Energy have a positive effect on stock return), their combination in the aggregate Brand Asset Index is not significant. A similar picture emerges in the Travel/Entertainment sector: Differentiation has a marginally negative effect while Esteem and Energy have marginally positive effects. The effect of Relevance is also highly significant in the Computer/Internet and Restaurants sectors and is marginally significant in the Distribution/Retail sector.

There is less consistency in the effects of other Brand Asset Index components across industrial sectors. While Differentiation has a negative effect in the Finance and Travel/Entertainment sectors, it has a positive effect in the Pharmaceutical sector. Esteem has a marginally positive effect in the Distribution/Retail and Travel/

Table 4 Disaggregate analyses: differential impact of brand perceptions by sector Dependent variable: abnormal stock return

	Finance	Distribution/ Retail	Computer/ Internet	Pharmaceutical	Restaurants	Travel/ Entertainment
<i>Panel A. Direct future impact of the BAIndex (equation model 9)</i>						
ΔDifferentiation	-0.100* (0.061)	0.010 (0.025)	-0.007 (0.030)	0.084* (0.044)	0.011 (0.060)	-0.197* (0.117)
ΔRelevance	0.197* (0.109)	0.109* (0.060)	0.199*** (0.064)	0.033 (0.070)	0.310** (0.130)	0.053 (0.198)
ΔEsteem	-0.000 (0.087)	0.075* (0.046)	0.039 (0.059)	0.004 (0.074)	-0.092 (0.110)	0.255* (0.147)
ΔKnowledge	-0.129 (0.190)	-0.101 (0.096)	-0.095 (0.113)	0.100 (0.135)	-0.145 (0.180)	0.107 (0.356)
ΔEnergy	0.110* (0.067)	-0.019 (0.030)	0.024 (0.027)	0.011 (0.042)	0.004 (0.067)	0.143* (0.087)
Unanticipated change in ROA	5.103*** (1.889)	4.132*** (0.564)	2.379*** (0.473)	1.971** (0.890)	6.152*** (1.126)	3.22 (2.389)
F-stat	2.74	11.01	6.23	2.43	6.55	1.48
N obs	172	379	420	61	106	91
<i>Panel B. Total Performance Impact of the BAIndex (equation model 10)</i>						
ΔDifferentiation	-0.123** (0.062)	0.017 (0.027)	-0.007 (0.031)	0.088* (0.045)	0.020 (0.068)	-0.216* (0.116)
ΔRelevance	0.230** (0.110)	0.105* (0.064)	0.159*** (0.065)	0.051 (0.072)	0.307** (0.155)	0.012 (0.196)
ΔEsteem	0.067 (0.085)	0.086* (0.049)	0.049 (0.061)	0.003 (0.076)	0.062 (0.122)	0.224 (0.146)
ΔKnowledge	-0.229 (0.190)	-0.110 (0.102)	-0.083 (0.116)	0.100 (0.140)	-0.156 (0.208)	0.208 (0.350)
ΔEnergy	0.116* (0.068)	-0.010 (0.032)	0.025 (0.028)	-0.001 (0.043)	0.017 (0.077)	0.141 (0.088)
F-stat	1.76	2.16	2.28	1.8	1.47	1.39
N obs	172	379	420	61	106	91

* $p < .10$; ** $p < .05$; *** $p < .01$

Standard errors in parentheses

Entertainment sectors only. Energy has a marginally positive effect in the Finance and Travel/Entertainment sectors only. Interestingly, while most models of brand equity focus on the construct of brand familiarity, there is no significant positive association detected between the stock returns and the measure of familiarity (Knowledge).

Conclusion

Brand assets are crucial to firm performance and are a significant component of firm value. They are, however, difficult to value and to quantify in financial terms. The key reasons valuation of brand assets is difficult are the following:

- (1) *They are intangible.* They reside in consumer minds and are represented by a mental structure of perceptions and attitudes. Individual components of brand associations and attitudes comprising customer mindset brand equity have been proposed and studied, but the research in this domain is not complete. At present, there is no general agreement on what exactly this mental structure is. Research continues into which components are relevant and which components are not relevant in customer mindset-based brand equity.
- (2) *They are an outcome of a complex psychological process.* The mechanism behind the brand equity formation is complex and probably highly individualistic.
- (3) *Nonseparable nature of brand and product equity.* The question of whether the value of brand asset is additive and can be viewed and assessed independently of its owner or the product it is associated with is unresolved. Brand characteristics and physical characteristics of the product carrying a brand name are not independent. Consumers exhibit significant biases in evaluating physical characteristics of branded products and also project their personal experiences and product perceptions onto the brand.

Brand equity is now measured at three main stages of the brand value chain (Keller and Lehman 2006). Consumer mindset approach uses survey methods to capture the psychological value consumers attach to a brand. Product market-based approach relies on scanner panel data or accounting data to estimate incremental market share, price, revenue, or profit premium attributable to the brand. Financial market-based approach uses secondary data from the stock market to estimate incremental firm value attributable to its brand assets. All these approaches involve a significant subjective component and judgement in deciding which perceptions to survey, how to define the market or generic benchmark, how much cash flow or market value to attribute to the brand versus other intangible assets the firm owns.

This chapter describes a tool that can be used to assess financial impact of brand perceptions and examine partial dynamics (immediate vs. future and direct vs. indirect) of this impact. This is conceptually and empirically different from measuring the value of brand equity. The presented approach has a limitation in that it allows estimating only partial (not full) dynamics of the effect. It does not allow estimating exactly how many years it will take for full benefits of brand development to be realized or what the pattern of the benefits in each future time period is. Also, the estimates are based on a set of monobrand firms. It is, however, likely (but remains to be confirmed) that the key insights are transferrable to individual product brands in multibrand firms.

With increasing availability of longer time series of brand metrics and brand performance outcome measures, addressing the full dynamics of brand impact will become feasible with standard time series and panel data approaches even at product brand level. Meanwhile, the approach described in this chapter allows assessing the long-term financial impact of marketing metrics with even limited time-series data to derive valuable insights. For example, in the empirical application of the method to disaggregate components of brand asset, important lessons are learned: most of the

financial impact of brand typically occurs in the future periods and brand perceptions have very different impact in different industrial sectors.

References

- Aaker, D. A. (1996). Measuring brand equity across products and markets. *California Management Review*, 38(3), 102–120.
- Aaker, D. A. (2012). *Building strong brands*. Free Press (November 8, 2011). Simon and Schuster Digital Sales Inc. https://www.amazon.com/Building-Strong-Brands-David-Aaker-ebook/dp/B005O315Z2/ref=la_B000APVZQI_1_9?s=books&ie=UTF8&qid=1504831274&sr=1-9&refinements=p_82%3AB000APVZQI
- Aaker, D. A., & Jacobson, R. (1994). The financial information content of perceived quality. *Journal of Marketing Research*, 31(2), 191–201.
- Aaker, D. A., & Jacobson, R. (2001). The value relevance of brand attitude in high-technology markets. *Journal of Marketing Research*, 38(4), 485–493.
- Ailawadi, K. L., Lehmann, D. R., & Neslin, S. A. (2003). Revenue premium as an outcome measure of brand equity. *Journal of Marketing*, 67(4), 1–17.
- Allison, R. I., & Uhl, K. P. (1964). Influence of beer brand identification on taste perception. *Journal of Marketing Research*, 1(3), 36–39.
- Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1), 47–82.
- Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2), 277–297.
- Austin, L. (2007). Accounting for intangible assets. *University of Auckland Business Review*, 9(1), 63–72.
- Bahadir, S. C., Bharadwaj, S. G., & Srivastava, R. K. (2008). Financial value of brands in mergers and acquisitions: Is value in the eye of the beholder? *Journal of Marketing*, 72(6), 49–64.
- Barth, M. E., Clement, M. B., Foster, G., & Kasznik, R. (1998). Brand values and capital market valuation. *Review of Accounting Studies*, 3(1–2), 41–68.
- Cañibano, L., Garcia-Ayuso, M., & Sanchez, P. (2000). Accounting for intangibles: A literature review. *Journal of Accounting Literature*, 19, 102–130.
- Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57–82.
- Cobb-Walgren, C. J., Ruble, C. A., & Donthu, N. (1995). Brand equity, brand preference, and purchase intent. *Journal of Advertising*, 24(3), 25–40.
- Daniel, K., & Titman, S. (1997). Evidence on the characteristics of cross sectional variation in stock returns. *The Journal of Finance*, 52(1), 1–33.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417.
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1), 3–56.
- Greene, W. H. (2002). *Econometric analysis* (5th ed.). Upper Saddle River: Pearson Education.
- Hrustic, E. (2012). Presentation at the “Brands and branding in law, accounting and marketing” conference. UNC.
- Kamakura, W. A., & Russell, G. J. (1993). Measuring brand value with scanner data. *International Journal of Research in Marketing*, 10(1), 9–22.
- Keller, K. L. (1993). Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing*, 57(1), 1–22.
- Keller, K. L., & Lehmann, D. R. (2003). How do brands create value? *Marketing Management*, 12(3), 26–31.

- Keller, K. L., & Lehmann, D. R. (2006). Brands and branding: Research findings and future priorities. *Marketing Science*, 25(6), 740–759.
- Lassar, W., Mittal, B., & Sharma, A. (1995). Measuring customer-based brand equity. *Journal of Consumer Marketing*, 12(4), 11–19.
- Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5), 881–894.
- Liu, L., Dzyabura, D., & Mizik, N. (2017). Visual listening in: Extracting brand image portrayed on social media. (May 8, 2017). Available at SSRN: <https://ssrn.com/abstract=2978805> or <http://dx.doi.org/10.2139/ssrn.2978805>.
- Madden, T. J., Fehle, F., & Fournier, S. (2006). Brands matter: An empirical demonstration of the creation of shareholder value through branding. *Journal of the Academy of Marketing Science*, 34(2), 224–235.
- Mizik, N. (2014). Assessing the total financial performance impact of brand equity with limited time-series data. *Journal of Marketing Research*, 51(6), 691–706.
- Mizik, N., & Jacobson, R. (2008). The financial value impact of perceptual brand attributes. *Journal of Marketing Research*, 45(1), 15–32.
- Mizik, N., & Jacobson, R. (2009). Valuing branded businesses. *Journal of Marketing*, 73(6), 137–153.
- Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3), 521–543.
- Pappu, R., Quester, P. G., & Cooksey, R. W. (2005). Consumer-based brand equity: Improving the measurement-empirical evidence. *Journal of Product and Brand Management*, 14(3), 143–154.
- Park, C. S., & Srinivasan, V. (1994). A survey-based method for measuring and understanding brand equity and its extendibility. *Journal of Marketing Research*, 31(2), 271–288.
- Pavlov, E. & Mizik, N. (2017). Values' voters and their brands, Working paper.
- Rangaswamy, A., Burke, R. R., & Oliva, T. A. (1993). Brand equity and the extendibility of brand names. *International Journal of Research in Marketing*, 10(1), 61–75.
- Simon, C. J., & Sullivan, M. W. (1993). The measurement and determinants of brand equity: A financial approach. *Marketing Science*, 12(1), 28–52.
- Sinclair, R. N., & Keller, K. L. (2014). A case for brands as assets: Acquired and internally developed. *Journal of Brand Management*, 21(4), 286–302.
- Srinivasan, V., Park, C. S., & Chang, D. R. (2005). An approach to the measurement, analysis, and prediction of brand equity and its sources. *Management Science*, 51(9), 1433–1448.
- Yoo, B., & Donthu, N. (2001). Developing and validating a multidimensional consumer-based brand equity scale. *Journal of Business Research*, 52(1), 1–14.



Measuring Sales Promotion Effectiveness

Karen Gedenk

Contents

Introduction	1056
Sales Promotion Tools and Their Effects	1056
Data for Measuring Sales Promotion Effectiveness	1059
Non-experimental Data on Observed Behavior	1059
Further Sources of Data	1061
Measuring Promotion Effectiveness with Panel Data	1062
SCAN*PRO	1063
PROMOTIONSCAN	1065
Decomposition Based on Single-Source Data	1067
Summary	1070
References	1070

Abstract

Sales promotions are an important marketing tool for both manufacturers and retailers. They include, for example, temporary price reductions, coupons, features, displays, sampling, and premiums. The bad news about promotions is that many of them are not profitable. The good news is that promotion effectiveness can be measured so that managers can identify the promotions which generate a profit and eliminate the ones that do not. This chapter presents data and models that can be used for this purpose. It focuses on panel data which is available at the aggregate (i.e., store) level and at the disaggregate (i.e., consumer) level. While aggregate data is more readily available and easier to analyze, disaggregate data allows for more detailed analyses. Several examples illustrate how models build on these data to measure promotion effectiveness. Since panel data has its limitations, it is often useful to complement it with surveys and/or experiments.

K. Gedenk (✉)
University of Hamburg, Hamburg, Germany
e-mail: karen.gedenk@uni-hamburg.de

Keywords

Sales promotions · Panel data · Surveys · Experiments · Decomposition of promotion effects

Introduction

Sales Promotions are a key marketing instrument for many companies. For example, firms sell their products with temporary price reductions (TPR), offer premiums, sweepstakes, or samples, and use feature advertising and displays to draw shoppers' attention to the promoted products. Manufacturer spendings on sales promotions are high, e.g., manufacturers of consumer packaged goods (CPG) in the USA spend 55% of their marketing budget on sales promotions (Cadent Consulting Group 2017). And retailers generate a large percentage of their sales with promoted products, e.g., in Europe 28% of volume sales in grocery retailing (IRI 2016).

Despite their widespread use, sales promotions are often not successful. McKinsey analyzed 5000 promotions in six European countries in 2002, finding that only 40% are profitable for the manufacturers of the promoted brands (N.N. 2002). Ailawadi et al. (2006) in their analysis of all promotions run by the drugstore chain CVS in 2003, find that less than half of them are profitable for the retailer.

Thus, managers may think that – similar to the famous quote for advertising – half the money they spend on promotions is wasted. Compared to advertising, however, it is much easier to find out which half this is. Effects of promotions are more immediate, so that, it is easier to establish causal relationships. There is plenty of data available for measuring promotion effectiveness, especially for CPG, and researchers have developed and tested models with which to analyze these data. Note that such analyses are not trivial – but given the huge investments in sales promotions, they are typically well worth the effort.

This chapter provides an overview of relevant data sources and approaches for analyzing these data. It focuses on promotions directed at consumers by manufacturers and retailers, as opposed to trade promotions, which manufacturers offer to retailers. The following section presents key sales promotion tools and their potential effects that need to be measured. Next, an overview of measurement approaches with a focus on relevant data sources and the opportunities they offer for analyzing sales promotion effectiveness. Since sales promotions are most heavily used for CPG, and the analysis of promotion effectiveness in this industry relies heavily on panel data, I describe some examples for this type of analysis in the remainder of the chapter.

Sales Promotion Tools and Their Effects

Promotions offered to consumers can be distinguished in price versus non-price promotions (Gedenk 2002; Gedenk et al. 2010). The most common type of price promotion is a temporary price reduction (TPR), where the product is offered at a

reduced price for a limited time (e.g., “normally 2.99 € – this week only 1.99 €”). Other types of price promotions include rebates (consumers pay full price and send in their receipt to receive a discount) and multi-item promotions (consumers only get a discount when they buy multiple units, e.g., “buy two – get one free”). Retailers and manufacturers can also distribute coupons through different media (e.g., newspapers, websites, or direct mail) and consumers receive a discount (for one or multiple items), when they present the coupon at the point-of purchase.

Non-price promotions can be “supportive” or “true” (Gedenk 2002; Gedenk et al. 2010). “Supportive” non-price promotions include, for example, displays, features (i.e., retailers’ weekly flyers), and other POS materials. They are typically used to draw attention to price promotions, and in most consumers’ minds, they are therefore closely linked to price cuts. Note, however, that “supportive” non-price promotions can also stand alone to highlight products at regular price. “True” non-price promotions, in contrast, clearly focus on the brand or store, not on price. Tools like samples, contests, and sweepstakes, as well as premiums, fall into this category. While managers typically use price and “supportive” non-price promotions to achieve short-term increases in sales and profit, “true” non-price promotions often focus more on long-term goals like building awareness for a brand and enhancing its image and thus increasing profit in the time after the promotion. Note that promotion campaigns by manufacturers and retailers often combine several of the above-mentioned tools.

Managers, who want to measure promotion effectiveness, need to make sure that they take into account all relevant effects. Even if the primary goal of a promotion is to increase sales in the short term, potential long-term effects must not be ignored. Also, the short-term sales bump needs to be decomposed to determine which part of it is truly incremental for manufacturers and retailers. Figure 1 details the different effects of promotions on sales of the promoted product in the focal store.

Suppose that chocolate by the brand Milka is on promotion at the German retailer Rewe. In this case, the sales of Milka at Rewe will evince a short-term increase in sales. This sales bump is made up of different components and can be decomposed by asking “what would consumers have done without the promotion?”. There are several answers to this question: The bump can occur because consumers otherwise would have bought their chocolate at Edeka (store switching), would have bought chocolate by Ritter Sport (brand switching), would have bought cookies instead of chocolate (category switching), or would have bought later or a smaller amount of chocolate (purchase acceleration). It is important to make this decomposition because effects that are advantageous for the retailer are not necessarily good for the manufacturer and vice versa. For example, brand switching increases sales for the manufacturer but not for the retailer, who only shifts sales from one brand to another. The opposite holds for store switching which increases sales for the retailer, while the manufacturer only shifts sales of its brand from one retailer to another. To determine whether purchase acceleration is advantageous, it needs to be decomposed further. If consumers buy more or buy earlier than they would have done without the promotion, in many product categories this leads to increased consumption, i.e., consumers eat more chocolate. This effect is positive for both manufacturers and retailers. However,

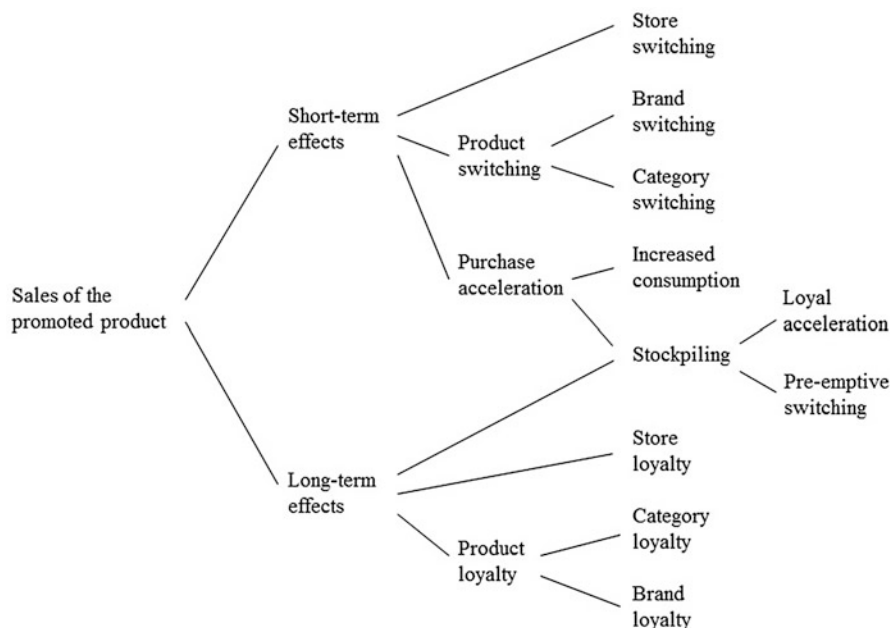


Fig. 1 Sales effects of promotions (Gedenk 2002, p. 104)

purchase acceleration may also result in consumers stockpiling the promoted product. In other words: The promotion steals sales from the future – without the promotion, the consumers would have purchased (more) at a later time. In this case, an interesting question arises: Does the promotion steal sales from the promoted brand in the focal store (loyal acceleration), or does it steal from the competition, i.e., other brands and/or other retailers (preemptive switching) (Ailawadi et al. 2007)? Preemptive switching results in an incremental sales effect for the focal brand/retailer, while loyal acceleration only shifts sales from one period to another.

In the long-term, promotions can affect consumers' loyalty to stores, brands, and product categories. The effect of price promotions on brand loyalty is often negative, in part because price cuts lead to lower reference prices and because consumers learn to buy on promotion (Gedenk 2002, pp. 245 ff.). This can result in conflicts within manufacturing firms, where the sales department often focuses primarily on short-term sales, whereas the marketing department is also interested in brand building. Store loyalty may improve because of a better price image but may also decrease because consumers learn to search for deals.

Note that Fig. 1 only captures effects of promotions on sales of the promoted brand in the store that runs the promotion. In addition, it can be interesting – especially for retailers – to look at effects on other product categories. Promotions can attract customers to the store who then buy other products that are not on promotion. Finally, in addition to effects on sales, managers need to consider prices and costs to determine the profitability of promotions.

Data for Measuring Sales Promotion Effectiveness

Figure 2 depicts a classification of relevant data sources, distinguishing between surveys versus observed behavior and between experimental versus non-experimental data.

Non-experimental Data on Observed Behavior

The large majority of academic research on sales promotions uses non-experimental data on observed behavior (cell I in Fig. 2), i.e., market data, especially panel data. For managers also, this is a key data source, especially for CPG sold through grocery retailers. Such data can have different levels of aggregation.

Sales data aggregated to the level of stores or retailers are available for many industries. Often retailers share their sales data with manufacturers, since both benefit from in-depth analyses of the data. Scanner panels provided by market research companies allow even more meaningful analyses, because they combine data from several retailers and include systematic information on sales promotions. The leading suppliers of such panels are IRI and Nielsen. They collect sales and price data from grocery retailers (including drugstores), who share the information from their databases. The retailers gather sales data by scanning consumers' purchases at the check-out. In addition, employees of the market research companies visit the participating stores each week to collect information on non-price promotions like displays and features.

Disaggregate data that captures purchases at the level of consumers is available to many companies in the form of consumer panels, provided, for example, by

	Non-experimental data	Experiments <ul style="list-style-type: none"> • Field experiments • Lab experiments
Observed behavior <ul style="list-style-type: none"> • Store-level data <ul style="list-style-type: none"> - Retailer sales data - Scanner panels • Consumer-level data <ul style="list-style-type: none"> - Consumer panels - Single-source panels - Loyalty card data 	I (majority of promotion research)	II
Survey	III	IV

Fig. 2 Data for measuring promotion success (Gedenk 2006, S. 586)

GfK. In a consumer panel, a large number of households record their purchases in the focal product categories (typically via in-home scanning), and send that information to the market research institute running the panel. For the analysis of sales promotions, consumer panel data need to be augmented with data on the promotions that were available in the stores where the panelists went shopping. Consumers' self-reports on the promotions they encountered are not very reliable and available only for the products purchased, but not for competing products. Thus, it is more promising to collect promotion information from other sources like retailer features. In Europe, this approach is used, for example, by the data provider FOCUS Marketing Research. It works well if many promotions are featured, which in turn requires coordination across the stores belonging to a retail chain.

A better matching of purchase with promotion data is achieved by single-source panels, like BehaviorScan by GfK. They combine a consumer panel with store panel data. In the German BehaviorScan market in Haßloch (Pfalz), 3000 households participate as well as most of the grocery retail stores in town. Households show ID cards when they go shopping, so that all their purchases are recorded. Retailers provide information on prices, and employees of GfK visit the stores to record non-price promotions. In addition, the participating households can receive targeted advertising – but that is typically only of minor relevance for the analysis of promotion effectiveness.

Loyalty card data also combine purchase with promotion information: Retailers know their prices and promotions, and can observe the purchase behavior of households participating in the loyalty program. The key problem here is that only purchases in the stores participating in the loyalty program are captured, while purchases at competing retailers cannot be observed.

Aggregate and disaggregate sales data allow for different types of analyses. Store-level data are good for measuring the short-term increase in sales caused by promotions. Researchers have also used aggregate data to decompose the short-term sales bump into its components and to measure long-term effects (e.g., Nijs et al. 2001; van Heerde et al. 2000, 2004). However, there is always the danger of an aggregation bias in these approaches. Neslin and Schneider Stone (1996) explain this in depth for the measurement of stockpiling effects. Their starting point is the observation that studies based on aggregate data typically do not find a post-promotion dip, while studies based on disaggregate data do find evidence for stockpiling. Neslin and Schneider Stone discuss several explanations for this phenomenon, mostly related to aggregation bias. For example, households have different interpurchase times, so that each household that stockpiles has its personal post-promotion dip at a different time, and an overall dip is not visible when data are aggregated across households. Also, deal-to-deal buying – an extreme form of stockpiling where households only buy a product when it is on promotion – cannot be detected in aggregate data.

A more precise decomposition of the short-term sales bump can be done with disaggregate data at the consumer level. It allows for a detailed analysis of consumer response to promotions: Researchers can not only model aggregate sales but rather

different consumer decisions like store choice, product category purchase incidence, brand choice, and decisions on purchase quantity. Often, researchers use models of purchase incidence, brand choice, and purchase quantity to identify brand switching and purchase acceleration (e.g., Gupta 1988; Bell et al. 1999; Ailawadi et al. 2007). Van Heerde et al. (2003) point out that this decomposition should be done based on unit sales rather than on elasticities. Otherwise, the importance of brand switching relative to purchase acceleration will be overstated.

Thus, researchers face a trade-off when choosing between aggregate and disaggregate data for analyzing promotion effectiveness. On the one hand, disaggregate data allow for more detailed analyses. On the other hand, aggregate data are cheaper, and the analysis requires less effort. Therefore, managers mostly analyze store-level data, while academics also use consumer-level data to provide deeper insights into promotion effectiveness (for a review, see Gedenk 2002; Gedenk et al. 2010). Academic research on sales promotions often relies on market data provided by market research institutes, manufacturers, and retailers. An example of a publically available database is described by Bronnenberg et al. (2008), and organizations like AiMark and the Kilts Center for Marketing at Chicago Booth make other databases available to academics.

In sum, non-experimental data on observed behavior can provide many insights into promotion effectiveness. However, it also has limitations:

- Market data is not readily available for all product categories. It abounds for CPG and grocery retailing, but may be difficult to come by in other industries.
- Market data is only available for products that are already on the market and for promotions that have already been used.
- Consumers' purchase behavior is driven by many factors beyond sales promotions, and it is often hard to control for these factors such that promotion effects can be properly identified. Also, endogeneity of the promotion variables may bias results.
- Market data can be used to measure consumers' behavior, but it does not contain information about attitudes, emotions, and motives that may explain this behavior.

Therefore, the measurement of promotion effectiveness is often supplemented by using other types of data.

Further Sources of Data

Surveys can be used to measure attitudes, emotions, motives, and behavioral intentions. They can solve all of the four problems mentioned above: Surveys can inquire into new products and new promotions in any product category and provide explanations for behavior, while controlling for extraneous factors. They can also easily capture characteristics of promotions, products, and consumers, as well as other factors which potentially moderate promotion response. A downside of surveys, of

course, is their limited external validity. Do respondents indicate their true attitude, and will they really behave as stated in a survey?

Non-experimental surveys (cell III in Fig. 2) have been used, for example, to measure deal proneness, i.e., to find out which consumers respond most strongly to promotions (e.g., Ailawadi et al. 2001). More frequently, however, surveys contain experiments, in which promotions are manipulated systematically to test the effects of different promotion designs (cell IV in Fig. 2). Experiments attempt to keep all variables besides the treatment variables constant, and therefore allow the diagnosis of causal effects, in this case, the identification of promotion effects and their moderators. For example, Arora and Henderson (2007) study the effect of cause-related marketing campaigns, which is difficult to isolate in field data. Also, many survey-based experiments have looked into the framing of price promotions. They have studied, for example, under which conditions a price cut should be indicated in percent versus as an absolute value (e.g., Chen et al. 1998).

Finally, field experiments with observed behavior (cell II in Fig. 2) are a valuable data source for measuring promotion effectiveness. They combine the high external validity of market data with the high internal validity of an experiment. Compared to non-experimental market data, promotion effects can be identified more clearly. Unfortunately, managers are often reluctant to run field experiments, maybe because they worry about high costs or about offending customers who receive unfavorable treatments. Yet, even simple store tests can yield very interesting results. For example, Wansink et al. (1998) show that quantity restrictions on price promotions (“maximum of X units”) can increase the average quantity bought. One explanation for this surprising effect is anchoring and adjustment, such that consumers use the quantity indicated in the restriction as an anchor and adjust their purchase quantity to it.

Note that the discussion so far has focused on promotions in offline retail channels. For digital promotions in the online and mobile channels, managers have some additional options for observing consumer behavior. In particular, they have information not only on purchases but also on consumers’ search behavior. In mobile marketing, they can make use of information on consumers’ location and target their promotions accordingly. Also, field experiments are much easier to implement and thus less costly than in the offline channel, so that we see more experiments for digital promotions. For example, Luo et al. (2014) and Fong et al. (2015) run field experiments to study how the design of mobile coupon campaigns affects redemption rates.

Measuring Promotion Effectiveness with Panel Data

In this section, I describe the basic ideas behind three frequently used approaches for analyzing non-experimental panel data (cell I in Fig. 2) and provide examples of applications. For a more detailed discussion and further promotion models, see van Heerde and Neslin (2017). Since managers mainly use store-level data to measure promotion effectiveness, I first present two common approaches for this type of analysis: the SCAN*PRO model developed in cooperation with Nielsen and the

PROMOTIONSCAN model developed in cooperation with IRI. At the core of the SCAN*PRO model (Wittink et al. 1987) is a multiplicative sales response function. In contrast, PROMOTIONSCAN (Abraham and Lodish 1993) uses a baseline approach. It determines the effect of a promotion as the difference between actual sales and baseline sales, where the baseline, i.e., the sales level that would have been achieved without the promotion, is estimated based on sales in promotion-free weeks. A third example (Ailawadi et al. 2007) illustrates a more in-depth analysis of promotion effects based on single-source panel data.

SCAN*PRO

In the SCAN*PRO model (Wittink et al. 1987; van Heerde et al. 2002), unit sales are a multiplicative function of price and non-price promotions:

$$Q_{ist} = \prod_{j=1}^J \left[\left(\frac{P_{jst}}{BP_{js}} \right)^{\beta_j} \cdot \gamma_{1j}^{F_{jst}} \cdot \gamma_{2j}^{D_{jst}} \cdot \gamma_{3j}^{FAD_{jst}} \right] \cdot \prod_{t=1}^T [\delta_t^{W_t}] \cdot \prod_{s=1}^S [\lambda_s^{Z_s}] \cdot e^{u_{ist}} \quad (1)$$

with:

- Q_{ist} = Unit sales of brand i in store s and time period t
- P_{jst} = Price per unit of brand j in store s and time period t
- BP_{js} = Base price per unit of brand j in store s (median of prices in weeks without sales promotion)
- F_{jst} = Indicator variable for feature (1 if brand j is featured but not on display in store s and time period t, 0 else)
- D_{jst} = Indicator variable for display (1 if brand j is on display but not featured in store s and time period t, 0 else)
- FAD_{jst} = Indicator variable for feature and display (1 if brand j is featured and on display in store s and time period t, 0 else)
- W_t = Indicator variable for time period (1 if observation is from week t, 0 else)
- Z_s = Indicator variable for store (1 if observation is from store s, 0 else)
- u_{ist} = Error term for brand i, store s, and time period t

Wittink et al. (1987) suggest aggregating the data to the level of weeks, stores, and brands, but different levels of aggregation are possible (e.g., days instead of weeks, retail chains instead of stores, and SKUs instead of brands). Typically, researchers pool across stores and time periods, and estimate one model per brand. Sales of the focal brand are a function of prices and promotions of this brand, but also of the prices and promotions of competing brands in the same product category. Prices are made comparable across stores by dividing actual price by base price. The base price captures the regular price in a store and is defined as the median of prices of a brand in a store in promotion-free time periods. Differences in unit sales levels between stores can be accommodated in a similar way: by dividing by base sales. Or

the researcher can include store dummies Z_s in the model, as in Eq. (1). SCAN*PRO models can also contain dummy variables for non-price promotions. Often these are “supportive” non-price promotions like features and displays (see Eq. 1), but of course, the effects of “true” non-price promotions can be modeled in the same way, if the respective data is available. Finally, indicator variables for weeks W_t control for seasonal effects and exceptional events.

The effect of price promotions is captured by the price elasticity β_i . For $i = j$, this is the direct elasticity (price of the focal brand i affects the sales of i), while for $i \neq j$ it is a cross-elasticity, (price of a competing brand j affects sales of the focal brand i). The coefficients γ_{1j} to γ_{3j} are multipliers indicating how much larger unit sales are with a non-price promotion than without.

An example for a SCAN*PRO model can be found in the study by Foekens et al. (1999). The authors estimate the model in Eq. (1) with scanner data from 28 stores of a US retail chain for three brands from one product category. Table 1 presents the parameter estimates for price and non-price promotion effects. Direct effects are highlighted by grey shading.

The three direct price elasticities (effect of P_i on sales of brand i) have typical values of around -3 , close to the mean short-term promotional price elasticity of -3.63 that Bijmolt et al. (2005) find in their meta-analysis. The cross-price elasticities (effect of P_j on sales of brand i) are also significant. Positive values indicate brand switching – if one brand decreases price, the other brands lose sales. Features only have a significant direct effect on sales for brand C (effect of F_C on sales of brand C). Here, unit sales are 2.12 times as high with a feature than without. Displays cause sales to roughly double for all three brands (effect of D_i on sales of brand i). Some of the multipliers for cross-effects of non-price promotions (effect of F_i , D_i , and FAD_i on sales of brand i) are smaller than 1, indicating brand switching.

Table 1 Parameters of the SCAN*PRO model by Foekens et al. (1999, p. 262)

Independent variable	Dependent variable: Unit sales of brand...		
	A	B	C
P_A	-2.96	0.26	0.54
P_B	0.69	-2.42	1.34
P_C	1.08	0.38	-3.21
F_A	n. s.	1.48	0.63
D_A	1.80	n. s.	0.82
FAD_A	1.75	1.34	0.54
F_B	0.38	n. s.	n. s.
D_B	n. s.	1.54	n. s.
FAD_B	0.33	2.08	0.61
F_C	n. s.	1.42	2.12
D_C	n. s.	n. s.	2.41
FAD_C	n. s.	1.26	3.22

Direct effects are shaded in grey, the remaining effects are cross-effects

P Price, F Feature only, D Display only, FAD Feature and Display, *n. s.* not significant

However, for brand B, Foekens et al. find significant multipliers larger than 1. Thus, brand B benefits from non-price promotions of competing brands. This can be explained by an attention effect. Features for brands A and C (alone or in combination with displays) draw consumers' attention to the product category, and customers with a preference for brand B buy more in the category but stay loyal to brand B. Only price promotions provide a sufficient incentive for them to switch brands. This suggests that brand B is a strong brand. Also, brand A should not pay an advertising allowance to retailers, since features for brand A only benefit brand B.

Many extensions of this simple SCAN*PRO model have been suggested over the years (van Heerde et al. 2002). For example researchers have estimated varying-parameters/hierarchical models where the price elasticity is a function of second-level variables like past promotions (Foekens et al. 1999) or store-size (Haans and Gijbrecchts 2011). Van Heerde et al. (2000) and Neslin and Macé (2004) add lead and lag effects (i.e., effects of past and future promotions) to model purchase acceleration. Van Heerde et al. (2001) show that price cuts with different depths have different effects on purchase behavior. And van Heerde et al. (2004) demonstrate that a decomposition of the short-term sales bump caused by a promotion is possible with aggregate sales data. Note that SCAN*PRO is not only frequently used by academics but also by managers, reflecting the wide availability of store-level panel data and the low effort required for the analysis. Van Heerde et al. (2002, p. 201) report more than 3000 commercial applications by Nielsen, and that number has certainly grown further by now.

PROMOTIONSCAN

With SCAN*PRO, researchers model the effect of promotions on sales directly and estimate it with panel data from all weeks. In contrast, Abraham and Lodish (1993) in their PROMOTIONSCAN model choose an indirect approach. They estimate a baseline which reflects sales that would have been achieved without promotions, using only data from weeks whose sales are not affected by promotions. Specifically, they proceed in five steps:

1. Adjustment of data for trend and seasonality,
2. Elimination of periods where sales is affected by promotions,
3. Elimination of outliers,
4. Calculation of preliminary baseline by smoothing over promotion-free periods and adding trend and seasonality back in,
5. Adjustment for out-of-stock situations and market-specific factors (e.g., marketing activities of the competition).

Some of these steps are performed several times in an iterative process. Together with final consistency checks, this leads to robust estimations that can be automated to a large degree. Once the researcher has estimated a baseline, he can compare it to actual sales to determine the effect of promotions. Typically, actual sales are higher than the baseline during a promotion, and lower afterwards because of stockpiling.

A key difficulty with PROMOTIONSCAN is the identification of periods where sales are not affected by promotions. The researcher needs to exclude not only the weeks during a promotion but also the weeks with a post-promotion dip caused by stockpiling afterwards. In product categories with frequent promotions, there are sometimes only few weeks left with which to estimate a baseline.

Baseline models are also frequently used in practice. A prominent example that combines academic research with a practical application is the study by Ailawadi et al. (2006). The authors use the relatively simple and robust method to analyze all promotions in 2003 of the US drug store chain CVS, i.e., 36 million promotions in 189 product categories and 3808 stores. They determine a baseline for a promotion by computing moving averages across promotion-free weeks before and after the promotion. The number of weeks taken into account varies, depending on the turnover and seasonality of the product. The difference between actual sales and the baseline yields the gross lift in unit sales of a product in a given store. Next, Ailawadi et al. decompose this short-term sales bump and consider a promotion's effect on other product categories. The retailer (CVS) needs to identify the components of the gross lift that are taken from other products in the category (brand switching) and from its future sales (loyal acceleration). These two components do not constitute incremental sales for the retailer, while the remaining components do. Furthermore, if a promotion has a halo effect, i.e., if it increases sales of other product categories (e.g., due to store switching), that is advantageous for the retailer. Therefore, Ailawadi et al. determine the extent of brand switching and halo effects on other product categories based on scanner data from CVS. They isolate loyal acceleration with the help of data from the retailer's loyalty program, since – as explained above – disaggregate data are more appropriate than store-level data for measuring stockpiling (Neslin and Schneider Stone 1996). Finally, the authors have access to data on retailer costs and on trade promotions, so that they can measure not only the net sales effect but also the net profit impact of promotions. Table 2 shows the medians for the various effects for four broad groups of product categories as well as for the full sample.

The gross sales lift, i.e., the short-term increase in sales, is about 300%. Given price cuts of 30% on average, this corresponds to price elasticities of roughly -10 . Note that these elasticities capture not only the effect of price promotions, but also the support by non-price promotions like features and displays. A little less than half of the gross sales lift (46% on average) derives from brand switching, and about 10% from loyal acceleration. The remainder – about 45% – constitutes incremental sales for CVS. In addition to these within-category effects, there is a halo effect: In three out of four groups of product categories, promotions in one category also advance sales in other product categories. Across the full sample, for each unit of the gross sales lift, there is a sales increase of 0.16 units of products from other categories. All these effects together (gross sales lift – brand switching – local acceleration + halo) result in a net sales impact of 1.05 units on average. That is, on promotion, CVS sells 1.05 units more per article, week, and store than the baseline of 0.86 units. However, the net profit impact is negative in two out of four groups of product categories. Here, the incremental sales and the trade promotions CVS receives from the manufacturers are not sufficient to compensate for the loss in margin. Overall, more than half of the promotions analyzed are not profitable for CVS.

Table 2 Results of Ailawadi et al. (2006, p. 527) (medians)

Effect	Full sample	Health	Beauty	Edibles	General merchandise
Gross sales lift	310%	264%	314%	308%	421%
Brand switching (fraction of gross lift)	0.46	0.50	0.47	0.40	0.43
Loyal acceleration (fraction of gross lift)	0.10	0.11	0.09	0.15	0.08
Halo (fraction of gross lift)	0.16	-0.04	0.30	0.05	0.28
Net unit sales impact	1.05	0.58	1.35	2.07	1.71
Baseline unit sales	0.86	0.80	0.67	2.00	0.75
Net profit impact	-0.62	-0.93	0.23	-1.14	0.08
Baseline profit	1.29	1.69	1.24	0.91	0.94

Ailawadi et al. also identify drivers of both the net sales impact and the net profit impact of promotions at CVS. Differences in promotion effectiveness are small between stores, but substantial between brands, product categories, and types of promotions. One of the most interesting results of the paper is that the drivers which explain this variance, typically have opposing effects on incremental sales versus incremental profit. For example, the depth of the price cut has a positive effect on the net sales impact of promotions (because it provides more incentive to purchase), but negatively affects their net profit impact (because it cuts more deeply into margins). Also, a brand’s market share has a positive effect on the net sales impact, but a negative effect on the net profit impact. This should remind managers of how dangerous it is to only look at sales effects when measuring promotion effectiveness. Also, corporate governance should not focus primarily on sales and market shares, but rather on profit goals. With the wrong incentive system, it is difficult for managers to eliminate unprofitable promotions, because typically this will only be possible by sacrificing sales.

The analysis of promotion effectiveness with models like SCAN*PRO or PROMOTIONSCAN has the big advantage that store-level data are easily available in many industries, especially CPG, and that the analysis can be automated to a large degree. On the other hand, disaggregate data allow for more detailed analysis of promotion effects. The example of Ailawadi et al. (2006) illustrates how both types of data and analyses can be combined. The authors use consumer-level data to measure purchase acceleration. The next subsection describes the study by Ailawadi et al. (2007) as an example of how an even more detailed decomposition of the sales promotion bump can be done with single-source data.

Decomposition Based on Single-Source Data

Ailawadi et al. (2007) analyze the effectiveness of price promotions and “supportive” non-price-promotions from the perspective of a manufacturer. Their model is based on data from a single-source panel and captures consumers’ decisions on category purchase incidence, brand choice, and purchase quantity:

$$P_{ht}(i \& q) = P_{ht}(inc) \cdot P_{ht}(i|inc) \cdot P_{ht}(q|inc \& i) \quad (2)$$

with:

$P_{ht}(i \& q)$ = Probability that household h purchases quantity q of brand i on shopping trip t

$P_{ht}(inc)$ = Probability that household h makes a purchase in the focal product category on shopping trip t

$P_{ht}(i|inc)$ = Probability that household h chooses brand i on shopping trip t , conditional on a purchase in the product category

$P_{ht}(q|inc \& i)$ = Probability that household h buys quantity q of brand i , conditional on a purchase in the category and on choosing brand i

Each shopping trip of a household constitutes an observation. On a given shopping trip, a household first decides on whether to make a purchase in the focal product category or not. If he decides to purchase, he next chooses a brand, and finally, he decides how much of it to buy. The authors use a nested logit model to capture category purchase incidence and brand choice, and a poisson model for purchase quantity.

Ailawadi et al. (2007) are primarily interested in a detailed analysis of purchase acceleration. Purchase acceleration as a whole becomes visible in category incidence and purchase quantity decisions: consumers buy earlier and/or more than they would have done without a promotion. Part of this extra quantity that consumers now have available goes into increased consumption, because consumers have fewer stock-outs and increase their consumption rates. Since inventory levels and consumption cannot be observed in panel data, the authors need to make assumptions about the quantities consumed and update inventory levels accordingly. They estimate a model with flexible consumption, and find that households consume more of a product if they have more of it in stock. That is, promotions that increase the inventory of households, lead to increased consumption.

In addition to increasing consumption, households also stockpile, i.e., sales are stolen from the future. Ailawadi et al. (2007) introduce the distinction between sales stolen from the brand's own future sales (loyal acceleration) versus from competing brands (preemptive switching). Finally, they also study how purchase acceleration affects brand loyalty, by including detailed feedback effects in their brand choice model. The respective parameters indicate that promotions decrease brand loyalty, but this effect is smaller when consumers stockpile. An explanation could be that consumers get more used to a brand if they use more of it, and thus are more likely to repurchase out of habit.

Finally, an analysis based on disaggregate data needs to take into account that consumers are heterogeneous in their brand preferences as well as in their response to marketing mix variables. If this heterogeneity is ignored, model parameters will be biased. Therefore, Ailawadi et al. (2007) use a continuous mixture model, where parameters follow a normal distribution (Train 2009). Once the authors have estimated the parameters of their model of purchase behavior, they determine the size of

the sales bump caused by promotions and decompose it. Their decomposition approach is based on unit sales and uses a Monte Carlo simulation. The authors simulate purchase behavior with and without a promotion – differences are caused by the promotion, and the authors assign them to the various components of the promotion effect.

Table 3 presents the results of this decomposition for one yogurt and one ketchup brand. The underlying model was estimated with data from a US single-source panel for these two product categories and used a dummy variable for promotions, i.e., temporary price reductions and/or features and displays.

As in the previous examples, the short-term sales bump is large. For the yogurt brand “Dannon,” sales increase by 12.45 units in the short run, compared to baseline sales in non-promotion periods of 3.88 units. About one third of this sales bump comes from brand switching and about two thirds from purchase acceleration. Most of the extra inventory that households acquire from purchase acceleration goes into increased consumption. That is, promotions make consumers eat much more yogurt, which is a favorable effect for manufacturers. There is only little stockpiling (loyal acceleration and preemptive switching), and when promotions steal sales from the future, they mostly steal from the competition (preemptive switching), and not from the focal brand’s future sales (loyal acceleration). This holds for three out of four yogurt brands investigated. For the ketchup brand “Del Monte,” there is less of a consumption effect than for yogurt, and the promotional sales bump is mostly driven by brand switching instead, which is plausible. Again, for three out of four ketchup brands, preemptive switching is larger than loyal acceleration.

Table 3 further shows that promotions affect consumption not only in the short, but also in the long run. Finally, there is a weak positive effect of purchase acceleration on brand loyalty. Note that this is not the total effect of promotions on

Table 3 Decomposition results by Ailawadi et al. (2007, p. 459)

		Yogurt (“Dannon”)		Ketchup (“Del Monte”)	
		Units	% of sales bump	Units	% of sales bump
Baseline sales		3.88		0.80	
Short-term sales bump		12.45	100%	3.85	100%
<i>Decomposition of short-term sales bump</i>					
Brand switching		4.78	38.4%	2.26	58.6%
Purchase acceleration	Short-term increase in consumption	6.50	52.2%	1.14	29.5%
	Loyal acceleration	0.41	3.3%	0.21	5.4%
	Preemptive brand switching	0.76	6.1%	0.25	6.5%
<i>Long-term effects from purchase acceleration</i>					
Long-term increase in consumption		-1.61	-12.9%	0.04	1.1%
Brand loyalty effect of purchase acceleration		0.20	1.6%	0.03	0.8%

loyalty but only the part that can be attributed to consumers purchasing and consuming more of the brand.

Summary

In order to determine sales promotion effectiveness, marketing researchers need to measure the short-term sales bump caused by the promotion, decompose it into its components, and capture the long-term effects of the promotion. Mostly, this is done based on panel data. Aggregate data at the store level are available either from retailers or from scanner panels, and the best disaggregate data at the household level is available from single-source panels. With respect to these data sources, there is a trade-off between costs and benefits. While disaggregate data allow a very detailed measurement of promotion effects, single-source data is not widely available, costly, and its analysis is difficult as well as time-consuming. Aggregate data yield fewer insights but are more readily available and easier to analyze. Examples illustrate how aggregate and disaggregate data can be used to measure promotion effectiveness. Since all panel data analysis has some limitations, it often is helpful to also use surveys and experiments. Finally, it is important to keep in mind that for monitoring promotions, it is not sufficient to focus on sales effects, since many promotions increase sales but decrease profits.

References

- Abraham, M. M., & Lodish, L. M. (1993). An implemented system for improving promotion productivity using store scanner data. *Marketing Science*, 12(3), 248–269.
- Ailawadi, K. L., Neslin, S. A., & Gedenk, K. (2001). Pursuing the value-conscious consumer: Store brands versus National Brand Promotions. *Journal of Marketing*, 65(1), 71–89.
- Ailawadi, K. L., Harlam, B. A., César, J., & Trounce, D. (2006). Promotion profitability for a retailer: The role of promotion, brand, category, and store characteristics. *Journal of Marketing Research*, 43(4), 518–535.
- Ailawadi, K. L., Gedenk, K., Lutzky, C., & Neslin, S. A. (2007). Decomposition of the sales impact of promotion-induced stockpiling. *Journal of Marketing Research*, 44(3), 450–467.
- Arora, N., & Henderson, T. (2007). Embedded premium promotion: Why it works and how to make it more effective. *Marketing Science*, 26(4), 514–531.
- Bell, D. R., Chiang, J., & Padmanabhan, V. (1999). The decomposition of promotional response: An empirical generalization. *Marketing Science*, 18(4), 504–526.
- Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42(2), 141–156.
- Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). The IRI marketing data set. *Marketing Science*, 27(4), 745–748.
- Cadent Consulting Group. (2017). 2017 Marketing spending industry study. <http://cadentcg.com/wp-content/uploads/2017-Marketing-Spending-Study.pdf>. Accessed 10 Aug 2017.
- Chen, S.-F. S., Monroe, K. B., & Lou, Y.-C. (1998). The effects of framing price promotion messages on consumers' perceptions and purchase intentions. *Journal of Retailing*, 74(3), 353–372.
- Foekens, E. W., Leeflang, P. S. H., & Wittink, D. R. (1999). Varying parameter models to accommodate dynamic promotion effects. *Journal of Econometrics*, 89(1/2), 249–268.

- Fong, N. M., Fang, Z., & Luo, X. (2015). Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research*, 52(5), 726–735.
- Gedenk, K. (2002). *Verkaufsförderung*. München: Vahlen.
- Gedenk, K. (2006). Controlling von Verkaufsförderungsmaßnahmen. In S. Reinecke & T. Tomczak (Eds.), *Handbuch Marketingcontrolling* (2nd ed., pp. 573–592). Gabler: Wiesbaden.
- Gedenk, K., Neslin, S. A., & Ailawadi, K. A. (2010). Sales promotion. In M. Krafft & M. K. Mantrala (Eds.), *Retailing in the 21st century* (2nd ed., pp. 393–407). Berlin: Springer.
- Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research*, 25(4), 342–355.
- Haans, H., & Gijbrecchts, E. (2011). “One-deal-fits-all?” on category sales promotion effectiveness in smaller versus larger supermarkets. *Journal of Retailing*, 87(4), 427–443.
- IRI. (2016). Price and promotion in Western Europe. <https://www.iriworldwide.com/IRI/media/IRI-Clients/International/Price-Promotion-in-Western-Economies-a-Pause-in-Promotion-Escalation.pdf>. Accessed 10 Aug 2017.
- Luo, X., Andrews, M., Fang, Z., & Phang, C. W. (2014). Mobile targeting. *Management Science*, 60(7), 1738–1756.
- N.N. (2002, May 24). Promotions – Fass ohne Boden? *Lebensmittelzeitung*, p. 36.
- Neslin, S. A., & Macé, S. (2004). The determinants of pre- and postpromotion dips in sales of frequently purchased goods. *Journal of Marketing Research*, 41(3), 339–350.
- Neslin, S. A., & Schneider Stone, L. G. (1996). Consumer inventory sensitivity and the post-promotion dip. *Marketing Letters*, 7(1), 77–94.
- Nijs, V., Dekimpe, M. G., Steenkamp, J.-B. E. M., & Hanssens, D. M. (2001). The category-demand effects of price promotions. *Marketing Science*, 20(1), 1–22.
- Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge: Cambridge University Press.
- Van Heerde, H. J., & Neslin, S. A. (2017). Sales promotion models. In B. Wierenga & R. van der Lans (Eds.), *Handbook of marketing decision models* (2nd ed., pp. 13–78). Berlin: Springer.
- van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2000). The estimation of pre- and postpromotion dips with store-level scanner data. *Journal of Marketing Research*, 37(3), 383–395.
- van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2001). Semiparametric analysis to estimate the deal effect curve. *Journal of Marketing Research*, 38(2), 197–215.
- van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2002). How promotions work: Scan*pro-based evolutionary model building. *Schmalenbach Business Review*, 54(3), 198–220.
- van Heerde, H. J., Gupta, S., & Wittink, D. R. (2003). Is 75% of the sales promotion bump due to brand switching? No, only 33% is. *Journal of Marketing Research*, 40(4), 481–491.
- van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2004). Decomposing the sales promotion bump with store data. *Marketing Science*, 23(3), 317–334.
- Wansink, B., Kent, R. J., & Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research*, 35(1), 71–81.
- Wittink, D. R., Addona, M. J., Hawkes, W. J., & Porter, J. C. (1987). Scan*pro: A model to measure short-term effects for promotional activities on brand sales, based on store-level scanner data, Working paper, Cornell University.



Return on Media Models

Dominique M. Hanssens

Contents

Introduction	1074
The Importance of Reference Points in Media Return Calculations	1075
Fundamental Advertising Response Phenomena	1076
Estimating Media Response Parameters	1078
The Shape of the Advertising-Response Function	1078
Advertising-Response Dynamics	1080
Data-Interval Bias	1082
Asymmetric Response	1083
Dealing with Reverse Causality	1083
Differences Across Media	1085
Advertising Copy and Creative Effects	1086
Intermediate Performance Metrics	1086
Deriving Media Returns from the Estimated Response Parameters	1087
Path-to-Purchase and Attribution Models for Digital Media	1088
Media Advertising and Asset Creation	1090
Brand Equity	1091
Customer Equity	1092
Other Response Metrics	1093
Conclusion	1094
References	1094

Abstract

The proliferation of marketing media, especially since the advent of digital media, has created an urgent need for marketers to understand their relative importance in generating revenue for their brands. Ultimately, this understanding should result in managers' ability to project returns from their media investments. This chapter will focus on quantitative methods that enable such media return calculations. We

D. M. Hanssens (✉)
UCLA Anderson School of Management, Los Angeles, CA, USA
e-mail: dominique.hanssens@anderson.ucla.edu

begin with a definition of “return on media” and show how it connects to the need of estimating top-line lift, i.e., consumer response to media, from various data sources. We introduce the standard media-mix response model and discuss the estimation of media response elasticities. We extend these models to include brand-building and customer-equity effects and intermediate-performance variables. Finally, we address return to media in the digital era, with specific reference to path-to-purchase models, and we describe how media returns are derived from sales response models.

Keywords

Return on media · Marketing mix · Sales response models · Attribution models · Marketing resource allocation

Introduction

Advertising is one of the most visible activities of companies and brands. Firms and brands advertise for a variety of reasons, among them to help launch new products, to announce price changes, to increase brand awareness, and to protect the brand franchise against competitive encroachment. These efforts are expensive. Worldwide advertising expenditures amounted to about \$600 billion in 2015. In relative terms, the advertising outlays in many firms are of an order of magnitude comparable to that of their profitability. For example, in 2015, the worldwide ad spending of Procter & Gamble was about \$8.3 billion, while their net income was about \$7 billion. Thus, knowledge of the economic impact of, and more specifically, the *return* on advertising spending is of paramount importance to managers and investors alike.

In recent years, there has been increasing pressure on marketing executives to demonstrate the shareholder value created by these investments, which are, after all, discretionary. Not surprisingly, a financial definition has emerged as the key metric for value, viz., return on investment (ROI). This motivates the focus of the current chapter on the models that are needed to obtain reasonable estimates of these media returns.

We start with a concise definition of return on media (ROM hereafter). Consistent with finance practice, return on media is the estimate of the incremental financial value (in monetary terms) to the firm generated by identifiable media expenditures, less the cost of those expenditures as a percentage of the same expenditures (Some firms do not subtract cost of media in the numerator. The resulting metric is still usable, as it merely shifts the break-even value from 0 to 1. However, that definition is, strictly speaking, not a “return” metric and runs the risk of being misinterpreted by financial executives.):

$$\text{ROM} = \frac{\text{Incremental Financial Value Generated by Media} - \text{Cost of Media}}{\text{Cost of Media}} \quad (1)$$

Unlike other types of investments, media funds are rarely tied up in inventories, fixed assets, or receivables, and most media expenditures come from what otherwise would be liquid funds. Therefore, great care will need to be taken to validate comparisons between the return on media and other ROI estimates. In particular, the effects of some media spending may be short-lived, while other media actions may generate revenue and profit returns over multiple years, building cumulative impact and creating assets with future value. What is needed is a strong focus on the first part of the numerator in (1), i.e., the *incremental* financial value attributed to the media spending. This attribution needs to, first, understand the *top-line* effects of media spending, which are typically expressed as unit sales or sales revenues. Unit sales are then multiplied by gross profit margins to obtain gross financial contribution. Sales revenues are multiplied by percent gross margin, or, in the case of relationship businesses, by the margin multiplier (Gupta et al. 2004) (The margin multiplier transforms short-term revenue to long-term revenue by incorporating the expected loyalty levels of newly acquired customers. The section on “*Customer Equity*” offers more specifics.). Once the incremental financial contributions are determined, it is straightforward to do the cost accounting part of the equation, i.e. subtract and divide by media spending.

With respect to top-line media effects, there exists a detailed marketing science literature on the sales response effects of advertising and other marketing drivers, see for example Hanssens et al. (2001), to which we turn next. We will make ample use of this literature in discussing the nature of consumer response to advertising response and its implications for model building. This focus on top-line media effects will also allow us to avoid some misinterpretations in industry’s use of media return estimates, which we discuss first.

The Importance of Reference Points in Media Return Calculations

Although the math is simple, the meaning and significance of the ROM metric is anything but straightforward. Above all, the first term in the numerator, “incremental financial value generated by media” needs careful attention. “Incremental” can only be measured if there is a baseline or reference point for comparison, i.e., “incremental compared to what.” Second, the metric makes it necessary to make an attribution with respect to media expenditures, i.e., there needs to be a causal link between the two. Finally, there is a time dimension with respect to “incremental value” that could influence the calculations.

Industry studies often result in stated conclusions such as “the return of our TV spending is 45%, whereas it is 32% for spending on print media. Therefore, TV works better for us.” Regardless of which medium is more effective, such statements are misleading, because they critically depend on the amount spent on each media. For example, if print is highly impactful, but the firm overspends on print media, its total return will be affected negatively, and could easily drop below that of other, less impactful media. Total ROM comparisons across media can only be made when the spending levels are the same, which is typically not the case.

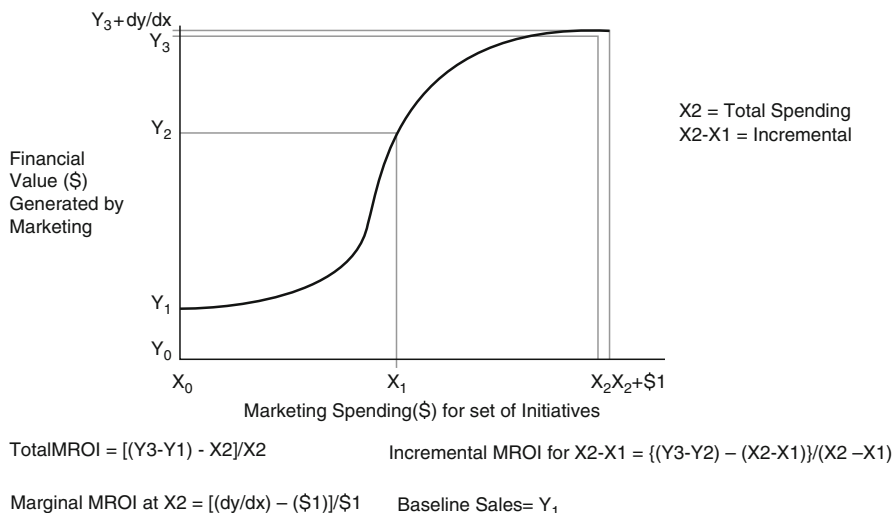


Fig. 1 Total, incremental, and marginal media returns (Source: Farris et al. (2015))

Figure 1 illustrates graphically the distinction between reporting total, incremental, and marginal ROM. Total ROM evaluates return on all spending, incremental for a specified additional spending “increment,” and marginal is the estimated return on the “last dollar” of media spending. Total and Incremental ROM are typically easier to estimate and often result from experimental designs (so-called A/B testing), or from models that use linear response functions. Evaluating the marginal returns to spending is more challenging and, with the exception of complex and expensive experiments, will usually involve models that include nonlinear response functions. Conceptually and practically, these three types of returns are different and they should not be compared to each other. Although diminishing returns will eventually be encountered, there is no general rule on which of the three measures of ROM will be higher or lower. Their relative values will depend on the shape of the response function and where on that function the return is evaluated. In other words, the critical difference among the three is the comparison or reference spending level. Because media impact on revenue is nonlinear, it matters a great deal which reference point is chosen. We therefore need to start the discussion with a summary of what is known about the advertising-to-sales response function.

Fundamental Advertising Response Phenomena

As media returns are derived, first and foremost, from the media’s impact on top-line revenue, we must recognize the specific nature of consumer response to media advertising. This response is *not* linear. Instead, it is characterized by the following five specific phenomena, which were first summarized in a seminal paper by John Little (1979):

- The steady-state response of sales to advertising is concave or S-shaped.
- Upward response is fast; downward response is slow.
- Competitive spending matters.
- Advertising effectiveness changes over time.
- Response can tail off even under constant spending.

Since this publication, a number of studies on advertising response have provided quantifications of advertising response that put these findings in a sharper perspective.

First, the predominant response function is concave, and the advertising elasticity (Advertising elasticity is the percent change in sales divided by the percent change in advertising. It implies that the sales change can be attributed to the advertising change.) empirical generalization is 0.1 (Sethuraman et al. 2011). Threshold effects that lead to S-shaped response functions may exist, but they are the exception (e.g., Rao and Miller 1975). In terms of overall sales sensitivity, advertising is the weakest of the marketing-mix instruments (see Hanssens et al. (2001) for details). That does not imply, however, that advertising is the least profitable instrument. It does imply that a profit-optimizing level of media spending exists, which we will elaborate on in a subsequent section.

Second, advertising elasticities are demonstrably higher for new products (elasticity about 0.3) than for established products (about 0.01) (Sethuraman et al. 2011). This is explained by the fact that, for both durables and consumables, advertising is stronger in creating awareness than in fostering preference. In particular, advertising has a stronger effect on trial rates than on repeat rates (Deighton et al. 1994). Initial awareness creation is a key for new products whereas, for more established products, prior consumer experience dominates. Indeed, the performance feedback loop (i.e., product usage experience or purchase reinforcement) is much stronger than advertising in determining future consumer choices. Hence, while advertising can be used to initiate trial, it alone is not sufficient to sustain repeat purchase without a favorable product evaluation. This helps explain the declining role of advertising over the product life cycle.

Third, visible short-term lifts are a *condition* for the existence of long-term effects. For example, an extensive experimental study by Lodish et al. (1995) showed that about *one-third* of television commercials showed a significant effect on sales in the first year. The long-term impact of these *effective* commercials is about twice the short-run effect. Thus ineffective media spending in the short run is unlikely to make a difference in the long run.

Fourth, while competitive spending matters, the ultimate effects of advertising are more influenced by the nature of consumer response itself than by the vigilance of competitors' reactions (Steenkamp et al. 2005). A straightforward approach to include competitive media spending is to use a "share of voice" metric in advertising response, i.e., brand spending divided by total (brand plus competition) spending in a certain time period.

Fifth, smaller competitors tend to have higher advertising-to-sales (A/S) ratios than market leaders (e.g., Tellis 2004). One explanation is that these smaller players

desire to grow their market share and need to invest in advertising more so than their competitors (in relative terms). Another explanation is that the market leaders have already developed strong assets in the form of extensive distribution and brand equity that make advertising less important for them, again in relative terms.

Finally, advertising wear-in and wear-out patterns help explain why ad response can tail off even under constant spending. For *consumables*, we know that people “learn faster than they forget,” which helps explain the different rise/decay rates in ad response. Over half a century after his published experiments, Zielske’s (1959) result that three to four exposures is best, still holds. If the brand continues to expose consumers beyond the fourth impression, the impact is expected to be drastically reduced. Furthermore, for *durables*, market rejuvenation is a key concept. Effective advertising for a durable product reduces the untapped market (buyers remaining), resulting in a loss of aggregate effectiveness. After some time has elapsed, the market is rejuvenated with new prospects, and a new campaign can once again be effective.

Many studies have focused on various qualitative aspects of advertising (see Vakratsas 2005 for a review of contributions). Among the most promising is work on eye movements that has revealed which aspects of a print ad (e.g., text, pictures, brand name, relative position on the page, etc.) are the most impactful (Pieters et al. 1999). Their results could well lead to a new, improved practice of copy writing. On the other hand, we know little about the *relative importance* of advertising quality and advertising quantity, e.g., can higher spending make up for poorer advertising quality?

Estimating Media Response Parameters

We now use these qualitative insights into the nature of advertising response to address the specification and estimation of media response parameters. In a subsequent section, we will address how these estimates of top-line impact are used in media return calculations that impact advertising decision-making.

The Shape of the Advertising-Response Function

All else equal, higher advertising spending is expected to increase sales for a variety of reasons. Acquiring previously unaware prospects, increasing purchase quantities, increasing brand switching in the direction of the advertised brand, and retaining a larger fraction of the existing customer base are among the major sources. At the same time, we expect there to be *diminishing returns* to these effects, again for several reasons. Consumers cease to be responsive once they have learned the basic message contained in the advertising (saturation effect), markets deplete as successful advertising causes purchasing which then removes the buyers from the market, at least temporarily (market-depletion effect) and, finally, there are natural ceilings to the number or percent of target customers that can be reached (ceiling effect). While

sales then still increase with increases in advertising support, each additional unit of advertising brings less in incremental sales than the previous unit did.

As a consequence, the basic advertising-response function is *nonlinear*. More specifically, it is expected to be concave, as in the following multiplicative model

$$S_t = e^c A_t^\beta X_t^\gamma Z_t^\delta e^{u_t}, \tag{2}$$

where S_t refers to sales or another performance metric in period t (for example, week t), A_t is the advertising support in that week, X_t refers to other elements of the marketing mix, Z_t corresponds to environmental factors, and u_t is an error term. For simplicity of exposition, we list only one X and one Z variable. The base response model may be estimated across time periods t but could also be specified over cross-sectional units $i = 1, \dots, I$ (for example geographical regions, market segments or individual consumers), or both. We expect $0 < \beta < 1$ in estimation, a condition which results in concavity of response.

The base model (2) implies that with infinite advertising comes infinite sales. In practice, however, there will be a limit or *ceiling* to sales, usually determined by prevailing market conditions. While there are other ways to represent concavity (see e.g., Hanssens et al. 2001, pp. 100–102), the multiplicative function is particularly appealing as it recognizes that media-mix effects interact with each other (i.e., the marginal sales effect of an incremental advertising dollar depends on the other elements in the equation). In addition, taking logarithms linearizes the model:

$$\ln(S_t) = c + \beta \ln(A_t) + \gamma \ln(X_t) + \delta \ln(Z_t) + u_t, \tag{3}$$

making it more easily estimable. Finally, the response parameters are easily interpreted as response elasticities. An example of an advertising response curve with elasticity $\beta = 0.13$ may be found in Fig. 2a.

In some cases, the response is S-shaped, i.e., there is a minimum or threshold-level of ad spend below which there is little or no impact, followed by a range of advertising spending with rapidly increasing sales response. At even higher spending levels (i.e., past a certain “inflection point”), the usual diminishing returns appear (see Fig. 2b). The base model (2) can readily be extended to an “odds” model that allows for S-shaped response, as demonstrated by Johansson (1979):

$$(S_t - I)/(K - S_t) = e^c A_t^\beta X_t^\gamma Z_t^\delta e^{u_t}, \tag{4}$$

where I is the minimum sales level (e.g., the level at zero media spend), and K is the ceiling level. For example, if sales is expressed in relative terms (i.e., market shares), I could be set at 0% and K at 100%. For advertising response parameters $0 < \beta < 1$, model (4) is still concave, but for $\beta > 1$, the function is S-shaped. Johansson (1979) discusses the formal estimation of (4) with maximum-likelihood methods, as well as an easy approximation based on ordinary least squares.

For all practical purposes, concavity and S-shape are sufficient functional forms to capture the essence of advertising response (We refer to Hanssens et al. (2001) for

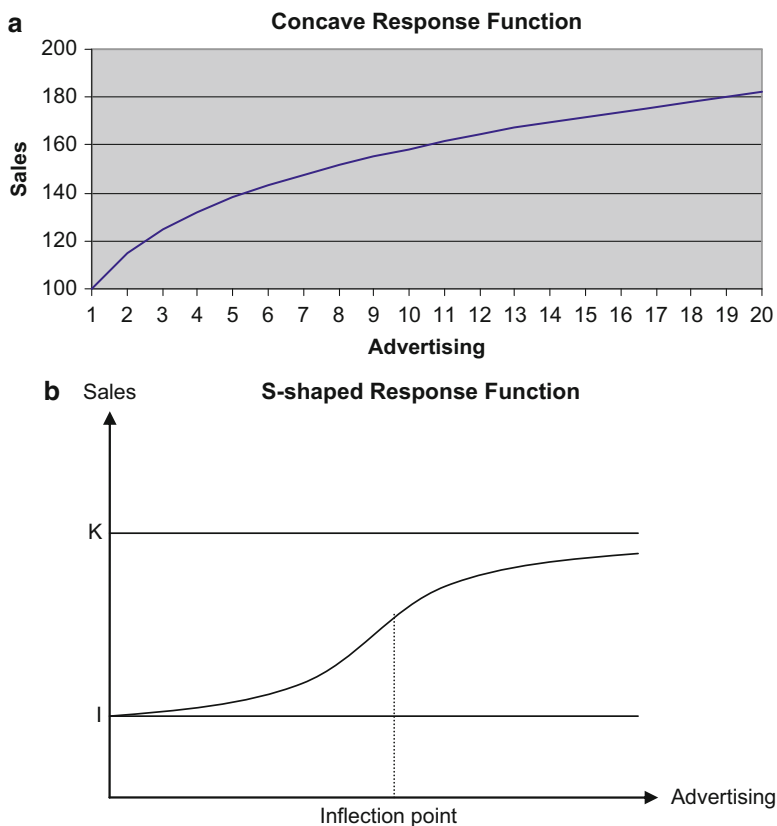


Fig. 2 (a) Concave response function (b) S-shaped response function

a review of other functional specifications that have been used in the literature.). In some cases, the response may be even simpler; if all the advertising spending observations lie in a restricted range of the data, the response function may well be approximated by a linear function. Of course, one should be careful not to make media effect inferences outside this restricted range.

Advertising-Response Dynamics

Advertising is communication, which can trigger memory, and thus its impact may easily extend beyond the campaign period. Similarly, customers could be exposed to a given advertising message (e.g., a magazine ad insert or pop-up digital ad) on multiple occasions. Therefore, response models that simply relate current sales to current advertising expenditures are likely to underestimate advertising's total impact. Provided good longitudinal data are available, advertising-response dynamics can be represented by extending the core model in (2) to a distributed-lag model over time t :

$$S_t = e^c A_t^{\beta(L)} X_t^\gamma Z_t^\delta e^{u_t}, \tag{5}$$

where $\beta(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \beta_3 L^3 + \dots$, and with L the lag operator (i.e., $L^2 A_t = A_{t-2}$).

In this model, the cumulative advertising effects are obtained as $\sum_i \beta^i$. The time-delayed effects of advertising captured in the lag-polynomial $\beta(L)$ are often referred to as carry-over effects and tend to vary across products, categories, and media. As it is hard to estimate a large number of β -parameters and/or to define a priori a precise cut-off point for the number of lags to consider, one often imposes a certain structure on the various β -parameters. A popular structure is the geometric decay implied in the Koyck model, in that $\beta_{i+1} = \lambda \beta_i$ (see e.g., Clarke 1976). Rather than a priori imposing a certain structure, one could also use empirical specification methods to determine the functional form of $\beta(L)$, such as the Box-Jenkins transfer-function method (see e.g., Helmer and Johansson 1977) or the Liu and Hanssens (1982) direct-lag specification method. A technical description of these methods is beyond the scope of this article but can be found in Hanssens et al. (2001, Chap. 7). We do want to point out, however, that the distributed lag-model in (5) is flexible enough to accommodate the well-known wear-in and wear-out effects of advertising. For example, if it takes two periods for any economic impact to take place, one can specify $\beta(L)$ as $\beta_2 L^2$. If, in addition, advertising effects wear out (decay) at a rate of 10% per period, the lag polynomial $\beta(L)$ can be written as $\beta_2 L^2 / (1 - 0.9L)$.

An alternative approach to assessing the dynamic effects of advertising is Adstock (Broadbent 1984), which has gained some popularity among practitioners. The Adstock model rests on the assumption that each advertising effort adds to a preexisting stock of advertising goodwill. Conversely, absent any current advertising, the stock decays at a constant rate. Thus, the empirical specification is

$$ADSTOCK_t = \alpha ADSTOCK_{t-1} + (1 - \alpha)A_t, \tag{6}$$

where α is the decay factor ($0 < \alpha < 1$). Adstock models are elegant in conceptualization and easy to use, as they circumvent the problem of empirical lag specification. However, the decay parameter is set subjectively and that may lead to a bias (usually an overestimation) in the advertising impact. Whenever possible, we recommend that decay parameters be estimated from the data. In addition, Adstock does not make the critical distinction between *purchase reinforcement* and *advertising carryover*. Suppose advertising generates trial for a new product, which leads to a positive consumption experience and, therefore, subsequent (repeat) sales. While advertising should be credited with generating the initial sales, it should share that credit with purchase reinforcement for subsequent sales. Givon and Horsky (1990) developed a distributed lag model to disentangle both effects. Their application in several product categories revealed that the purchase feedback effect dominates the advertising carryover effect.

Research on the dynamics of advertising impact has intensified in recent years, in part because of the availability of high-quality advertising response data, especially from digital media. This has created a need for more advanced estimation methods such as persistence modeling, dynamic linear models, and Bayesian

estimation. A detailed comparative review of these methods may be found in Leeflang et al. (2009).

Data-Interval Bias

The determination of advertising response dynamics is closely related to the chosen temporal data interval. In the past, data intervals posed a serious econometric problem because they were too coarse to allow for detailed inference. For example, sales and advertising movements sampled annually, or even quarterly, will almost inevitably represent a mixture of consumer response effects, firm decision, and competitive reaction effects. Thus, the nature of a *contemporaneous* correlation between advertising and sales is difficult to ascertain. Moreover, marketing researchers have long been puzzled by the empirical observation that estimated carry-over, and hence, advertising-duration effects, differed depending on whether dynamic advertising models were estimated on monthly, quarterly, or annual data. Conventional wisdom was then to use as “preferred” data interval the one that most closely resembled the brand’s interpurchase time (see e.g., Leone (1995) for a review). This view was challenged by Tellis and Franses (2006), who showed that one can retrieve the correct carryover effects with aggregate data provided one has information about the exact interexposure times. However, in many instances, applied marketing researchers do not know the exact media insertion patterns, let alone the interexposure time of consumers. In that case, using data at the lowest level of aggregation was found to be a good heuristic.

Fortunately, advertising databases have become available at much more disaggregate sampling levels. For example, weekly media advertising spending data are routinely collected by Kantar Media and Nielsen. When matched with weekly sales numbers, distributed lag models of advertising effects may readily be estimated. Contemporaneous correlations between advertising and sales can then safely be interpreted as consumer response effects, since it would be difficult for most organizations to react to competitive moves and/or to incorporate sales feedback in advertising spending within 1 week. In some cases, the data interval is reduced even further, for example to hourly measurements in the case of direct-response television campaigns (e.g., Tellis et al. 2000). Overall, it is fair to conclude that advances in data collection technology are gradually obsoleting the problem of temporal data interval bias.

Another form of data aggregation bias is cross-sectional. For example, national data may be used in an advertising response model, even though some regional markets receive different advertising exposures than others. In nonlinear models (as the multiplicative core model (2)), linear aggregation (e.g., when data of the individual regions are summed to a national aggregate) will create biased estimates of advertising’s effectiveness if advertising spending differed between the regions across which the summation took place. We refer to Christen et al. (1997) for an in-depth discussion of these issues.

Asymmetric Response

The core response model assumes that advertising effects are symmetric, so if adding 10% to the spending increases sales by 2%, then cutting the spending by 10% implies that sales will decrease by 2%. In reality, however, asymmetries may exist.

Consider the case where a successful new advertising campaign first acquires several new customers quickly, then brings in a lower, but still positive number of new customers. Similarly, at some point, a mature brand's sales may have become resistant to further advertising increases but quite sensitive to advertising reductions (Vakratsas 2005), a situation many managers may feel applies to their brand. Representing such a process requires that we model, separately, the effects of "starting," "continuing," and "ending" of a campaign. That problem was tackled by Simon (1982), who extends the core response model (with only contemporaneous response effects for simplicity of exposure) as follows:

$$S_t = e^c A_t \beta + \theta I_t X_t^\gamma Z_t^\delta e^{u_t}, \quad (7)$$

where $I_t = 1$ if the period corresponds to the beginning of a campaign, and $I_t = 0$ elsewhere. Similarly, one could add another time-varying dummy variable to denote the ending of the campaign. Thus, the start of a campaign has a total positive impact of $(\beta + \theta)$, and the remainder of the campaign has a "level" impact of β . Simon refers to the asymmetric effect θ as a "differential-stimulus" effect. This model was tested by Hanssens and Levien (1983) in the context of Navy advertising for recruits. Their lead generation model showed, for example, that radio advertising had a θ effect of 0.08 added to its level elasticity β of 0.26.

New data sources could provide even better insights into the asymmetric response nature of advertising. For example, in the case of a durable product or service, a company may be able to separate its weekly sales into two components: sales coming from new customers and sales from existing customers. Since we know that advertising has a stronger impact on trial sales than on repeat sales (e.g., Deighton et al. 1994), we may discover the asymmetries by applying the core model (2) separately to both components.

Dealing with Reverse Causality

Scientists like to make inferences from experimental data, i.e., where the treatment and control conditions are assigned randomly, so that the impact of the treatment is readily isolated from other influences on the dependent variable of interest. Marketing managers, however, would not last long if their allocation decisions were made randomly. In the case of advertising spending, for example, some months of the year may be favored over others, and some products or regional markets will likewise receive preferential allocation over others. In most cases, these allocations are made based on *ex ante expected sales performance*. For example, motion picture studios

routinely allocate substantial prelaunch advertising budgets to high-cost productions that are expected to become blockbusters, sometimes as part of their contractual obligations to high-profile actors.

These well-known managerial conditions create an inference problem called endogeneity or reverse causality. Econometrically, it means that the error term in model (2) may be correlated with some of the explanatory variables, in Casu advertising. Ordinary least-squares estimation will then lead to biased estimates of advertising's effects on sales. For example, if management picks successful products to advertise more heavily, an OLS estimator may confuse a true advertising effect with a "popular-product" effect, and thus overestimate the advertising impact.

Endogeneity is most problematic with cross-sectional data, where the modeler lacks the natural passing of time to establish the direction of effects. Standard econometric tests such as the Hausman test are available to diagnose endogeneity, and when detected, alternative estimators such as instrumental-variable methods may be used. We refer to Van Heerde et al. (2005) for an in-depth discussion of these methods. Even so, the most reliable inference of advertising impact on cross-sectional data comes from experimental designs in which some markets or customers are deliberately given different advertising treatments than others. A good example in this context are the V-8 advertising experiments reported in Eastlack and Rao (1986). Advertising experiments are now enjoying a resurgence because of the relative ease with which they can be executed with digital media, see for example Lambrecht and Tucker (2017).

On time-series data, endogeneity is less problematic provided one has sufficiently short data intervals (e.g., weekly data). It is generally the case that consumers can respond to advertising stimuli a lot faster than companies can adjust their advertising spending to changes in consumer demand. Thus, if an application of core model (2) reveals a 0.2 contemporaneous elasticity of advertising, we can be reasonably confident that this is a true advertising response effect. In addition, any lagged advertising response effects are by definition free of reverse causality effects.

It is important to understand that the presence of decision feedback loops and competitive reactions can create a long-run outcome of an advertising campaign that may be quite different from its short-run impact. In order to assess such long-run effects, the single-equation advertising response model in (2) is replaced by a dynamic system with multiple equations, a method called "persistence modeling." A discussion of persistence models is beyond the scope of this chapter and may be found in Dekimpe and Hanssens (1995) (Persistence modeling uses Vector Autoregressive Modeling (VAR) techniques.). Taken together, the combined chain-like effects of consumer response, decision feedback and competitive response to various marketing initiatives, including media advertising, can generate long-run sales impacts that are up to five times that of their short-term response, as demonstrated by Pauwels (2004) in the frozen-dinner market.

Differences Across Media

Not all advertising media are created equal, and in many cases, a marketer will want to understand the differential impact of spending in different media. The core response model allows for this by including as many explanatory variables in the response model as there are relevant media. For example, a media-mix model may be

$$S_t = e^c TV_t^{\beta_1^{(L)}} PRINT_t^{\beta_2^{(L)}} EMAIL_t^{\beta_3^{(L)}} X_t^\gamma Z_t^\delta e^{u_t}, \quad (8)$$

where the combined advertising expenditures A_t are now replaced by the respective expenditures in three media (as an illustration): TV, Print, and Email. Each medium has its own response elasticity and lag structure. Provided the database has sufficient degrees of freedom, and provided there is natural variation in the media spending patterns, media-mix models can be used to disentangle the relative contribution of each medium in explaining observed sales variations.

Media mix models have a long history starting with Montgomery and Silk (1972) who analyzed a pharmaceutical market and showed the differential impact of print advertising, direct mail, and sampling and literature. Dekimpe and Hanssens (1995) considered both print and TV advertising for a major home-improvement chain and found the medium with the lowest short-run effect (TV) to have the highest long-run impact.

At the qualitative level, one can make predictions of differential media impact by considering the information style delivered by each medium. Advertising is persuasive insofar as it fosters consumers' cognition, affect, and experience (Vakratsas and Ambler 1999). Specifically, print or internet advertising is strong on *cognition* because readers can absorb factual information about a product as long as they like (for example, reading about the performance characteristics of an automobile). TV, by contrast, is better at delivering *affect* through the use of sound and motion (for example, portraying the thrill of a car's swift acceleration). Communicating *experience* is often achieved with specific advertising content such as consumer testimonials and celebrity endorsements, for which several media may be used. By matching the delivery of cognition, affect, and experience to the needs of the advertised brand, managers can formulate initial hypotheses about media differences that are testable using model (8) and its variants.

Should a company invest all of its advertising spending in the medium with the highest lift? While such a decision rule is simple, it would be incorrect. Not only do sales benefit from synergistic effects of spending across the media but the laws of diminishing returns would make such a rule suboptimal. Instead, managers should allocate their advertising budget to media *in proportion to the media response elasticities*, the so-called Dorfman-Steiner conditions. While a formal proof of this proposition is beyond the scope of this chapter, it is readily obtained by maximizing the profit function accompanying response model (8). See, for example Hanssens et al. (2001) for a more detailed discussion. Explicitly allowing for synergistic effects

between two media which are each characterized by diminishing returns, Naik and Raman (2003) demonstrated that, as synergy increases, advertisers should not only increase the total budget but also allocate more funds to the less effective medium.

Advertising Copy and Creative Effects

Creativity of communication is an integral part of advertising design and execution and can have a substantial impact on the persuasive appeal of advertising. From a response modeling perspective, creative impact can be measured in a number of ways. First, one could use categorical variables to distinguish between different creative executions (e.g., campaigns). A simple example with two campaign executions may be

$$S_t = e^c A_t^{\beta_1(L) + \beta_2(L) E_t} X_t^\gamma Z_t^\delta e^{u_t}, \quad (9)$$

where $E_t = 0$ for the base execution, and $E_t = 1$ for the new execution. Insofar as the new copy had a different impact on sales than the previous execution, the terms in $\beta_2(L)$ will be different from zero. Alternatively, one could try to directly import creative-quality metrics into the market-response model. For example, suppose a panel of experts rate the creative execution of each advertising campaign on a 5-point scale. Then the basic advertising response model may be extended to distinguish between spending elasticities and creative-execution elasticities as follows:

$$S_t = e^c A_t^{\beta_1(L) + \beta_2(L) Q_t} X_t^\gamma Z_t^\delta e^{u_t}, \quad (10)$$

where the $\beta_1(L)$ terms now measure the pure spending effects of advertising, and the $\beta_2(L)$ terms measure how these effects increase with higher-quality creative, measured by Q_t . This approach was used in a Bayesian framework for measuring advertising copy effects by Gatignon (1984).

The internet age promises to bring new insights from the measurement of digital communication content in general and the creative quality of advertising in particular. For example, text-mining algorithms are used by Culotta and Cutler (2016) to monitor brand messages on Twitter and by Trusov et al. (2016) to target consumers for advertising based on their web-surfing behavior.

Intermediate Performance Metrics

So far, the core response model (2) and its extensions have focused on cases where the advertising→sales relationship can be inferred in a direct way, which is the most relevant for assessing accountability and financial planning. While these represent the majority of applications, there are instances of relatively long sales cycles, notably in some business-to-business markets, where advertising's impact is better represented in stages. For example, media spending may create awareness of a

certain business-to-business offering; however, that awareness is insufficient to create a sale. Instead, there need to be follow-up sales calls to an initial inquiry, sometimes multiple calls, to guide the customer to an eventual product adoption.

Intermediate media performance metrics may include consumer awareness, consumer consideration, and consumer preference metrics. They are generally survey based and are not financial in nature and therefore do not readily lend themselves to return calculations. However, they can be used as intermediate dependent variables in a chain (system of equations), where each equation is similar in design to the core response model (5). For example, media spending impacts consumer awareness in an awareness equation. This, in turn, converts to sales in a separate sales equation. The ultimate return of media spending then depends on their initial impact on awareness multiplied by the awareness-to-sales conversion rate. Hanssens et al. (2014) developed such a system for several brands in four consumer product categories over a 7-year period. They reported that the model with intermediate performance metrics provided superior performance in terms of sales prediction accuracy as well as quality of managerial recommendations.

A major challenge with intermediate performance metrics is that they should be collected at similar data intervals as the sales data, which is often unrealistic. For example, sales data are available on a weekly or even daily basis, but consumer preference surveys are conducted only quarterly. The advent of digital data sources, however, provides new metrics that can prove to be useful. These include Google queries (a proxy for consumer interest) and Facebook likes (a proxy for consumer sentiment). Figure 4 illustrates how different digital response metrics impact the estimation of media effectiveness, as discussed below.

Deriving Media Returns from the Estimated Response Parameters

Having obtained some reliable estimates of sales response to media spending (i.e., the *top-line* effects), how are media *return* estimates derived? Assuming a constant profit margin, the net cash flows (CF) in period t – excluding nonmedia costs – may be expressed as

$$CF_t = S_t * \text{margin} - A_t, \quad (11)$$

where A_t is total advertising media spending.

The *total* return on media spending A is then obtained as

$$ROM = [CF(A) - CF(A = 0)]/A \quad (12)$$

Note that ROM is a ratio, which is useful for an ex-post assessment of the return of a specific media campaign or investment. However, as pointed out earlier, total ROM should *not* be used to determine optimal levels of media spending. Doing so will often result in underinvesting on media, because ROM typically declines monotonically with higher spending (see Ambler and Roberts 2008 for an

elaboration). Instead, the optimal media spend A^* may be derived from maximizing the cash-flow function (11) based on the response model (5):

$$A^* = \left[e^{c'} \beta(L)^* \text{margin} \right]^{1/[1-\beta(L)]}, \quad (13)$$

where we have incorporated the effects of other firm-controlled variables X and environmental conditions Z into the adjusted baseline $e^{c'}$ for ease of exposition. It is easy to demonstrate that, at this optimal spend level, the *marginal* ROM is zero.

Importantly, the relationship between media spending and cash flow generation depends on (i) the natural size (the baseline) of the business, (ii) the productivity of media spending $\beta(L)$, and (iii) the prevailing profit margin. Taken together, they fully determine optimal short-run media-resource allocation. At the same time, these determinants are exogenous; for example, it is assumed that more aggressive media spending has no impact on either the baseline or media effectiveness itself. Thus, the decision rule in (13) may be thought of as a harvesting or reactive view of media resource allocation. We will discuss in a subsequent section how to incorporate *asset-building* effects of media spending, in particular brand equity and customer equity.

The important take-away from the analysis so far is that only *one* ROM definition is useful for media allocation decisions: the *marginal* media return should be zero across all the media in the mix. Positive marginal returns indicate *underspending* and negative values suggest *overspending* in that medium.

A corollary of this basic result is that, while total ROM estimates can always be made, they are useful only for comparing the impact of *equal* spend levels. For example, if total media ROI is 75% for TV spending and 60% for internet spending, you cannot infer that TV is a more effective medium *unless* the two have the same spending levels. Indeed, the internet spending may *appear* to be a less impactful medium, whereas, in reality, the medium is highly effective but offers a lower ROM due to high spending relative to TV.

Finally, note that it is difficult to obtain marginal ROM estimates – and, therefore, optimal advertising spending levels – from advertising experiments, because these generally result in only two or three points of the response curve. Media experiments are more useful for assessing the causal impact of media on sales and for testing different creative executions.

Path-to-Purchase and Attribution Models for Digital Media

The media return measurement approach described above applies to all media and as such is widely used. In recent years, the availability of so-called *path to purchase* data in the online world has resulted in a different approach for ROM assessment. The central idea is that, at any point in time, individual consumers have different purchase probabilities for a given category and brand, and they reveal these probabilities by their online activities, including searching, blogging, and clicking on

various online ads. These methods are generally known as *digital attribution* methods.

Initially, digital attribution models used a simple *last click* heuristic, i.e., whichever medium the consumer used just prior to online purchase was given full credit for the purchase conversion. Subsequently, more sophisticated models recognized that prior media interventions may deserve some conversion credit as well, by recognizing that consumers move through various stages in the purchase funnel over time. A well-known digital attribution technique (Li and Kannan 2014) makes the distinction between customer-initiated web actions (e.g., conducting a search, visiting a website) and firm-initiated channels (e.g., delivering a pop-up display ad, sending an e-mail). These actions change the consumer’s purchase probability over time, to different degrees. Figure 3 presents the conceptual framework underlying the authors’ digital attribution method.

The key challenge in digital attribution is to estimate the *incremental* purchase probability achieved by a certain media intervention. To do so, the authors set up a model that incorporates consumers’ movement from brand consideration stage to visit stage to purchase stage and estimate it with Bayesian methods. Their empirical analysis in the hospitality industry revealed strong differences across the media in their ability to convert consumers’ apparent interest in a product to actual purchase.

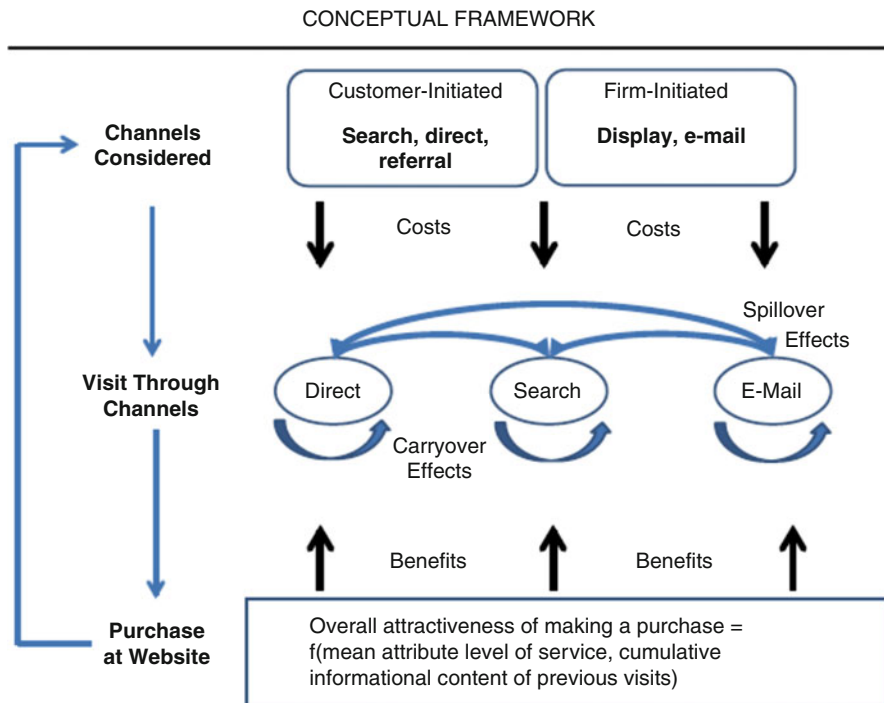


Fig. 3 Conceptual digital attribution model (Source: Li and Kannan (2014))

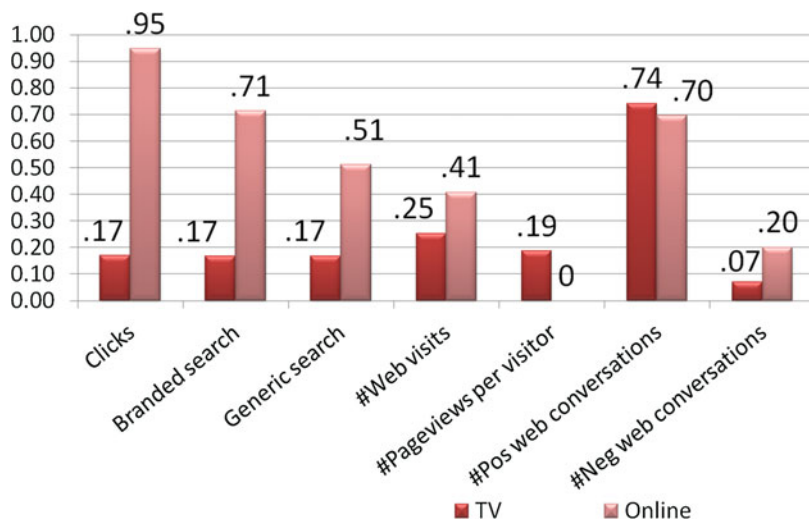


Fig. 4 Comparison of TV and online media elasticities on online performance metrics (Source: Pauwels and van Ewijk (2013))

The differences in estimated response effects of digital media inside the consumer purchase funnel are illustrated in Fig. 4 (Pauwels and van Ewijk 2013). One important conclusion from the figure is that, the deeper one goes into the consumer purchase funnel, the lower the estimated media elasticities. The figure also shows that digital elasticities are generally higher than TV elasticities, *except* on the criterion of generating positive web conversations. This illustrates the principle that traditional media (such as TV) can “drive consumers to the web” and, as such, still provide an important role in the overall media mix.

An in-depth review of digital marketing, including attribution models, may be found in Kannan and Li (2017). Overall, digital attribution models offer substantial promise in the analysis of digital media ROI. At the same time, they depend on accurate clickstream data and, as such, are limited to online purchasing. The use of digital attribution models in industry will increase with their ability to act on consumer actions in a “just-in-time” fashion. For example, *mobile* advertising messages are particularly adept at reaching consumers who are close to a purchase occasion.

Media Advertising and Asset Creation

A prevailing belief among practitioners and academics is that well-placed media spending not only stimulates sales but also builds *future assets* for the firm. In order to represent that capability of media, we must extend the core response model to account for endogenously created assets that, in turn, will generate future cash flows.

Not doing so may result in underestimating media returns, because only the cash flows coming from incremental sales effects are incorporated. By contrast, by considering *stock metrics* of business performance, in addition to cash flows, we may represent important long-term media effects. Among those stock metrics, the most important are brand equity and customer equity.

Brand Equity

Perhaps the most frequently studied stock metric in the marketing literature is the concept of brand equity. An excellent review is given in Ailawadi et al. (2003), who propose the *revenue premium* as a financially relevant measure for the value of a brand in a given industry. The revenue premium is defined as the difference in revenue realized by branded vs. unbranded competitors, i.e.,

$$\text{Revenue premium} = \text{volume}_{\text{brand}} * \text{price}_{\text{brand}} - \text{volume}_{\text{nonbrand}} * \text{price}_{\text{nonbrand}} \quad (14)$$

This reflects the idea that brand equity may boost sales volume, allow for a price premium, or both. Put differently, brand-building activities may enhance future cash flows as a result of realizing a higher sales volume and/or a higher price. The measure is shown to be actionable, stable over time, and to have considerable diagnostic value in terms of the brand's long-run health, thereby conforming to our earlier criteria. Interestingly, Ailawadi et al. (2003) also demonstrate how branded products exhibit asymmetric upward and downward (market-share) price elasticities. Using data from a variety of consumer-packaged products, they derive that low-revenue premium brands have an average down price elasticity of -1.195 , and an average up elasticity of -0.921 . High-equity brands, in contrast, have an average down share elasticity of -0.747 , and an up elasticity of only -0.183 . Hence, while brands with a higher revenue premium gain some share when they reduce their prices, they lose comparatively less share when they increase their price. As such, brand equity is a stock metric that enhances future cash flows through three different routes described earlier: higher baseline sales (volume premium), higher profit margins (price premium), and increased marketing effectiveness (differential $\beta(L)$).

Note that some marketing activity may deteriorate brand equity. For example, Mela et al. (1997) used time-varying response models to demonstrate that increasing the frequency of sales promotions may increase customers' price sensitivity to the brand. As a result, either a smaller percent of sales is generated at full price, or the brand's price premium is lowered. Both scenarios result in damage to the brand's equity.

From a media return perspective, the revenue premium that captures brand equity in (14) is typically estimated using the sales-response model (5) for different brands in a category and examining differences in the intercept and slope parameters. A time-varying-parameter version of this model may also be used in this context.

Customer Equity

While brand equity focuses on the supply side, i.e., the offerings of the firm, customer equity (CE) is an asset valued on the demand side, with specific reference to the firm's customer base. Customer lifetime value (CLV) is generally defined as the present value of all future profits obtained from a customer over his/her life of relationship with a firm (Gupta et al. 2004):

$$CLV = \sum_{t=0}^T \frac{(p_t - c_t)r_t}{(1+i)^t} - AC \quad (15)$$

where p_t = revenue generated by a consumer at time t , c_t = direct cost of servicing the customer at time t , i = discount rate or cost of capital for the firm, r_t = probability of customer repeat buying or being "alive" at time t , AC = customer acquisition cost, and T = time horizon for estimating CLV.

Customer equity (CE) is the sum of the firm's customers' lifetime values. CLV and CE measure "net present value" from a customer asset perspective, and thus speak to both shareholder value and customer value.

Marketing spending may impact customer equity in several ways: through acquiring new customers (at a cost AC per customer), through retaining existing customers (at a servicing cost c_t in each period), and through increasing per-customer revenue, which is sometimes referred to as *share of wallet*. In relationship businesses such as insurance and financial services, these effects can be quantified through direct counting of customers and aggregation of their CLVs. In most cases, however, the direct-count approach is not feasible or practical, and we should infer marketing's impact on customer equity at a more aggregate level (see e.g., Rust et al. 2004). This may be achieved by examining marketing's role in *purchase reinforcement*, i.e., using an existing sale to create more future sales from that customer. Purchase reinforcement modeling applies mainly in frequently purchased product and service categories, where consumers have reason to expect a similar-quality experience between one purchase occasion and the next. Givon and Horsky (1990) developed a market-share model that contrasts the impact of purchase experience (β) relative to advertising-induced retention (λ) as follows:

$$\text{Share}_t = \alpha(1 - \lambda) + (\beta + \lambda)\text{Share}_{t-1} - \beta \lambda \text{Share}_{t-2} + \gamma \text{Adshare}_t + e_t \quad (16)$$

This model is a special case of the dynamic core response function (5) with two-period dynamics. Thus it lends itself to calculations of the cash-flow impact (and therefore return) of investments in advertising vs. customer service provision. In their empirical investigation of four frequently purchased product categories, the authors reported that $\beta > \lambda$, i.e., the impact of purchase experience exceeds that of advertising spending. As such, even without renewed instantaneous media support, a stock effect is at work that results in future sales.

Since then, more complex models have been developed that infer movements in customer equity from sales transactions data and brand-related marketing actions

in a variety of sectors. For example, Hanssens et al. (2008) explored the impact of various marketing activities and external factors on the growth in customer equity for a major financial institution. Customer equity has, in various studies, been found to be an actionable and stable metric, which offers reliable guidance and an explicit linkage to financial performance (see e.g., Gupta and Lehmann 2005 for a review).

Other Response Metrics

Our discussion of return on media would not be complete without considering other, indirect ways in which media spending can enhance firm performance and, therefore, be subject to return calculations. We list below a number of important alternative performance metrics, discussed in Dekimpe and Hanssens (2011), and we make some observations about how advertising is known to affect them. Note that the response models for these alternative metrics are generally similar in nature to the models we have discussed in this chapter.

Protecting or enhancing price premiums. There is evidence that, ceteris paribus, nonprice advertising leads to lower price sensitivity and hence the ability to charge higher prices (e.g., Farris and Albion 1980). Note that, by the same token, price advertising may increase price sensitivity (see also Kaul and Wittink 1995).

Enhancing sales-call effectiveness. Advertising support may pre-educate a prospect so that subsequent sales calls have a higher chance of success. For example, Gatignon and Hanssens (1987) found this to be the case in military recruitment, i.e., the effectiveness of recruiting efforts increases with higher levels of media support.

Building distribution. When the trade makes stocking decisions based on anticipated consumer demand and when they perceive that demand to be influenced by advertising, higher distribution levels may be obtained (e.g., Parsons 1974). Since the distribution-to-sales elasticity is known to be high, this indirect route of media impact can be quite substantial.

Motivating employees. Advertising may have an “internal” audience in addition to its usual external audience (e.g., Gilly and Wolfenbarger 1998). This is particularly important in service-intensive industries, for example the hospitality sector, where “happy employees” can contribute to customer satisfaction.

Lifting stock price. Investors are exposed to advertising as much as consumers are. Evidence from the PC industry suggests that advertising may increase stock prices above and beyond the effect expected from an increase in sales and profits (Joshi and Hanssens 2010). In addition, advertising can reduce earnings volatility, which also impacts stock prices positively. This is because media spending is relatively easily manipulated and is fully amortized in the period in which it occurs. That makes it a convenient expenditure category to expand in times of strong earnings and reduce in times of weak earnings, thereby reducing the firm’s earnings volatility and its cost of capital. This strategic use of advertising is enhanced by the finding that, over long time periods, higher media expenditures are associated with lower systemic risk of the firm (McAlister et al. 2007).

Conclusion

This chapter has reviewed the analytical aspects of determining the return on media (ROM), which is a managerial task of increasing importance given the proliferation of new, technology-enabled advertising media. We first provided a tight definition of return on media that allows marketers to speak the language of finance, which is important in budget allocations. We then reviewed some common mistakes in ROM interpretation found in industry and concluded that media return measurement should start with assessing the top-line effects (i.e., on brand sales revenue) of media spending. This led us to formulate the key phenomena of advertising response that should be present in models of media return. Next, we reviewed several challenges that arise in the measurement of advertising effects, including the shape of the advertising response function, dynamic effects of advertising, optimal data intervals for measurement, asymmetric response, reverse causality, differences in media effects, copy and creative impact, and intermediate performance metrics. Next, we derived media return values from sales response models and discussed their use in media mix allocation. Finally, we extended the measurement of media returns to purchase path analyses of digital media, to the incorporation of asset building effects such as brand equity and customer equity, and to the use of various alternative business performance metrics.

Acknowledgment I am grateful to coauthors in other publications that have helped shape the content of this chapter, in particular material from Hanssens et al. (2001), Dekimpe and Hanssens (2007, 2011), Hanssens and Dekimpe (2008), and Farris et al. (2015).

References

- Ailawadi, K. L., Lehmann, D. R., & Neslin, S. A. (2003). Revenue premium as an outcome measure of brand equity. *Journal of Marketing*, *67*, 1–17.
- Ambler, T., & Roberts, J. H. (2008). Assessing marketing performance: Don't settle for a silver metric. *Journal of Marketing Management*, *24*(7–8), 733–750.
- Broadbent, S. (1984). Modeling with Adstock. *Journal of the Market Research Society*, *26*, 295–312.
- Christen, M., Gupta, S., Porter, J. C., Staelin, R., & Wittink, D. R. (1997). Using market level data to understand promotional effects in a nonlinear model. *Journal of Marketing Research*, *34*(3), 322–334.
- Clarke, D. G. (1976). Econometric measurement of the duration of advertising effect on sales. *Journal of Marketing Research*, *16*, 286–289.
- Culotta, A., & Cutler, J. (2016). Mining brand perceptions from twitter social networks. *Marketing Science*, *25*(3), 343–362.
- Deighton, J., Henderson, C., & Neslin, S. (1994). The effects of advertising on brand switching and repeat purchasing. *Journal of Marketing Research*, *31*, 28–42.
- Dekimpe, M. G., & Hanssens, D. (1995). The persistence of marketing effects on sales. *Marketing Science*, *14*, 1–21.
- Dekimpe, M. G., & Hanssens, D. M. (2007). Advertising response modeling. In G. Tellis & T. Ambler (Eds.), *Handbook of advertising*. London: Sage Publications.

- Dekimpe, M. G., & Hanssens, D. M. (2011). The hidden powers of advertising investments. In J. Wierenga, P. Verhoef, & J. Hoekstra (Eds.), *Liber Amicorum in honor of Peter S.H. Leeflang*. Groningen: Rijksuniversiteit.
- Eastlack, J. O., Jr., & Rao, A. G. (1986). Modeling response to advertising and pricing changes for 'V-8' cocktail vegetable juice. *Marketing Science*, 5, 245–259.
- Farris, P. W., & Albion, M. S. (1980). The impact of advertising on the price of consumer products. *Journal of Advertising Research*, 44, 17–35.
- Farris, P., Hanssens, D. M., Lenskold, J., & Reibstein, D. R. (2015). Marketing return on investment: Seeking clarity for concept and measurement. *Applied Marketing Analytics*, 1(3), 267–282.
- Gatignon, H. (1984). Toward a methodology for measuring advertising copy effects. *Marketing Science*, 3(4), 308–326.
- Gatignon, H., & Hanssens, D. M. (1987). Modeling marketing interactions with application to salesforce effectiveness. *Journal of Marketing Research*, 24, 247–257.
- Gilly, M., & Wolfinbarger, M. (1998). Advertising's internal audience. *Journal of Marketing*, 62, 69–88.
- Givon, M., & Horsky, D. (1990). Untangling the effects of purchase reinforcement and advertising carryover. *Marketing Science*, 9(2), 171–187.
- Gupta, S., & Lehmann, D. R. (2005). *Managing customers as investments*. Upper Saddle River: Wharton School Publishing/Pearson-Financial Times.
- Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. *Journal of Marketing Research*, 41, 7–18.
- Hanssens, D. M., & Dekimpe, M. G. (2008). Models for the financial performance effects of marketing. In B. Wierenga (Ed.), *Handbook of marketing decision models*. New York: Springer Science. Second edition forthcoming, 2017.
- Hanssens, D. M., & Levien, H. A. (1983). An econometric study of recruitment marketing in the U.S. navy. *Management Science*, 29, 1167–1184.
- Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2001). *Market response models* (2nd ed.). Boston: Kluwer Academic Publishers.
- Hanssens, D. M., Pauwels, K. H., Srinivasan, S., Vanhuele, M., & Yildirim, G. (2014). Consumer attitude metrics for guiding marketing mix decisions. *Marketing Science*, 33, 534–550.
- Hanssens, D. M., Thorpe, D., & Finkbeiner, C. (2008). Marketing when customer equity matters. *Harvard Business Review*, 86, 117–123.
- Helmer, R. M., & Johansson, J. K. (1977). An exposition of the box-Jenkins transfer function analysis with application to the advertising-sales relationship. *Journal of Marketing Research*, 14, 227–239.
- Johansson, J. K. (1979). Advertising and the S-curve: A new approach. *Journal of Marketing Research*, 16, 346–354.
- Joshi, A., & Hanssens, D. M. (2010). The direct and indirect effects of advertising spending on firm value. *Journal of Marketing*, 74(1), 20–33.
- Kannan, P. K., & Li, H. (2017). Digital marketing: A framework, review and research agenda. *International Journal of Research in Marketing*, 34, 22–45.
- Kaul, A., & Wittink, D. R. (1995). Empirical generalizations about the impact of advertising on price sensitivity and price. *Marketing Science*, 14(3), G151–G160.
- Lambrecht, A., & Tucker, C. E. (2017). Field experiments. In N. Mizik, & D. M. Hanssens (Eds.), *Handbook of marketing analytics: Methods and applications in marketing management, public policy and litigation support*. London: Edward Elgar, forthcoming.
- Leeflang, P., Bijmolt, T., van Doorn, J., Hanssens, D., van Heerde, H., Verhoef, P., & Wierenga, J. (2009). Lift versus base: Current trends in marketing dynamics. *International Journal of Research in Marketing*, 26(1), 13–20.
- Li, H., & Kannan, P. K. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research*, 51, 40–56.

- Leone, R. P. (1995). Generalizing what is known of temporal aggregation and advertising carryover. *Marketing Science, 14*(3), G141–G150.
- Little, J. D. C. (1979). Aggregate advertising models: The state of the art. *Operations Research, 27*, 629–667.
- Liu, L.-M., & Hanssens, D. M. (1982). Identification of multiple-input transfer function models. *Communication in Statistics Theory and Methods, 11*(3), 297–314.
- Lodish, L. M., Magid, A., Kalmenson, S., Livelsberger, J., Beth, L., Richardson, B., & Stevens, M. E. (1995). How TV advertising works: A meta-analysis of 389 real world split cable TV advertising experiments. *Journal of Marketing Research, 32*, 125–139.
- McAlister, L., Srinivasan, R., & Kim, M. C. (2007). Advertising, research and development, and systematic risk of the firm. *Journal of Marketing, 71*(1), 35–48.
- Mela, C. F., Gupta, S., & Lehmann, D. R. (1997). The long-term impact of promotion and advertising on consumer brand choice. *Journal of Marketing Research, 34*(2), 248–261.
- Montgomery, D. B., & Silk, A. J. (1972). Estimating dynamic effects of marketing communication expenditures. *Management Science, 18*, B485–B501.
- Naik, P. A., & Raman, K. (2003). Understanding the impact of synergy in multimarketing communications. *Journal of Marketing Research, 40*, 375–388.
- Parsons, L. J. (1974). An econometric analysis of advertising, retail availability, and sales of a new brand. *Management Science, 20*, 938–947.
- Pauwels, K. (2004). How dynamic consumer response, competitor response, company support and company inertia shape long-term marketing effectiveness. *Marketing Science, 23*(4), 596–610.
- Pauwels, K., & Bernadette, van E. (2013). Do online behavior tracking or attitude survey metrics drive brand sales? An integrative model of attitudes and actions on the consumer boulevard, Report 13–118. Cambridge, MA: Marketing Science Institute.
- Pieters, R., Rosbergen, E., & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research, 36*, 424–438.
- Rao, A., & Miller, P. B. (1975). Advertising/sales response functions. *Journal of Advertising Research, 15*, 7–15.
- Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing, 68*, 109–127.
- Sethuraman, R., Tellis, G., & Briesch, R. (2011). How well does advertising work? Generalizations from meta-analysis of brand advertising elasticities. *Journal of Marketing Research, 48*(3), 457–471.
- Simon, H. (1982). ADPULS: An advertising model with wearout and pulsation. *Journal of Marketing Research, 19*, 352–363.
- Steenkamp, J.-B. E. M., Nijs, V. R., Hanssens, D. M., & Dekimpe, M. G. (2005). Competitive reactions and the cross-sales effects of advertising and promotion. *Marketing Science, 24*(1), 35–54.
- Tellis, G. (2004). *Effective advertising*. Thousand Oaks: Sage Publications.
- Tellis, G., Chandy, R. K., & Thaivanich, P. (2000). Which ads work, when, where, and how often? Modeling the effects of direct television advertising. *Journal of Marketing Research, 37*, 32–46.
- Tellis, G., & Franses, P. H. (2006). Optimal data interval for estimating advertising response. *Marketing Science, 25*(3), 217–229.
- Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science, 25*(3), 405–426.
- Vakratsas, D. (2005). Advertising response models with managerial impact: An agenda for the future. *Applied Stochastic Models in Business and Industry, 21*(4–5), 351–361.
- Vakratsas, D., & Ambler, T. (1999). How advertising works: What do we really know? *Journal of Marketing, 63*(1), 26–43.
- Van Heerde, H. J., Dekimpe, M. G., & Putsis, W. P., Jr. (2005). Marketing models and the Lucas critique. *Journal of Marketing Research, 42*, 15–21.
- Zielske, H. (1959). The remembering and forgetting of advertising. *Journal of Marketing, 23*, 239–243.

Index

A

Abnormal stock return, 1043
A/B test, 4, 32, 39, 43, 44, 46, 47, 51–55, 57, 60, 823, 825
Accessibility-diagnostics theory, 72
Acquiescence, 83, 84
Adapted etic model, 125
Adaptive Conjoint Analysis (ACA), 785
Adjacency matrix, 698
Adjusted R-squared (Adjusted R^2), 304, 305, 308
Advertising, 671–672
 elasticity, 187
 endogeneity, 189
 visuals, 680
Aggregate regressors, 169
Aggregation bias, 1060
Akaike information criteria (AIC), 259, 485, 499
Akaike weights, 611
Alpha inflation, 276
Ambiguous temporal precedence, 49
Amoroso-Robinson theorem, 188
Analysis of covariance (ANCOVA), 267, 292
Analysis of variance (ANOVA), 266
Anderson-Hsiao, 1047
A priori contrasts, 276
Arellano-Bond approach, 460, 1047
Artefactual field experiments, 19, 42
Attenuation bias, 443
Attribute importance, 804
Attribute levels, 783
Attribution models, 1089
Augmented Dickey Fuller (ADF) method, 473, 474, 495, 497, 520
Autocorrelation, 310, 311, 313–316, 321, 325
Autocorrelation function (ACF), 475–477, 495, 497, 498
Automated content analysis, 645, 657, 658

Automated text analysis, *see* Text analysis
Autoregressive (AR), 471, 472, 497, 498
 coefficient, 1047
Autoregressive integrated moving average (ARIMA), 474
 ACF, 475–477, 495, 497–499
 log transformation, 495
 PACF, 475–477, 495, 497–499
 stationary tests, 495–497
Autoregressive moving average (ARMA), 471–474
Average treatment effect (ATE), 44–47
Average variance extracted (AVE), 565, 578, 604

B

Back-translation, 135
Barplot, 268
Baseline model, 382–383, 1066
Bayesian estimation, 387
Bayesian inference
 beta-binomial model, 732–734
 binomial probit likelihood, 736–737
 conditional posterior distributions, 742–744
 conditional posterior distributions in hierarchical models, 745–746
 data augmentation, 740–741
 Gibbs sampler, 738–740
 Metropolis-Hastings algorithm, 746–748
 normal-normal model, 734–736
 software implementation, 765
Bayesian information criterion (BIC), 259, 485, 499, 540, 611, 1025
Bayesian missing data, 169
 problems, 149–152
Bayesian model
 Bayesian estimation, 721, 731–732
 computational resources, 721

- Bayesian model (*cont.*)
- developments in, 759–761
 - goals, 723
 - hierarchical multinomial logit model, 766–769
 - inference about parameters, 721
 - likelihood function, 722
 - mediation analysis, 769–771
 - popularity, 720
 - predictions, 744–745
 - prior distributions, 722–724
- Bayesian shrinkage, 161
- Bayesian structural equation modeling (BSEM), 570
- Becker–DeGroot–Marschak (BDM) mechanism, 798, 979, 980
- BehaviorScan, 1060
- Bernoulli experiment, 732
- Best-worst scaling, 797
- Beta-binomial model, 732–734
- Beta-geometric/NBD model, 1012
- See also* Pareto/NBD model
- Beta-prior, 733
- Between-subjects design, 25, 266
- Between variance, 376, 377, 379, 382, 384, 395
- Bias, 187, 192, 201, 214
- Bias-corrected and accelerated (BCa) method, 605
- Biases, in survey research
- measurement errors, 74–85
 - representation errors, 85–89
- Big data, 720, 854
- analytics, 38
- Binary data, 164
- Binomial logit model, 209
- Binomial probit likelihood, 736–737
- Blind tests, 1039
- Blocked Gibbs samplers, 741–742
- Blundell–Bond approach, 460
- Bootstrapping, 606
- approach, 865–867, 875
 - confidence interval, 605
 - techniques, 208
- Business field experiments, 38
- Business optimization, 60
- C**
- Canonical correlation coefficient, 338
- Carryover effects, 27, 527
- Categorical data, 163, 167, 168
- Category equivalence, 127
- Cauchy prior, 157
- Causal inference, 44, 186
- blockage design, 893
 - causal design, 39
 - causal indicators, 592
 - causal statements, 828
 - causal steps approach, 863
 - challenge to, 844
 - concurrent double randomization design, 893
 - consistency design, 893
 - double randomized experiments, 892
 - effect, 40, 43, 44
 - of marketing, 191
 - of prices, 194, 195
 - enhancement design, 893
 - experimental methods, 891, 893
 - measurement-of-mediation design, 893
 - in mediation analysis, 891
 - from non-equal groups, 829
 - parallel design, 893
 - pattern matching, 893
 - prerequisite of, 853
 - in quasi field experiments, 841
 - requirements for, 839
 - specificity design, 893
 - within-subject design, 894
- Cause-effect relationships, 5, 14, 19, 32
- Cause indicators, 552
- Centering, 395, 389–390, 896–897
- group mean centering 389, 390, 395, 403
 - grand mean centering 389, 403
 - standardization 306–307
- Centroid scheme, 596
- K*-centroids cluster analysis, 245
- Certainty approach, 979
- Chi-square goodness-of-fit test, 558, 561, 567
- Choice-based conjoint analysis (CBC), 783, 976
- Choice design, 795
- Choice experiment, 795
- Choice model, 787
- Choice set, 784
- Classification matrix, 686
- Closed questions, 98
- Closeness centrality, 701
- Cluster analysis
- applications, 226
 - k*-centroids, 240
 - characteristics, 222
 - history, 222
 - of market basket data, 229–244
 - partitioning, 229
 - steps to conduct, 224

- Cluster-robust standard errors, 428, 430, 433, 443, 445, 449, 453, 455
- Cluster sample, 51
- Coefficient of determination (R^2), 304, 609
- Cointegration test, 481, 483, 521
- Collaborative translation, 135, 136
- Collinearity assessment, 606
- Color histograms, 681
- Common method bias (CMB), 571, 580
 - categories, 75
 - conceptualization, 75–76
 - definition, 75
 - different data sources, 77
 - proximal separation, 78
 - psychological separation of scale formats, 78
 - reasons for occurrence, 76–77
 - statistical remedies, 78
 - temporal separation of measurement, 77–78
- Communality, 604
- Competitive market structure (CMS) analysis, 222
- Competitive mediation, 869
- Competitive reaction effects, 528
- Complete mediation, 869
- Composite indicators, 552, 593
- Composite reliability, 564, 565, 578, 603
- Computer-assisted text analysis, *see* Text analysis
- Computer vision, 686
- Conditional direct effect, in mediation models, 881
- Conditional Granger causality test, 482
- Conditional independence assumption, 166
- Conditional indirect effect, in mediation models, 881
- Conditional moderated mediation, 888
- Conditional posterior distributions, 742–744
- Conditional process analysis, 885–887
 - conceptual description, 879
 - statistical description, 881
- Condition (experimental), 63
- Confidence interval, 268
- Configural equivalence, 139
- Configural properties, 379
- Confirmatory composite analysis, 601
- Confirmatory factor analysis (CFA), 575
- Confirmatory factor model, 555–556
- Confound checks, 13
- Confounding variables, 13, 41, 49, 859
- Confusion matrix, 164
- Congeneric factor model, 564
- Conjoint analysis (CA), 57, 574, 782, 949, 976, 1039
 - adaptive, 785
 - choice-based, 783, 976
 - traditional, 976
- Consistent Akaike information criterion (CAIC), 259
- Consistent estimate, 194
- Consistent PLS (PLSc) approach, 598
- Constant conditional variance, 489
- Constrained prior, 723
- Construct(s), 590
 - equivalence, 126
- Consumer images, 676
 - direct elicited images, 677–679
 - images from internet and social media, 676–677
- Consumer preferences, 971
- Consumption-specificity, 945, 947–948
- Contemporaneous effects, 527
- Content analysis, 635
 - automated, 645, 657, 658
 - computer-assisted, 635
 - modern, 635
 - traditional, 635, 637
- Context effect, 50
- Control function (CF) approach, 200, 207–209
- Control group, 41, 45–47, 51, 58, 61
 - nonequal, 829
 - suitable, 833
- Control variables, 13, 41, 49, 94, 859
- Convenience sampling, 23, 51
- Conventional lab experiment, 42
- Convergent validity, 564, 565, 567, 604, 606
- Convolutional neural networks, 682
- Cook's distance, 317
- Correction-based marker variable technique, 78
- Counterfactual, 832
- Covariance-based SEM, 588
- Covariates, 13, 41, 49, 94, 203, 859, 859
- Cox model, 1009
- Cronbach's alpha, 603
- Cross-cultural research, 130, 132
- Cross-level interaction, 380, 385, 390–392, 398, 400, 403
- Cross-level models, 380
- Crossover designs, 27
- Culti-units, 129
- Cultural dimensions, 130
- Customer base management
 - data and methodology, 1016–1018
 - estimation, 1018–1022

- Customer base management (*cont.*)
 parameter estimation, 1023–1025
 purchase prediction validity, 1025–1028
 results, 1023
- Customer lifetime value (CLV)
 continuous mixed models, 1012
 measurement models, 1015
 migration models, 1010–1012
 retention models, 1008–1009
 taxonomy, 1003–1006
- Customer loyalty
 antecedent of, 912
 conceptualization of, 914
 customer satisfaction and, 910
 definition, 911
 measurement, 921–925
 satisfaction and, 911
- Customer mindset, 1037
 associations, 1039
 blind tests, 1039
 conjoint analysis, 1039
 surveys, 1039
- Customer relationship management (CRM),
 947, 949, 954, 960
 data, 148, 170
- Customer retention-churn modeling, 1005
- Customer satisfaction
 conceptualization of, 910, 914
 CRM-Outcome Chain, 912
 cumulative perspective, 917
 and customer loyalty, 910
 definition, 911
 indices, 918
 level of, 918, 919
 loyalty, relationship of, 911
 and loyalty, 910
 measurement, 102, 910, 915
 multi-item scales, 96, 915
 novel and existing, 910
 ratings, 170
 requirements, 910
 scales, 915
 surveys, 102
 transaction-specific, 915
- Customer scoring, 160
- Customer targeting, 160
- Cut point model, 154, 166, 169
- D**
- Data augmentation, 151, 166, 740–741
- Data collection, 23, 32, 104
- Data equivalence, 124
- Data fusion, 152, 164, 167–169
 classic problem, 148–151, 167
 mixed levels of data aggregation, 151–152
 multivariate normal model, 153–163
 multivariate probit model, 163–165
- Data generating process (DGP), 194, 721
- Data visualizations, 493–494
- Debriefing, 31
- Deception, 31, 32
- Decomposition, 1057, 1061
- Deep neural networks, 682, 683, 686
- Degenerate prior, 723
- Degree centrality, 699
- Degree distribution, 703
- Degrees of freedom, 271
- Delta method, 864
- Demand curve, OLS, 185
- Demand effect, 27, 49, 75
 categories, 75
 conceptualization, 75–76
 definition, 75
 different data sources, 77
 proximal separation, 78
 psychological separation of scale
 formats, 78
 reasons for occurrence, 76–77
 statistical remedies, 78
 temporal separation of measurement, 77–78
- Dependent variable, 266, 300, 302, 304–306,
 310, 311, 314, 316, 319, 321,
 324, 326
- Design effect, 378
- Design matrix, 54
- Design of experiment (DOE), 55
- Detailing endogeneity, 190
- Deterministic models, 1007–1008, 1010
- Deviance, 388
- Dichotomous choice method, 971
- Dictionary-based methods, 636
- Difference in coefficients approach, 863
- Difference-in-differences estimator, 456,
 834–835
 application area for, 831
 approach, 829, 832
 critical assumptions, 832–834
 designs, 829, 832
 methodology, 830
 regression, 830
 requirements, 829
 staggered designs, 831
- Difference-in-means, 45
- Differential attrition, 48–49
- Diffuse prior, 723

- Digital experiment, 39
- Direct data fusion method, 165, 168
- Directed network, 699
- Direct effect, in mediation analysis, 860, 862, 863
- Direct surveys, 973
- Disacquiescence, 83–85, 91
- Disaggregated approach, 377
- Discontinuity
 - plausibility, 839
 - regression designs, 835
- Discrete mixture model, 167
- Discriminant analysis, 331, 951, 954
 - applied examples, 357–359
 - concept of, 332
 - discriminant function coefficients, 339–340
 - discriminant function estimation, 334–337
 - discriminant function performance, 337–339
 - model formulation, 334
 - objectives, 331–332
 - prediction, 341–343
 - problem and group definition, 333
- Discriminant validity, 565–567, 604
- Distance-based clustering methods, 227
- Distance concept, 341
- Distribution of the product approach, 865
- Domain-sampling model, 98
- Double-asymmetric structural VAR (DASVAR) model, 529
- Double randomized experiments, 892
- Dummy-coding, 293
- Dummy variables, 324
- Durbin-Watson statistic, 314
- Dutch auctions, 980
- Dyad, 697
- Dyadic data, 77
- Dynamic multiplier analysis, 535
- Dynamic panel, 1048
 - models, 457–461
- Dynamic systems, 471, 477, 481, 482, 485, 486, 491
- E**
- Effect-coding, 278, 293, 800
- Effect indicators, 552, 592
- Effect size, 294
- Eigenvector centrality, 701
- Elasticities, 307, 309, 312, 319, 355
- Emergent variables, 593
- Emic concepts, 125, 126, 135
- Empirical identification, 167
- Endogeneity, 301, 309, 321–324, 413, 567–569
 - advertising, 189
 - advertising elasticity, 187
 - correction, 204
 - description, 183–187
 - detailing, 190
 - Hausman test, 202
 - interaction term, 210–211
 - measurement error, 444, 463, 464
 - multiple endogenous regressors, 210
 - omitted variable bias, 441, 442, 449, 450, 456, 458, 459
 - price, 188
 - price elasticity, 187
 - problem on survey data, 191
 - regression model, 192–198
 - in survey research, 106–110
 - of treatment, 844
- Endogenous latent variable, 591
- Endogenous variables, 527–529
- Engle and Granger approach, 522
- English auctions, 979
- Error term, 184, 302
- Ethnocentrism, 140
- Etic concepts, 125, 126, 135
- Event study, 832
- Evolving series, 519
- Exact fit test, 600
- Excludability, 45, 46
- Exclusion restriction, 192, 196, 846, 847
- Exogeneity, 847
- Exogenous latent variable, 591
- Exogenous variables, 523, 526, 527
- Experimental bias, 982
- Experimental-causal-chain designs, 16
- Experimental design, 6, 47, 53, 793
 - dependent variable, operationalization of, 14–15
 - environmental setting (*see* Experimental environment)
 - experimental units and assignment to treatments (*see* Experimental units)
 - extraneous variables, 13–14
 - independent variable, 7–13, 29
 - mediators, 15–17, 30
 - preliminary testing, 27–28, 30
 - research question, definition of, 6–7, 29
- Experimental environment
 - field experiments, 19–20
 - laboratory experiments, 17–19
 - online experiments, 20–21
- Experimental factors, 266

- Experimental units, 21–22, 30
 experimental treatments, participants to, 25–27
 incentivization of participants, 24–25
 participants and sampling procedure, number of, 22–24
- Experimenter bias, 49
- Experiments, in market research
 and causality, 5
 ethical issues, 31–32
 experimental design (*see* Experimental design)
- Explanatory power, 595, 609
- Exploratory structural equation modeling (ESEM), 570
- External validity, 18, 44, 51, 85
- Extraneous variables, 13
- Extreme response style, 84, 91
- Eye-tracking, 15, 22
- F**
- Factor, 53, 54
- Factor-based SEM, 588
- Factorial Design, 793
- Factor weighting scheme, 596
- F-distribution, 271
- Feasible generalized least squares estimator, 436
- Feature extraction, 681–683
- Feedback effects, 523, 527
- f^2 effect size, 609
- Field experiments, 19, 40, 192, 215, 825–828
 average treatment effect, 44
 in business, 57–60
 causality and internal validity, 48
 classification, 825
 definition, 40–41
 experimental design and multivariate experiments, 53
 features of, 42–43
 generalizability of findings and external validity, 50
 vs. lab, 41–42
 natural, 826
 offline, 60
 online experiments, 43
 quasi, 826
 quasi-natural, 835
 sample size, 51
 types of, 825
- Financial impact, 1043, 1046, 1048–1049
- Finite mixture models
 applications in multivariate methods, 260–262
 description, 252
 example, 253–254
 and likelihood function, 254–256
 segmentation variable and mixed density function, 256–257
 segment determination, 258–260
- Firm images
 advertising databases, 680
 firm brand communications, 680
 images on social media pages of firms, 680
 product images on retail websites, 679–680
- First difference estimator, 455
- First-price auctions, 980
- First-stage regression, 193, 195, 207
- Fisher's classification functions, 341
- Fixed-effects estimator, 211, 432–435, 451, 454
- k -fold cross-validation, 610
- Forecast error variance decomposition (FEVD), 504–505, 519, 537
- Forecasting, 685
- Formative measurement model, 552, 592
 assessment, PLS-SEM, 606–608, 617–619
- Four-way interaction, 56
- Fractional factorial design, 8, 56
- Fraction of missing information, 163
- Framed field experiments, 19, 42
- F-test, 195, 273, 305–308
- Full factorial design, 8, 29
- Full information maximum likelihood (FIML) test, 483, 522
- Full model, 351
- Functional equivalence, 127
- Fusing media and purchase data, 148
- G**
- Gabor filter, 681
- Gabor Granger method, 974
- Gaussian copulas, 215
- Generalized autoregressive conditional heteroskedasticity (GARCH), 490, 491
- Generalized forecast error variation decomposition (GFEVD), 487–488, 537, 538
- Generalized impulse response functions (GIRFs), 485–487, 534

- Generalized method of moments (GMM) estimator, 460
- General linear model (GLM), 270
- Geo-experiments, 61
- Geographic proximity, 703
- Geweke and Meese's criterion (GM), 611
- Gibbs sampler, 155, 166, 168, 738–740
 blocked, 741–742
 Metropolis-Hastings (MH) algorithm, 746–748
 Monte-Carlo-Markov-Chain (MCMC), 748–749
- Gibbs sampler proposal density, 749–759
- GLOBE study, 131
- Goodness of fit, 304–305, 319, 600
- Grand mean, 270
 centering, 389, 403
- Granger causality tests, 481–482, 519, 535
- Group mean centering, 389, 390, 395, 403
- H**
- p*-Hacking, 607
- Halo effect, 1066
- Hamiltonian-Monte-Carlo (HMC), 759
- Hannan-Quinn (HQ) criterion, 485
- Hausman-Taylor approach, 450
- Hausman test, 200, 202, 203, 206, 208, 442, 444, 449
- Hawthorne effect, 49
- Hazard function, 1009
- Heterogeneity, 942, 943, 951–955, 1049–1051, 1068
- Heterogeneous customer group, 955
- Heteroskedasticity, 310, 315–317, 325, 326, 444, 454
- Heterotrait-monotrait ratio (HTMT), 604
- Hidden Markov models, 1011
- Hierarchical Bayes models, 730
- Hierarchical Bayes (HB) procedure, 811
- Hierarchical clustering, 253, 952, 953
 agglomerative algorithms, 228
 agglomerative clustering, 232
 application of, 240
 K-medoid clustering, 236–240
 linkage criteria in, 233
 outcome of, 233
 techniques for, 228
- Hierarchical data structure, 371, 376, 381, 403
- Hierarchical models, 745–746
- Hierarchical (or multilevel) random effect models, 573, 574
- Hierarchy of effects (HoE) model, 541
- Hit rate, 352, 799
- HLM, 396, 397
- Hofstede cultural dimensions, 131
- Holdout-sample validation, 204, 799
 sample, 610
- Homogeneity, 942
- Homologous models, 380
- Homophily, 695, 702
- Homoskedasticity, 309
- Hosmer-Lemeshow test, 353, 359
- Hotdeck, 150
- Hubs, 696, 699
- Hue, saturation, value (HSV), 681
- Human-coded features, 682
- Human development indicator (HDI), 130
- Hypothetical bias, 978, 979, 982, 994
- I**
- Ideal market segment, 944–945
- Ideal point model, 789
- Identifiability, 944–945, 958
- Identification strategy, 192
- Ignorability, 150, 167, 169
- Image processing, 668, 678, 679
- Image tagging, 668, 669, 687
 advertising, 671–672
 branding, 672–674
 consumer perspectives, 675
 consumer vs. firm images, 675–681
 feature extraction, 681–683
 future research in marketing, 669
 model application, 685
 model evaluation and validation, 684–685
 model training, 683–684
 online shopping experience, 674–675
 product design, 670–671
- Immediate effect, 486, 507
- Imposed-etic approach, 125
- Impulse response function (IRF), 478, 485, 486, 505–509, 519, 530
- Incentive(s), 89
 alignment, 798
- Incentivization, 24
- Independent variable, 266, 300–302, 304–307, 309–314, 316, 317, 319, 321, 324, 326
 coding scheme, 897–898
 fractional factorial designs, 8–9
 multi-factor designs, 8
 number of levels, 9
 operationalization of treatments, 10–13
 single-factor designs, 8

- Index of moderated mediation, 882, 886
- Indicators, 591
- Indirect effect, in mediation analysis, 860, 862
- Indirect surveys, 976
- Individual-item reliability, 564
- Individual-level constructs, 378
- Inference-by-eye, 269, 276
- Informative prior, 723
- Informed consent, 31
- In-sample model fit criteria, 204
- In-sample predictive power, 609
- Instructional manipulation checks, 11
- Instrumental variables (IVs), 183, 192, 195, 568, 1047
 - applications in business research, 850
 - area of application, 845
 - in field experiment, 849
 - graphical illustration of, 845
 - method, 853
- Interaction(s), 8, 9, 20, 28, 29
 - effect, 49, 50, 53, 319, 324, 356, 790
 - term, 210
- Internal consistency reliability, 603
- Internal validity, 17, 44, 47, 86
- International market research
 - challenges, 123
 - conceptual framework, 124
 - data analysis, 138
 - data collection, 133
 - interpretation, 140
 - units and drivers, 127
- Interpretable features, 668, 679, 682, 687
- Intervention, 825, 826, 830, 831
 - analysis, 478
- Intraclass correlation coefficient (ICC), 377, 395, 436
 - ICC1, 379
 - ICC2, 379
- Invariance of parameters across groups, 572
- Inverse square root method, 599
- Item parcels, 570
- Item reversals, 100
- Iterative translation, 135, 136
- J**
- Jaccard similarity coefficient, 226
- Jackknife method, 356
- JAGS, 166
- John Henry effect, 49
- Joint data fusion method, 167
- Joint endogeneity, 480
- K**
- K-centroids cluster analysis, 240–244
- Key informant bias, 79
 - conceptualization, 79
 - procedural remedies, 80
 - reasons for occurrence, 79
 - statistical remedies, 81
- K-means clustering, 639, 642
- K-medoid clustering, 236–240
- KPSS tests, 520
- L**
- Laboratory experiment, 17, 41, 48, 63
- Lagrange multiplier test, 439, 562
- Latent class analysis (LCA), 809, 954
- Latent constructs, 292
- Latent curve models, 573
- Latent Dirichlet allocation (LDA), 639, 642
- LatentGold[®], 261
- Latent growth models, 380
- Latent instrumental variables (LIV), 215
- Latent semantic analysis (LSA), 639, 642
- Latent variable model, 344, 551, 556, 564, 567
 - empirical example, 579–581
 - extensions of, 571
- Latent variables, 154, 163, 590
- Least squares dummy variables (LSDV) regression, 433
- Least squares method, 302, 307, 311, 315
- Levene's test, 295
- Likelihood-ratio test, 351, 353, 354, 388
- Linear mixed models (LMMs), 294
- Linear regression model, 300, 301, 304, 307, 312, 317, 319, 321, 322, 324, 326, 722
 - aim of, 302
 - benchmark, 610
 - goal of, 317
 - goodness of fit for, 304
 - results of, 308, 318
- LKJ prior, 157
- Logistic probability unit, 345
- Logistic regression
 - applications, 343
 - applied examples, 359–364
 - coefficients, 353–356
 - function estimation, 349
 - latent variable model, 344–345
 - model formulation, 348
 - model performance, 349–353
 - prediction, 356
 - probability model, 345–347
 - problem and group definition, 348

- Log-likelihood, 1025
 Long format, 286, 287
 Longitudinal data, 374, 375, 380
 Longitudinal mediation analysis, 890
 Long-run cointegrating equilibrium, 523
 Long-term effect, 486, 504, 507
- M**
- Machine learning, 729
 Mahalanobis distance, 317
 Main effect, 56
 Manipulation, 10, 13, 14, 16–18, 23, 27–29
 checks, 11–13, 29
 effectiveness of, 30
 Marginal effects, 354
 Marketing and mindset metrics models, 541
 Market response modeling, 148
 Market scenarios, 783
 Market segmentation, 222, 783, 941
 accessibility, 957–958
 a priori idea, 955
 conjoint analysis, 949
 consumption-specificity, 947–948
 creation of market segments, 952–953
 CRM databases, 949
 customers as unique entities, 940
 data-driven, 951
 digitalization, 949
 formation of segments, 951–957
 goals of, 940
 heterogeneity and homogeneity, 942
 ideal market segment, 944–945
 identifiability, 944–945
 implementation, 960–962
 observability, 946
 profiling of market segments, 953–954
 responsiveness, 958
 segment-of-one, 942–943
 sizing of market segments, 954–957
 stability, 958–959
 standardized, 940, 942, 950
 substantiality, 959
 surveys, 949
 Market simulations, 806
 Markov chain Monte Carlo (MCMC), 151, 153,
 155, 159, 160, 166
 Markov processes, 1011
 MATLAB, 157, 166
 Maximum likelihood estimation (MLE)
 approach, 254, 386, 462, 802, 1018
 Mean absolute error (MAE), 610
 Mean centering, in conditional process analysis,
 896–897
 Measured latent factor, 82
 Measured response style technique, 85
 Measurement equivalence, 139
 Measurement errors, survey research, 186
 CMB, 75–78
 key informant bias, 78–81
 response styles, 83–85
 social desirability, 81–83
 Measurement models, 564–567
 confirmatory factor analysis, 575
 extensions, 569–571
 observed variables, 563
 types of, 552
 Measurement-of-mediation design, 16
 Measurement theory, 72–74, 592–594
 Mechanical Turk (MTurk), 23, 91, 93
 Mediated moderation, 880
 Mediation analysis, 769–771, 859
 causal inference, 891–894
 classification, 869–870
 complete mediation, 869
 conditional direct effect, 881
 conditional indirect effect, 881
 conditional process analysis, 885
 direct effect 860, 862, 863
 effect size measures, 870
 indirect effect, 860, 862
 longitudinal and multilevel, 890
 regression analysis vs. structural equation
 modeling, 898–899
 sample size and power, 895–896
 software tools, 900
 time and nested data in, 890–891
 variable metrics, 870–871
 Mediation models, with multiple predictor/
 outcome variables, 889–890
 Mediator, 15, 859
 Meta-analysis, 375
 Method effects, 570, 571
 Metric equivalence models, 139
 Metropolis-Hastings (MH) algorithm, 746–748
 Midpoint responding, 84
 Missing by design, 151, 165, 167
 Missing completely at random (MCAR), 150
 Missing data, 149–150, 152–155, 158–159,
 161, 163, 165, 169–170
 mechanism, 167, 169
 Missing regressors, 169
 Mixed aggregate-disaggregate data, 152, 167,
 168
 Mixed-ANOVA, 289

- MLwiN, 397
- Model chi-square, 351
- Model comparisons, 611
- Model estimation, 385, 390, 556–558
- Model fit, 387, 388, 394, 399–401
- Model identification, 554
- Model modification, 562
- Model overfit, 609
- Model training, 683–684
- Moderated mediation, 879, 880
- Moderated moderated mediation, 888
- Moderation-of-process designs, 16
- Moderator, 325, 859
- Modification index (MI), 562, 572, 577
- Modified Akaike information criterion, 259
- Monadic approach, 974, 995
- Monte-Carlo-Markov-Chain (MCMC)
 techniques, 748–749
 binomial probit without data augmentation,
 772–778
 convergence, 754
 graphical inspection, 759
 iterations, 766
 traces, 753
- Monte Carlo simulation approach, 534
- Moving average (MA), 472, 497, 498
- Mplus, 397, 399
- MTurk, *see* Mechanical Turk (MTurk)
- Multicollinearity, 310–313, 326
- Multi-factor designs, 8
- Multilevel mediation analysis, 890
- Multilevel modeling, 398–400
 aggregation, 379
 assumptions, 386
 baseline model, 382–383
 conceptual relevance, 372–375
 configural properties, 379
 cross-level interaction effects, 385
 cross-level models, 380
 global properties, 379
 homologous models, 380
 independent variables, at level 1, 383
 independent variables, at level 2, 383–384
 individual-level constructs, 378
 intraclass correlation, 379
 latent growth models, 380
 longitudinal data, 380
 model estimation and assessing model fit,
 386–389
 multilevel structural equation modeling,
 392–396
 random slopes, testing for, 384–385
 sample size considerations, 390–392
 shared properties, 379
 single-level models, 379
 software, 396–397
 statistical relevance, 375–378
 unobserved heterogeneity, 379
 variable centering, 389–390
- Multilevel structure, 372
 equation modeling, 392–396
 in marketing organizations, 374
- Multinomial logit (MNL) model, 791
- Multinomial regression model, 209
- Multiple endogenous regressors, 209–210
- Multiple imputation, 167
- Multiple linear regression (MLR), 495,
 500, 502
- Multiple regression model, 500
- Multiple time series models, 479–481
 cointegration test, 483
 GFEVD, 487–488
 GIRFs, 485–487
 Granger causality tests, 481–482
 VAR model, 483–485
 VEC model, 484
 volatility models, 489–491
- Multiplicative sales response function, 318, 326
- Multi-sample models, 573
- Multi-stage random sampling, 23
- Multivariate analysis of variance (MANOVA),
 267, 293
- Multivariate experiment, 53–57
- N**
- Natural experiment, 41
- Natural field experiments, 19, 42
- Nested data structures, 371–373, 375–377, 379,
 383, 389
- Net acquiescence, 83
- Network analysis, 694
- Network density, 703
- Nonconvergence, 558
- Nonexperimental approaches, 192
- Non-hierarchical clustering, 228, 952
- Non-ignorability, 151, 169
- Non-interference, 45, 47
- Non-overlapping clustering approaches, 228
- Non-parametric model, 166
- Non-probability sampling, 104
- Non-response bias, 85, 87
 conceptualization, 87–88
 organizational factors, 88
 personal factors, 88
 procedural remedies, 88–89

- statistical remedies, 89
 - survey-related factors, 88
- Non-sampling bias, 85
 - conceptualization, 85–86
 - procedural remedies, 87
 - reasons for occurrence, 86
 - statistical remedies, 87
- Normal distribution of the residuals, 309, 311
- Normal-Inverse Wishart prior, 746
- Normal prior, 157
- Normal regression likelihood, 734
- No U-turn Sampler (NUTS), 759, 761

- O**
- Object detection, 677, 686, 688
- Observational data, 94, 183
- Observational investigations, 40, 41
- Omitted variables, 107, 196, 203
- Omnibus test, 276, 293
- Online collages, 668, 672
- Online experiment, 20, 43–44, 47, 51
- Online social networks, 950
- Open-ended questions, 98
- Operational segmentation, 942, 943, 952, 962
- Opinion leaders, 696
- Optimal design, 56
- Optimal number of clusters, 953
- Ordered-categorical (discrete-ordinal) observed variables, 569
- Ordinary least squares (OLS) approach, 184, 484
- Orthogonality, 56
- Out-degree, 699
- Outlier detection, 317
- Outliers, 310, 317–318, 321, 325, 326
- Out-of-sample predictive power, 610
- Over-identification test, 200

- P**
- Pairwise Granger causality test, 482
- Panel data, 211–213, 380, 1056, 1059, 1060, 1062, 1065, 1070
- Panel-internal instrumental variables, 450, 459
- Panel vector autoregression (PVAR), 484, 541
- Parallel multiple mediator model, 871–875
- Parallel trends assumption, 832, 833
- Parametric probabilistic models, 1008
- Pareto/NBD model, 1012–1014
 - customer base management (*see* Customer base management)
 - regression coefficients, 1023
- Partial autocorrelation function (PACF), 475–477, 495, 497, 498
- Partial least squares (PLS) estimation, 557
- Partial least squares structural equation modeling (PLS-SEM)
 - algorithm, 595–598
 - bias, 594
 - considerations using, 598–601
 - corporate reputation model, 612–614
 - distributional assumptions, 598
 - formative measurement model assessment, 606–608
 - goodness-of-fit, 600–601
 - measurement theory, 592–594
 - methodological reason for, 589
 - model complexity and sample size, 599–600
 - model estimation, 615
 - procedure for evaluation, 601–602
 - reflective measurement model assessment, 603–605
 - research application, 612–621
 - results evaluation, 616–621
 - statistical power, 599
 - structural model assessment, 608–612
 - structural theory, 591–592
- Partial mediation, 869
- Partial moderated mediation, 888
- Partworth model, 789
- Partworth utilities, 784
- Path, 700
 - coefficients, 592
 - weighting scheme, 596
- Path model
 - estimation with PLS-SEM, 594–601
 - with latent variables, 590–591
- Payment card method, 973, 975
- Percentile method, 605
- Persistence modeling, 481, 527
- Pilot tests, 28
- Placebo effect, 49
- PLSpredict procedure, 610
- Poisson regression model, 209
- Polar extreme approach, 348
- Policy simulation analysis, 481
- Pooled OLS (POLS) estimator, 427, 430, 452
- Population heterogeneity, 572–574
- Posterior, 151, 153, 161, 163, 164
- Post hoc tests, 276
- Power, 52, 286, 558, 567
- Prediction, 205
 - accuracy, 684, 687
 - error, 610
 - power, 610
- Preference measurement, 782
- Preliminary testing, 27

- Pretesting, 28
 Pre-trained models, 683, 685
 Price elasticity, 182, 187
 Price endogeneity, 188
 Price experiment, 981
 Price promotions, 1057, 1058, 1062, 1066
 Price sensitivity, 959, 963, 973, 988, 989
 Prior, 157, 166
 Probabilistic latent semantic analysis (PLSA), 639
 Probabilistic models
 parametric, 1008
 semi-parametric, 1009
 Probability concept, 342
 Probability density function, 273
 Probability model, 345
 Probability samples, 104
 Probit model, 209
 Product attribute, 976, 977
 Product of coefficients approach, 863
 Profiling of market segments, 952–954
 Prolific Academic (ProA), 23, 92, 93
 Promotion effectiveness, 1056, 1057, 1059–1062, 1067, 1070
 PROMOTIONSCAN model, 1065–1067
 Proportional chance criterion (PCC), 352, 353, 358, 359
 Proportional hazard model, 1009
 Proportional random sampling, 23
 Proximity measures, 225
 Pseudo-etic, 125
 PSPP, 325
 Psychology of survey response, 71–72
 Pulse effect, 478
 P-value, 273
 Python, 157, 166
- Q**
- Quantification, 952
 Questionnaire design, survey research process
 pre-test of questionnaire, 103–104
 question content, 95–98
 question format, 98–100
 question sequence, 101–103
 question wording, 100–101
 survey content, 94–95
 survey layout, 103
 Question wording
 neutrality, 101
 simplicity, 100
 unambiguousness, 101
 Quota sampling, 24
- R**
- R^2 , 304, 305, 307, 308, 311, 318, 319, 388
 Random assignment, 40, 41, 43, 45, 46, 58
 of prices, 191
 Random cluster-specific slope
 coefficients, 461
 Random coefficient modeling, 371
 Random effects, 213
 estimator, 432, 436, 438, 452
 Random intercept, 382–385, 387, 388, 403
 Randomization, 5, 13, 19, 30, 40, 42, 44, 47, 60
 Randomized experiment, 61
 Randomized response techniques, 82
 Random slope models, 461, 462
 Random slopes, 384–385, 389, 394, 400, 403
 Random utility theory, 787
 Receiver operator characteristic (ROC), 352
 Recursive model, 553
 Reduced-form VAR, 483, 486, 519, 527
 Redundancy analysis, 606, 618
 Reflective indicators, 592
 Reflective measurement model, 552, 592
 assessment, PLS-SEM, 603–605, 616–621
 Regression analysis, 56, 192, 270, 726
 autocorrelation, test for, 313
 dependent and independent variables, 310
 efficiency of estimators, 311
 endogeneity, 309, 321
 goodness of fit, 304
 heteroskedasticity, test for, 315
 implications, 320–321
 interpretation of results, 307
 multicollinearity, test for, 311
 objective function and estimation of
 regression coefficients, 301–303
 outliers, identification of, 317
 problem statement, 300–301
 residuals, 309
 results, 307–309
 significance testing, 305
 software, 325
 standardization of coefficients, 306
 transformation of variables, 318
 Regression discontinuity
 approach, 838, 839
 bandwidth selection, 839
 designs, 835, 838
 method, 841
 setting, 835
 Regression weights, 597
 Regressor-error dependencies, 187
 Relevance condition, 846

- Reliability, 72, 73, 75, 79–81, 96–99, 104–106, 557, 564, 565, 567, 570
 coefficient, 603
- Relief from royalty, 1042
- Repeated measures, 375
- Repeated-measures analysis of variance (RM-ANOVA), 267, 286
- Representation errors, survey research
 non-response bias, 87–89
 non-sampling bias, 85–87
- Representativeness, 85
- Residual, 302, 303, 305, 309, 311, 313–317, 321, 322
 analysis, 562
- Response range, 84
- Response rates, 87
- Response styles, 83
 acquiescence, disacquiescence and net acquiescence, 83–84
 extreme responding, midpoint responding and response range, 84
 procedural remedies, 84–85
 statistical remedies, 85
- Responsiveness, 958
- Return on media (ROM), 1074–1076
 advertising copy and creative effects, 1086
 advertising response dynamics, 1080–1082
 advertising-response function, shape of, 1078–1080
 asymmetric response, 1082–1083
 brand equity, 1091
 building distribution, 1093
 customer equity, 1091–1093
 data-interval bias, 1082
 enhancing sales-call effectiveness, 1093
 estimated response parameters, 1087–1088
 fundamental advertising response phenomena, 1076–1078
 intermediate performance metrics, 1086–1087
 lifting stock price, 1093
 path-to-purchase and attribution models for digital media, 1088–1090
 protecting/enhancing price premiums, 1093
 reverse causality, 1083–1084
- Revealed preference methods, 972, 979, 994
 direct, 979
 experimental, 981
- Reverse pricing, 980
- Root mean squared error (RMSE), 610
 of approximation, 560
- Root-mean-square deviation (RMSE), 501
- Root mean square residual covariance (RMStheta), 600
- R-package bayesm, 721, 728, 745, 750
- RStan, 155
- R statistical language, 157, 166
- S**
- Sales contests (SCs), 357
- Sales promotions
 academic research on, 1059, 1061
 analysis of, 1060
 characteristics, 1056
 effectiveness, 1059–1062
 investments in, 1056
 manufacturer spendings on, 1056
 marketing budget on, 1056
 systematic information on, 1059
 tools, 1056–1058
- Sales response function, 318, 320, 326
- Sales response models, 1094
- Sample size, 22, 42, 51–53
 considerations, 390–392
- Sampling frame equivalence, 136
- Sampling frames, 86
- Sampling methods, 151
- Sargan-test, 201
- SAS, 325, 397
- Satisficing, 11
- Sawtooth[®], 261
- SCAN*PRO model, 1063, 1065
- Scanner panel, 1059, 1070
 data, 1041
- Schwartz' Bayes Information Criterion (SBIC), 259, 485, 499, 540, 611, 1025
- Schwartz cultural dimensions, 131
- Second-price auctions, 980
- Segmentation criteria, 945–948
- Segmentation, targeting, positioning (STP) framework, 941
- Segment-of-one, 942–943
- Segment size, 963
- Selection bias, 46, 48
- Self-reports, 14, 29, 71
- Semi-parametric probabilistic models, 1009–1010
- Sensitivity analysis, 355
- Sentiment analysis, 640
- Sequential approach, 990, 994
- Serial correlation, 426, 428, 430, 436, 437, 444, 454
- Serial multiple mediator model, 875–878

- Significance testing, 52, 53, 305–306
- Simple random sampling, 23, 51
- Simplicity, 100
- Simultaneity, 187, 714, 1045
- Single-factor designs, 8
- Single items, 96
- Single-level models, 379
- Single mediator model
 - assumptions, 867–869
 - conceptual description, 879
 - description of, 881
 - statistical description, 881–885
- Single-source panels, 1060, 1063, 1067, 1069, 1070
- Singular value decomposition (SVD), 639
- Sizing of market segments, 952, 954–957
- SMARTPLS, 261
- Snowball sampling, 705
- Sobel test, 864
- Social desirability, 81
 - conceptualization, 81
 - procedural remedies, 81–82
 - reasons for occurrence, 81
 - statistical remedies, 82–83
- Social influence, 696
- Social learning, 696
- Social network, 695
- Social normative pressure, 696
- Software, 396–397
- Specific indirect effect, multiple mediator model, 878
- Sphericity, 288
- Split-plot ANOVA, 289
- Split questionnaire, 149, 153, 163, 166–168
- SPSS, 307, 325, 326, 397, 825
- Stability, 947, 958–959
- Stable Unit Treatment Value Assumption (SUTVA), 47
- Stan, 155–157, 163–166, 721, 759, 775–776
- Standard error of mean, 268
- Standard errors, 202, 208
- Standardized regression coefficients, 306–307
- Standardized root mean square residual (SRMR), 600
- STATA, 307, 325, 326, 397, 825, 849
- Stated preference method, 972
 - direct, 973
 - experimental, 976
 - hypothetical bias, 978
- Stated preference methods, 973, 978
- Stationary, 473, 475, 478, 481, 483, 495–497, 519
- Statistical power, 391
- Step effect, 478
- Stimuli, 795
- Strategic answering, 976
- Strategic segmentation, 942, 952, 962
- Stratified sample, 51
- Structural breaks, 521
- Structural equation modeling (SEM), 392–396, 550
 - empirical example, 574–581
 - extensions of core structure equation model, 569–574
 - latent variable model, 567
 - measurement model, 564–567
 - problem of endogeneity, 567–569
 - submodels, 551–556
- Structural model, 169
- Structural model assessment, PLS-SEM, 608–612, 619–621
- Structural theory, 591–592
- Structural vector-autoregressive model (SVAR), 519, 529
- Student samples, 104
- Subcultures, 129
- Subgraph sampling, 706, 707
- Substantiality, 959
- Sum of squares, 271
- Supervised classification task, 683
- Supervised learning, 684
- Support vector machine (SVM), 638
- Suppressor variables, 13
- Survey(s), 949
 - design, 69, 88, 110
 - non-response, 151, 170
 - research, 149–151, 163, 170
 - subsampling, 150, 167
- Survey bias
 - fundamentals of survey research, 70–71
 - measurement theory, 72–74
 - psychology of survey response, 71–72
- Survey research process
 - data analysis, 106
 - data collection, 104–105
 - measurement evaluation, 105–106
 - questionnaire design, 93–104
 - selection of research variables, 89–90
 - selection of survey method, 91–93
- Synthetic controls, 833
- Systematic biases, 74
- Systematic errors, in survey research
 - measurement errors, 74–85
 - representation errors, 85–89
- Systematic random sampling, 23

T

- Target variable, 686, 687
- T-distribution, 270
- Technology acceptance model (TAM), 554, 574, 580
- Temporal causality, 536
- Testing the global fit of model, 558–561
- Text analysis
 - approaches to, 635–636
 - classification methods, 638
 - construct, 647–648
 - data collection, 646–647
 - dictionary-based methods, 636–638
 - history, 635
 - interpretation and analysis, 650–657
 - operationalization, 648–650
 - organization and firm environment, measurement of, 642–643
 - research question, 645–646
 - sentiment analysis, 640–641
 - textual data, 643–644
 - topic discovery and positioning maps, online text, 642
 - topic modeling, 638–639
 - validation, 657–659
 - word of mouth communication, 641–642
- Three-way interaction, 56
- Ties, 697
- Tie strength, 701
- Time effects, 49
- Time fixed effects, 213
- Time sampling, 150, 167
- Time-series analysis, in marketing
 - multiple time series models, 479–491
 - univariate time series treatments and diagnostics, 471–479
- Time series processor (TSP), 539
- Time trend, 831
- Tolerance, 311
- Topic modeling, 638
- Total effect, in mediation analysis, 860, 862
- Total indirect effect
 - multiple mediator model, 878
 - parallel multiple mediator model, 872, 873
 - serial multiple mediator model, 876, 877
- Transfer function, 478, 479
- Transformation of variables, 310, 318–320
- Treatment effect, 830
 - local average treatment effect (ATE), 44–47, 829, 848
- Treatment group, 40, 41, 45, 46, 825, 826, 828–831
- T-test, 270

- Two-level regression model, 381–392
- Two-stage least squares (2SLS) approach, 194, 199, 202
- Two-way fixed effects model, 212
- Two-way interaction, 56

U

- U-method, 356
- Unambiguousness, 101
- Undirected network, 699
- Unit root test, 473, 481, 483, 494, 502, 519
- Univariate time series models
 - autoregressive and moving average process, 471–473
 - autoregressive integrated moving average model, 474–476
 - evolution vs. stationarity, 472–474
 - single equation time-series models, exogenous variables, 476–479
- Unobserved demand shocks, 189
- Unobserved effects models, 211
- Unobserved heterogeneity, 379
- Unobserved population heterogeneity, 574
- Unsupervised learning, 684, 687
- Utility, 782
 - function, 783, 977
 - model, 787

V

- Validity, 69, 72–75, 81, 85, 86, 95, 96, 104–106, 785
- Value comprehension, 988
- van Westendorp method, 973
- Variable centering, 389–390
- Variance inflation factor (VIF) value, 311, 312, 347, 606
- Vector autoregressive (VAR) models, 483, 484, 502–504, 518
 - dynamic system of equations, 523
 - investor response models in marketing-finance, 539
 - in levels, 484
 - marketing studies, 524
 - order of lags, 484–485
 - software programs for estimation, 539
 - VAR-in-difference model, 483, 484
- Vector-error correction model (VECM), 481, 484, 523, 530
- Vector model, 788
- Vector moving average (VMA), 486
- Vickrey auctions, 980

Videos, 668
Visual data, 668
Visual features, 672
Volatility, 471, 489–491
Volunteer sampling, 23

W
Website test, 57
Weighted PLS-SEM approach, 598
Weight matrix, 698

Wide-format, 286, 287
Wilks's lambda, 338
Willingness-to-pay (WTP), 798, 970, 971, 973,
976–982
 definition, 971
 drivers of, 982–989
 measurement methods, 972–986
WinBUGS, 155, 166, 765
Within-estimator, 212
Within-subjects, 25–27, 266
Within variance, 376, 377, 382, 384