MULTIVARIATE APPLICATIONS SERIES

Longitudinal Structural Equation Modeling

A COMPREHENSIVE INTRODUCTION

Jason T. Newsom

ROUTLEDGE

"The breadth of topics covered is perfect for my course. ... Connections are drawn between types of models and their similarities/differences ... [and] ... students often ask questions related to "when to use what model." ... I am also excited about inclusion of continuous and discrete variables. It is difficult to find texts that include both. ... I would absolutely ... recommend it to colleagues and students. I also would consider using this text as a primary textbook for my courses."

– Natalie D. Eggum, Arizona State University, USA

"Newsom is a particularly gifted writer. ... He explains complex material clearly, without over-simplifying it ...The book will be very popular among applied scientists. ... [and it is] ... appropriate for graduate courses in either SEM or longitudinal modeling. … I would strongly consider adopting this book ... as the primary textbook."

– David L. Roth, Johns Hopkins University, USA

"I could see this text being ... required for ... SEM ... [in] ... all of the social sciences. ... While there are other ... texts on longitudinal analysis, they do not emphasize the use of SEM. ... Strengths: ... addresses confusion that [readers] have [when] deciding which analytic technique to choose ... [and includes] analysis on both continuous and discrete variables. ...The author is a good writer, who can readily explain these advanced statistical topics."

– Brian A. Lawton, George Mason University, USA

"Many of the current texts on longitudinal data analysis only tangentially mention SEM or do not cover the full spectrum of approaches. ...This book is really needed in the social science field. ... I like the organization ... Newsom starts with the basic issues in longitudinal SEM and moves to more advanced topics through the volume. I also like the way the chapter was formatted with respect to defining terms/concepts and then providing an example with data for each concept."

– Kristin D. Mickelson, Kent State University, USA

This page intentionally left blank

Longitudinal Structural Equation Modeling

This comprehensive resource reviews structural equation modeling (SEM) strategies for longitudinal data to help readers see which modeling options are available for which hypotheses. The author demonstrates how SEM is related to other longitudinal data techniques throughout. By exploring connections between models, readers gain a better understanding of when to choose one analysis over another. This book explores a range of models, from basic to sophisticated, including the statistical and conceptual underpinnings that are the building blocks of the analyses. The book is accessibly written, and research examples from the behavioral and social sciences and results interpretations are provided throughout. The emphasis is on concepts and practical guidance for applied research rather than on mathematical proofs. New terms are highlighted and defined in the Glossary. Figures are included for every model along with detailed discussions of model specification and implementation issues. Each chapter also includes examples of each model type, comment sections that provide practical guidance, model extensions, and recommended readings.

Highlights include:

- Covers the major SEM approaches to longitudinal analysis in one resource.
- Explores connections between longitudinal SEM models to enhance integration.
- Gives numerous examples that help readers match research questions to appropriate analyses and interpret results.
- Reviews practical issues related to model specification and estimation to reinforce connections.
- Analyzes continuous and discrete (binary and ordinal) variables throughout for breadth not found in other sources.
- Reviews key SEM concepts for those who need a refresher (Ch. 1).
- Emphasizes how to apply and interpret each model through realistic data examples.
- Provides the book's data sets at www.longitudinalsem.com along with the Mplus and R-lavaan syntax used to generate the results.
- Introduces the LISREL notation system used throughout (Appendix A).

Ideal for graduate courses on longitudinal (data) analysis, advanced SEM, longitudinal SEM, and/or advanced data (quantitative) analysis taught in the behavioral, social, and health sciences, this text also appeals to researchers in these fields. Intended for those without an extensive math background, prerequisites include familiarity with basic SEM. Matrix algebra is avoided in all but a few places.

Jason T. Newsom is professor of community health at the Institute on Aging at Portland State University, USA.

Multivariate Applications Series

Sponsored by the Society of Multivariate Experimental Psychology, the goal of this series is to apply statistical methods to significant social or behavioral issues, in such a way so as to be accessible to a nontechnical-oriented readership (e.g., non-methodological researchers, teachers, students, government personnel, practitioners, and other professionals). Applications from a variety of disciplines such as psychology, public health, sociology, education, and business are welcome. Books can be single- or multiple-authored or edited volumes that (1) demonstrate the application of a variety of multivariate methods to a single, major area of research; (2) describe a multivariate procedure or framework that could be applied to a number of research areas; or (3) present a variety of perspectives on a topic of interest to applied multivariate researchers.

Anyone wishing to submit a book proposal should send the following: (1) author/title; (2) timeline including completion date; (3) brief overview of the book's focus, including table of contents and, ideally, a sample chapter (or chapters); (4) a brief description of competing publications; and (5) targeted audiences.

For more information, please contact the series editor, Lisa Harlow, at Department of Psychology, University of Rhode Island, 10 Chafee Road, Suite 8, Kingston, RI 02881-0808; phone (401) 874-4242; fax (401) 874-5562; or e-mail LHarlow@uri.edu.

What If There Were No Significance Tests? co-edited by Lisa L. Harlow, Stanley A. Mulaik, and James H. Steiger (1997)

Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming, written by Barbara M. Byrne (1998)

Multivariate Applications in Substance Use Research: New Methods for New Questions, co-edited by Jennifer S. Rose, Laurie Chassin, Clark C. Presson, and Steven J. Sherman (2000)

Item Response Theory for Psychologists, co-authored by Susan E. Embretson and Steven P. Reise (2000)

Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, written by Barbara M. Byrne (2001)

Conducting Meta-Analysis Using SAS, written by Winfred Arthur, Jr., Winston Bennett, Jr., and Allen I. Huffcutt (2001)

Modeling Intraindividual Variability with Repeated Measures Data: Methods and Applications, co-edited by D. S. Moskowitz and Scott L. Hershberger (2002)

Multilevel Modeling: Methodological Advances, Issues, and Applications, co-edited by Steven P. Reise and Naihua Duan (2003)

The Essence of Multivariate Thinking: Basic Themes and Methods, written by Lisa Harlow (2005)

Contemporary Psychometrics: A Festschrift for Roderick P. McDonald, co-edited by Albert Maydeu-Olivares and John J. McArdle (2005)

Structural Equation Modeling with EQS: Basic Concepts, Applications, and Programming, Second Edition, written by Barbara M. Byrne (2006)

A Paul Meehl Reader: Essays on the Practice of Scientific Psychology, co-edited by Niels G. Waller, Leslie J. Yonce, William M. Grove, David Faust, and Mark F. Lenzenweger (2006)

Introduction to Statistical Mediation Analysis, written by David P. MacKinnon (2008)

Applied Data Analytic Techniques for Turning Points Research, edited by Patricia Cohen (2008)

Cognitive Assessment: An Introduction to the Rule Space Method, written by Kikumi K. Tatsuoka (2009)

Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Second Edition, written by Barbara M. Byrne (2010)

Handbook of Ethics in Quantitative Methodology, co-edited by Abigail T. Panter and Sonya K. Sterba (2011)

Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Sciences, co-edited by Jason T. Newsom, Richard N. Jones, and Scott M. Hofer (2012)

Structural Equation Modeling with MPlus: Basic Concepts, Applications, and Programming written by Barbara M. Byrne (2012)

Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis, written by Geoff Cumming (2012)

Frontiers of Test Validity Theory: Measurement, Causation and Meaning, written by Keith A. Markus and Denny Borsboom (2013)

The Essence of Multivariate Thinking: Basic Themes and Methods, Second Edition, written by Lisa L. Harlow (2014)

Longitudinal Analysis: Modeling Within-Person Fluctuation and Change, written by Lesa Hoffman (2015)

Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment, co-edited by Steven P. Reise & Dennis Revicki (2015)

Longitudinal Structural Equation Modeling: A Comprehensive Introduction, written by Jason T. Newsom (2015)

This page intentionally left blank

Longitudinal Structural Equation Modeling

A Comprehensive Introduction

Jason T. Newsom

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2015
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
27 Church Road, Hove, East Sussex BN3 2FA

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging in Publication data
Newsom, Jason T.
Longitudinal structural equation modeling : a comprehensive introduction / Jason T. Newsom.
pages cm. – (Multivariate applications series)
1. Longitudinal method–Mathematical models. 2. Social sciences–Statistical methods. 3. Structural equation modeling. I. Title.
H61.25.N49 2015
519.5′3–dc23
2014046412

ISBN: 978-1-84872-696-3 (hbk)
ISBN: 978-1-84872-697-0 (pbk)
ISBN: 978-1-315-87131-8 (ebk)

Typeset in Sabon
by Out of House Publishing

In memory of my mother, Autumn Newsom, who instilled a sense of wonder, and my great aunt, Florence Glenn, and my grandparents, Lee and Dora Newsom, without whose financial assistance I would have never gone to college.

This page intentionally left blank

Contents

Figures

Tables

This page intentionally left blank

Preface

This book is a thorough examination of structural equation modeling (SEM) strategies for longitudinal data. SEM is an analysis approach that combines path analysis and confirmatory factor analysis, and its ability to model complex hypotheses and account for measurement error creates a truly remarkable, flexible method for analyzing data collected over time. The book is intended for advanced students or researchers in behavioral and social sciences and other related fields. Throughout the book readers will find examples relevant for students and researchers in psychology, gerontology, public health, sociology, education, social work, and economics. It is appropriate as a text for a second course in SEM or as a reference for researchers conducting longitudinal research. Familiarity is assumed with basic concepts of SEM, such as path analysis confirmatory factor analysis, model identification and fit, and general structural models.

Reasons for Writing this Book

At the time this book was conceived there were no books devoted exclusively to a structural modeling approach to longitudinal data. Although there were several books on growth curve analysis, there had been nothing that documented how SEM might be applied to a more complete range of questions about change over time. Introductory texts on SEM do not often deal with the longitudinal case, understandably, and there are quite a number of books on longitudinal analysis that use other statistical methods.

Since first being introduced to SEM in graduate school, I have been conducting longitudinal research and teaching statistics and structural equation modeling courses for more than 20 years. For the last several years, I have been teaching summer statistics workshops, alternating between courses on SEM, multilevel regression, and a more general course on longitudinal data analysis. In teaching the latter course I have included a compendium of analytic approaches, and this course especially made me realize that analysts are faced with an intimidating set of choices when analyzing longitudinal data. Even for the analysis of just two time points, there are an astounding number of possible analyses – repeated measures ANOVA, MANOVA, ANCOVA, regression, difference scores, McNemar's chi-square, marginal homogeneity tests, conditional logistic, and loglinear models. When considering just regression analysis, one has to decide whether to use a lagged model, controlling for an early time point, or try to predict difference scores. During most of my career, I have been reading about and making use of many impressive developments in SEM for analyzing longitudinal data, including time series analysis, survival analysis, latent state-trait models, nonlinear growth curve models, and latent difference score models. In doing so, I thought "Wouldn't it be great to try to bring all of those longitudinal models together under one roof to provide a resource for researchers who want to address any number of

longitudinal hypotheses with SEM?" Well, after more than a year of very intense work, here it is: *Longitudinal Structural Equation Modeling: A Comprehensive Approach*.

In writing this book, I have learned an amazing amount about how the very large family of longitudinal analyses strategies can fit together under one roof. My hope is that readers will too. I hope that in reading this book you will not only see which modeling options are available for which particular hypotheses but will also develop a deeper understanding of structural modeling more generally. The early contributions of Jöreskog, Keesling, and Wiley and others put decades of work on path analysis and factor analysis into one system. In its current form, SEM is an even broader unifying system that encompasses an enormous family of statistical analyses, involving discrete variables, latent class factors, and aspects of multilevel analyses. Like other entities, biological organisms or sophisticated electronics, SEM's basic building blocks, simple and multiple regressions, are fully reducible, but, in my opinion, they come together to form a statistical modeling system that has some emergent properties. It is a powerful system for analyzing many kinds of hypotheses, and this book is an attempt to paint a more complete picture of how this system can be applied to analysis of longitudinal data.

Special Features

There are several special features in this book. Binary and ordinal variables are discussed throughout, a topic that is far too rare in other SEM texts. For those without much background on structural models with these types of variables, the first chapter includes a review of the most important concepts. Comment sections are included frequently within chapters to give general overviews, evaluation of the topic from a balanced perspective, and some practical guidance. So that readers can acquire further knowledge or gain other views and instructors can supplement the text, an annotated list of recommended readings is given at the end of each chapter. New terms are highlighted in the text and definitions are summarized in the Glossary in the back of the book to facilitate use as a course text.

Within every chapter, there is extensive focus on how to apply and interpret each type of model. Every major modeling approach introduced within a chapter is also illustrated with one or more realistic data sets. The data sets used for the analyses are derived from four longitudinal studies that illustrate a variety of measures and study designs. Measures consisting of continuous, binary, and ordinal variable types assess constructs related to social relationships, psychological factors, physical health, and consumer confidence. Study designs capture change across various time frames – three semi-annual waves, six biennial waves, twenty-four daily diary assessments, and a single-case study with 164 monthly economic index values. More details about the data sets are included in the following section titled "Example Data Sets" An accompanying website at www.longitudinalsem.com includes all data sets used in the book and syntax for Mplus and lavaan, the R package, software programs for every example. This book is not about how to use a particular software program, however. There are just too many SEM programs that researchers use these days to tie explanations to just one program.

Many efforts have been made for the material to be detailed and thorough while remaining highly accessible. The book is intended for behavioral and social scientists without any extensive math background, not for statisticians (though I hope some more statistically minded readers will also find it valuable). Perhaps most importantly, model specifications are conveyed through extensive illustration of each major approach. The figures also include the notation for parameters in the model that aid connections between the model specification and the equations. I believe it is essential to present the equations associated with the fundamental concepts in order to engender a solid

understanding of each model. This can be accomplished with simple algebraic formulas, however, and little more math than what most readers learned in high school (or even eighth or ninth grade!) is required. I avoid matrix algebra in all but a few places to increase accessibility of the material. There are a few exceptions when matrices are presented simply to give common matrix notation associated with a model. I use the LISREL notation system in this book, because this notation is widely used in the statistical literature, and I want readers to be able to learn beyond this book. I realize that many readers have been introduced to SEM without being introduced to this notation system. Anticipating that many readers may be unfamiliar with LISREL notation, each symbol in each formula is clearly defined and Appendix A presents a detailed introduction to the notation. Readers should not be deterred by the notation. At least as implemented in this book, it represents little more than replacing English letters used in other texts with Greek letters (e.g., β instead of b and η instead of F). Knowledge of calculus is not needed either, although Appendix B gives a very brief and very basic introduction to some of the concepts of derivatives that may help supplement an understanding of curves that represent change over time.

Outline of the Book

The order of chapters in this book has been carefully considered, and though they can be read out of order, more will be gained if the order of the chapters is followed. The first four chapters certainly should be read before the others, because nearly every later chapter refers back to them. But readers will benefit most by keeping within the order after that as well, as nearly all chapters make use of concepts introduced in prior chapters. For example, the latent transition chapter (Chapter 10) refers back frequently to simplex models from the cross-lagged panel chapter. And the missing data chapter (Chapter 13) requires some knowledge of mixture modeling introduced in Chapter 10.

Chapter 1 reviews some of the key concepts of latent variables, such as latent variable variance and mean identification and analysis of binary and ordinal variables. Chapter 2 applies much of the information from Chapter 1 to assessing longitudinal measurement invariance. SEM tests of dependent means and proportions over few time points are explored in Chapter 3, and basic concepts of stability and change, difference scores, and lagged regression are covered in Chapter 4. The remaining chapters are each primarily devoted to one major type of longitudinal structural equation model. Chapter 5 builds on the preceding chapter by exploring full cross-lagged panel models and simplex models in depth. Chapter 6 focuses on modeling stability with several versions of state-trait models. Chapters 7 and 8 are devoted exclusively to latent growth curve models, with one chapter on linear change and the other chapter on nonlinear change. Latent difference score models are the topic of Chapter 9. Chapter 10 introduces latent class analysis concepts generally and then applies them to latent transition models and growth mixture models. Structural modeling analysis of time series for multiple cases and single cases are discussed in Chapter 11. Chapter 12 shows how SEM can be used to for survival analysis, the approach to modeling observations of events that are censored. Missing data issues are covered in Chapter 13 where they can be discussed in the context of many of the preceding models.

Acknowledgments

Teachers learn far more from their students than either party ever realizes. We benefit not only from student questions but from their insightful suggestions, enthusiasm, and misunderstandings (ours as well as theirs). I am grateful to all the students from my courses over the years (too many to mention), participants in the Portland State Summer Quantitative Method Series, and the PSU quantitative interest group members. I feel unbelievably fortunate to have lucked into an eminent list of statistics teachers during my training to whom I owe an incalculable debt: Ed Wike, Priscilla Hancock, David Thissen, Sandy Braver, Steve West, Leona Aiken, and Dave MacKinnon. I have also learned much from workshops taught by Paul Allison on missing data and survival analysis, Bengt and Linda Muthén on longitudinal analysis, and Stephanie Lanza on latent class and latent transition analysis.

Profound thanks are due to those who took precious time away from their own work to provide feedback on one or more chapters: Janne Adolf, Charles Driver, Craig Enders, Carrie Furrer, Andrew Green, Beth Green, Greg Hancock, Nathalie Huguet, Rex Kline, Stephanie Lanza, Ben Ratliff, Joel Steele, and Manuel Voelkle. I would also like to thank the reviewers Natalie D. Eggum, Arizona State University, Emilio Ferer, University of California, Davis, Brian A. Lawton, George Mason University, Kristin D. Mickelson, Kent State University, and David L. Roth, Johns Hopkins University and the series editor Lisa L. Harlow. Mary Ann Parris Stephens, Melissa Franks, and Karen Rook graciously shared their data set with me and wonderful ideas about possible analyses. Thanks also to several people who swiftly responded with copies of needed chapters or software code: Kevin Grimm, Rex Kline, Jack McArdle, Roger Millsap, Davood Tofighi, and Zhiyong Johnny Zhang. I appreciate the assistance of Linda Muthén, who could always be counted on for prompt replies to my stupid questions about Mplus. I will be forever grateful to Amanuel Zimam, who devoted countless hours to reading all the chapters and translating my Mplus code into lavaan code. Your ability to go from 0 to 60 in about 5 seconds driving an SEM roadster is truly amazing. I appreciate the help of Abdi Hamid for formatting references and Lindsey Cochran who aided in last minute PowerPoint file disaggregation. A huge thank you goes to Debra Riegert for advocating for this project and for patience with deadlines as well as Angela Halliday, the Routledge/Taylor and Francis editorial assistant, for guiding it through the final stages of publication. I am highly appreciative of all the painstaking copy editing by Len Cegielka and by Emma Hart and the team at Out of House Publishing Solutions. Finally, I would not have had the courage to embark on or complete such an enormous task if it was not for the love and support of my wife, Beth, and daughter, Zeta. As much, I appreciate their patience with my more than usual level of distractedness and irritability.

Example Data Sets

The analysis examples in this book are derived from four different sources, though most involve two longitudinal studies. All data sets are available for download from the book's website, http//:www.longitudinalsem.com, along with the syntax used to generate the results. In order to provide the data to readers and protect privacy agreements for these studies at the same time, I created simulated data sets that replicate the sample means and variance–covariance matrices of the real data. Simulating the data enabled me to create realistic examples that would have results that closely parallel the actual results. Statistical conclusions have matched analyses with the real data in every instance I have examined, but readers should not cite the results in these examples as real research findings. Data for most chapters were based on complete cases in order to simplify the examples, but Chapters 12 and 13 use data with missing cases simulated to mirror the characteristics of missingness in the real data.

Social Exchanges Data Set

Many examples use simulated values based on the Later Life Study of Social Exchanges (LLSSE; PI: Karen Rook; see Sorkin & Rook, 2004 for a description). LLSSE is a national longitudinal study of 916 older adults (ages 65 years and over) designed to assess changes in positive and negative social exchanges. Positive social exchanges are social interactions with family, friends, or others that include emotional and instrumental support and companionship. Negative social exchanges involve conflictual interactions with others, such as unwanted advice, criticisms, or neglect. Several key variables from this study are used in the examples in the text, including subscales from the Positive and Negative Social Exchanges (PANSE) scale (Newsom, Rook, Nishishiba, Sorkin, & Mahan, 2005) and a brief version of the Center for Epidemiologic Studies-Depression (CES-D) scale (Radloff, 1977; Santor & Coyne, 1997). One version of this data set also contains information on mortality for up to nine years after the beginning of the study.

Health and Aging Data Set

The health and aging data were simulated from values of the Health and Retirement Study (HRS; Heeringa & Connor, 1995). The HRS is a national health interview of 12,654 adults ages 50 and over that includes a wide range of questions about health, psychosocial variables, and income. Key variables from this study that are used in examples are body mass index (BMI), depression (from the CES-D), and self-rated health, a common single-item measure of perceived global health. The study began in 1992, but the data from the examples was generated from results from just six waves beginning with data collected in 1994, containing data from fewer respondents than the original sample.

Diabetes Data Set

The diabetes data set consists of simulated values derived from a study conducted by Stephens, Rook, and Franks (Iida, Stephens, Rook, Franks, & Salem, 2010), which collected daily diary data over 24 days from 129 patients with Type 2 diabetes. Patients were surveyed repeatedly about a variety of constructs, including daily emotions, perceptions of dietary control, and encouragement from their spouses. The focus of the examples is daily assessment of positive affect measured by the Positive and Negative Affect Scale (Watson & Clark, 1994).

Consumer Confidence Data Set

The consumer confidence data used for single case time series examples in Chapter 11 are actual values from publicly available data collected as part of the Conference Board Consumer Confidence Index. The U.S. Consumer Confidence Index is a monthly assessment of general optimism about the economy based on attitudes and interest in purchasing goods. These data are for 164 months from June 1997 to January 2011, downloaded from http://www.pollingreport.com/consumer.htm.

References

Heeringa, S. G., & Connor, J. (1995). *Technical description of the Health and Retirement Study sample design*. Online version; originally published as HRS/AHEAD Documentation Report DR-002. Retrieved from http://hrsonline.isr.umich.edu/index.php?p=pubs.

Iida, M., Stephens, M. A. P., Rook, K. S., Franks, M. M., & Salem, J. K. (2010). When the going gets tough, does support get going? Determinants of spousal support provision to type 2 diabetic patients. *Personality and Social Psychology Bulletin*, *36*, 780–791.

Newsom, J. T., Rook, K. S., Nishishiba, M., Sorkin, D., & Mahan, T. L. (2005). Understanding the relative importance of positive and negative social exchanges: Examining specific domains and appraisals. *Journals of Gerontology: Psychological Sciences*, *60B*, P304–P312.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychology and Measurement*, *1*, 385–401.

Santor, D. A., & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment*, *9*, 233–243.

Sorkin, D. H., & Rook, K. S. (2004). Interpersonal control strivings and vulnerability to negative social exchanges in later life. *Psychology and Aging*, *19*, 555–564.

Watson, D., & Clark, L. A. (1994). *THE PANAS-X Manual for the Positive and Negative*
6 Affect Schedule – Expanded Form. Unpublished manuscript, University of Iowa.

About the Author

Jason T. Newsom, professor of community health at the Institute on Aging at Portland State University, is a social psychologist. His research has focused on social relationships, health behaviors, and statistics. He has extensive experience working with national longitudinal studies and over 20 years of experience teaching statistics, including courses on structural equation modeling, multilevel modeling, and longitudinal analysis.

This page intentionally left blank

1 Review of some Key Latent Variable Principles

To lay the groundwork needed for acquiring an in-depth understanding of longitudinal structural equation models, this chapter reviews some of the key concepts related to latent variables, including principles of reliability, measurement error, factor variance identification, measurement intercepts and latent variable means, and analysis of non-continuous indicators. There is particular emphasis on how factor variances and means are defined. I emphasize these issues partly to bridge some of the gaps that may exist between the knowledge gained from an introductory course on structural equation modeling (SEM) and the understanding needed to discerningly apply many of the longitudinal structural equation models discussed in this book. My hope is that nearly all readers will find this chapter at least a worthwhile review if not the kernel of more meaningful insights into longitudinal structural models. As a starting point, I assume the reader is familiar with introductory concepts of latent variables and path analysis, including basic principles of path analysis, confirmatory factor analysis, model identification, estimation methods, modeling diagrams, and model fit. Many good sources exist for an introduction to these concepts (e.g., Bollen, 1989; Hayduk, 1987; Kline, 2010; Maruyama, 1997; Mulaik, 2009). This chapter and the remainder of the book uses the Greek-symbol notation system that evolved through the early works of several authors (Jöreskog, 1973, Keesling, 1972; Wiley, 1973), now often referred to as the "LISREL" notation system. It is a worthwhile investment to learn this system, because it is the most widely applied notation in the SEM statistical literature. For those not familiar with the notation system, Appendix A provides a gentle introduction and a convenient reference.

Latent Variables with Continuous Indicators

Variance, Measurement Error, and Reliability

The original conception of reliability and one applied in most physical sciences is that the measure itself is unchanging over time. Upon repeated measurement, given that the underlying construct has not changed, the observed value returned will be identical unless measurement error is present. In other words, an instant retest would produce the same value if the measure is perfectly reliable. The test-retest concept gives rise to the idea of the split-half correlation, developed to assess the reliability of tests with two equivalent forms administered simultaneously. The split-half correlation concept of reliability, in turn, gives rise to the more general idea of internal reliability for measures with multiple observations of the same construct. The classical test theory equation (Novick, 1966) partitions the variance of an observed score into variance due to the true score and variance due to *measurement error*, $\text{Var}(X) = \text{Var}(T) + \text{Var}(e)$. Reliability is defined as the proportion of observed score variance that is true score variance, $\text{Var}(T)/\text{Var}(X)$, with perfect reliability occurring when there is no variance due to error. Alternatively, then, one can think of reliability as the absence of measurement error.

The SEM measurement model parallels the classical test theory equation (Bollen, 1989; Lord & Novick, 1968). If we assume there is only a single indicator of a latent variable, the measurement model can be stated as $y_1 = \lambda_{11}\eta_1 + \varepsilon_1$. In this equation, y_1 is an observed variable, the Greek lambda, λ_{11}, is the factor loading, the Greek eta, η_1, is the latent variable, and ε_1 is the residual. The subscript for the loading, λ_{11}, stands for the first measured variable predicted by the first factor. This equation is the same as the simple regression equation. It is common to set the loading, λ_{11}, equal to 1 when there is only one indicator. This constraint implies a single unit ratio of the variances of the true score to the variance of the observed score, reducing the equation to $y_1 = \eta_1 + \varepsilon_1$. It is then instructive to conceptualize the equation as parallel to the classical test theory equation, where y_1 is the observed score, η_1 is analogous to the true score, and ε_1 is analogous to the measurement error. Inclusion of multiple items assessing a single construct in practice allows for the estimation of true score variance, $\text{Var}(\eta_k) = \psi_{kk}$, and error variance, $\text{Var}(\varepsilon_j) = \theta_{jj}$. The subscript k indexes a particular factor and the subscript j indexes a particular measured variable. This very simple measurement model is not identified without additional constraints, of course, and serves only to illustrate the idea that the observed score variance theoretically can be partitioned into two components, $\text{Var}(y_1) = \text{Var}(\eta_1) + \text{Var}(\varepsilon_1)$. If there is no measurement error, then the variance of the latent variable will be equal to the variance of the observed variable. If there is measurement error, then the variance of the observed variable will be larger than the variance of the latent variable.

An important difference in the meaning of the error term in the SEM measurement model and the error term in the classical test formula is that, for the SEM measurement model, we cannot assume that all the variance unaccounted for by η_k is purely measurement error (Lord & Novick, 1968, p. 535). Because the latent variable is defined by multiple items in practice, the *measurement residual* term incorporates any variance, systematic variance or error variance, that is unique to each measured variable and, therefore, cannot be due to the latent variable. Stated in the form of an equation, $\varepsilon_j = s_j + e_j$. *Systematic variation*, s_j, is any value of the observed score that varies systematically as a function of any variable once the true score is taken into account. Measurement error, e_j, is any remaining random variation, which by virtue of being random cannot be correlated with anything.

Within this latent variable framework, the error term contains information about systematic variance unique to each indicator as well as random measurement error. Accordingly, I will use the term "measurement residual" throughout the book to avoid the implication that the variance unaccounted for by the factor, represented in the error term ε_j, is comprised only of measurement error. The terms "unique variance" or "unique factors" are also appropriate and commonly used. As an example, depression scales typically include an item about sleep disturbance. Response to a question about trouble sleeping may be a function of the underlying construct that the latent variable attempts to represent (e.g., depression) as well as other systematic factors that might affect sleep patterns, such as a physical health problem, that are unique to the question and are not part of the underlying construct represented by the latent variable. If the physical health problem impacts responses to multiple items from the depression scale, the estimate of the latent variable variance will include the systematic variance due to the health problem unless it is otherwise accounted for in the model. If the systematic variance is not modeled, the estimated "true score" variance may represent a construct differing from what the researcher intends (Jöreskog, 1974).[1] A general diagram for a latent variable with four indicators is shown in Figure 1.1. Notation is added for the factor variance, ψ_{11}, loadings, λ_{jj}, and measurement residual ε_j, where ψ_{kk}, $= \text{Var}(\eta_k)$ and $\text{Var}(\varepsilon_j) = \theta_{jj}$.

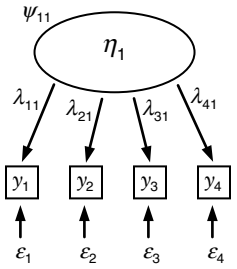

Figure 1.1 Single Factor Measurement Model, no Mean Structure.

Factor Variances and Covariances

Latent variables are unobserved entities that must be assigned meaning through the measurement of their observed indicators. In a certain statistical sense, the variance of any variable provides the basis of its interpretation and meaning, so considering how the variance of a latent variable is defined provides the key to a more intuitive understanding of latent variables. Latent variable variances are derived from the variances and covariances of the observed variables. We know this from the familiar decomposition of the covariance matrix into the constituent matrices of loadings, factor variances, and measurement residuals. The decomposition is summarized in the matrix expression of the structural equation measurement model.

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{1.1}$$

$\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the variance–covariance matrix of observed variables, $\boldsymbol{\Lambda}$ is the factor loading matrix, $\boldsymbol{\Lambda}'$ is the transposed factor loading matrix, $\boldsymbol{\Psi}$ is the factor variance matrix, and $\boldsymbol{\Theta}$ is the measurement residual matrix. At the individual item level, the variance of an observed variable is a function of the factor variance, the loading, and the measurement residual variance. The decomposition of the covariance matrix implies two important equations among factor loadings, factor variances, observed variable covariances, and measurement residual variance given by Equations (1.2) and (1.3) below (e.g, Bollen, 1989).

$$\mathrm{Var}\left(y_j\right) = \lambda_{jk}^2 \mathrm{Var}\left(\eta_k\right) + \mathrm{Var}\left(\varepsilon_j\right) \tag{1.2}$$

$$\mathrm{Cov}\left(y_j, y_{j^\circ}\right) = \lambda_{jk}\lambda_{j^\circ k}\mathrm{Var}\left(\eta_k\right) \tag{1.3}$$

The superscript $^\circ$ represents a variable other than j (i.e., $j \neq j^\circ$). Equation (1.4) illustrates that the factor variance can be derived from the square of any loading and the associated measurement residual variance.

$$\mathrm{Var}\left(\eta_k\right) = \frac{\mathrm{Var}\left(y_j\right) - \mathrm{Var}\left(\varepsilon_j\right)}{\lambda_{jk}^2} \tag{1.4}$$

For each of the above equations, $\mathrm{Var}(\varepsilon_j) = \theta_{jj}$ and $\mathrm{Var}(\eta_k) = \psi_{kk}$ are individual elements of the $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ matrices of the measurement model given by Equation (1.1). It may seem odd that Equation (1.4) can be true, given that there is only one factor variance and each loading will tend to have a different value. The loadings, measurement residual variance, and factor variance, however, are estimated so that the equation is true for each indicator, and, thus, constraining one of these values will affect the other values.

Equation (1.3) shows that the covariance between two indicators for a latent variable, denoted by j and j° subscripts, is equal to the product of their respective loadings, λ_{jk} and $\lambda_{j^{\circ}k}$, and the variance of the factor, ψ_{kk}. This equality can be explained in terms of Wright's (1934) path analysis tracing rules, where a covariance between any two variables connected in a path diagram can be decomposed into a product of the constituent paths (Loehlin [2004] and Maruyama [1997] provide accessible introductions). There are some underlying assumptions to this equation, such as independence of the residual from the latent variable and independence of the residuals, which I leave to others to detail (e.g., Bollen, 2002).

Factor Variance Scaling. The equations above are not of any practical use without additional information. The factor variance cannot be identified unless we impose a scaling constraint.[2] Typically, either the variance of the factor is constrained to be equal to 1, which I will call the *factor variance identification* approach, or one of the loadings is constrained to be equal to 1, which I will call the *referent loading identification* approach (also commonly known as the "marker variable" approach). A third identification approach, effects coding identification (Little, Slegers, & Card, 2006), will be explained later in the chapter. The effects coding identification approach is less commonly used, but has some distinct advantages in the context of longitudinal models.

The choice of identification approach is an arbitrary one in some respects. The same model tested using any of the three identification approaches will have identical fit and standardized loadings. Although the unstandardized loadings will not be equal under the three identification approaches, they are algebraically equivalent – the values obtained with one identification approach can be computed from the values obtained with another identification approach. I begin with a discussion of key concepts and connections for the two simpler and more commonly used identification approaches, the factor variance identification and referent loading identification approaches, followed by a discussion of the effects coding approach.

Based on Equations (1.2) through (1.4), it can be shown that loadings obtained under the factor variance identification approach can be derived from the estimated factor variance when the referent loading identification approach is used (Hayduk, 1987, p. 181). For example, the value of the first loading, λ_{11}, obtained when the factor variance is set equal to 1 for identification, is equal to the square root of factor variance obtained when a referent loading is used, so that $\lambda_{11} = \sqrt{\psi'_{11}}$. I use the prime symbol to indicate values obtained from the referent loading approach. The value of any other loading, say λ_{21}, estimated when the factor variance is set to unity can be calculated using the estimate of the loading, λ'_{21}, and the factor variance, ψ'_{11}, obtained under the referent loading approach to identification.

$$\lambda_{21} = \lambda'_{21}\sqrt{\psi'_{11}} \tag{1.5}$$

Within a particular identification method, loadings from the same factor are related to one another through their covariances and the factor variance. If we assume a factor with only two indicators as a simple (albeit underidentified) case, Equation (1.6) helps illustrate that one loading can be derived from another loading, if the covariance among the two variables and the factor variance are known.

$$\lambda_{21} = \frac{\mathrm{Cov}(y_1, y_2)}{\lambda_{11}\psi_{11}} \tag{1.6}$$

This relation indicates that the second loading is proportionate to the first loading. Combining quantities from Equations (1.3), (1.5), and (1.6), and using a little algebra

reveals that any loading from one identification approach can be obtained from a simple ratio of two loadings obtained from the other identification approach. For example,

$$\lambda'_{21} = \frac{\lambda_{21}}{\lambda_{11}} \tag{1.7}$$

where λ_{11} and λ_{21} are freely estimated loadings when the factor variance identification approach is used, and λ'_{21} is the loading when the referent loading identification approach is used.

Example 1.1: Factor Loadings and Factor Variance

To demonstrate the algebraic equivalencies, I specified a three-indicator latent variable using three items from the social exchanges data set ($N = 574$). The example data sets are described in the Example Data Sets section at the beginning of the book. Syntax and data sets used in the examples are available at the website for the book. The three items comprise a measure of perceived companionship (ratings of frequency with which network members "were good company," "included you," and engaged in "recreational activities"), which can be viewed as one domain of supportive social exchanges (Newsom, Rook, Nishishiba, Sorkin, & Mahan, 2005). Two of the three loadings and the factor variance are presented in the first section of Table 1.1 and the relation among these parameters using the referent loading identification and the factor identification approaches are demonstrated. As the table makes clear, loadings obtained from analysis using one method of identification can be used to obtain loadings from the other method of identification. Factor means and measurement intercepts, which are discussed next, are also included in the table.

Means and Intercepts

Structural means are integral to many longitudinal models, so it is essential to develop a thorough understanding of how latent variable means are determined. Although many structural equation models omit means, they are the focus of certain types of longitudinal modeling hypothesis tests. Just as latent variable variances are a function of observed variable variances, latent variable means are a function of observed variable means.

Table 1.1 Computational Examples Illustrating the Algebraic Equivalences of Referent and Factor Identification Solutions (Social Exchanges Data Set)

Referent identification	*Factor identification*	*Computations*
$\lambda'_{11} = 1$	$\lambda_{11} = .918$	$\lambda_{21} = \lambda'_{21}\sqrt{\psi'_{11}} = 1.152\sqrt{.844} = 1.058$
$\lambda'_{21} = 1.152$	$\lambda_{21} = 1.058$	$\lambda'_{21} = \dfrac{\lambda_{21}}{\lambda_{11}} = \dfrac{1.058}{.918} = 1.152$
$\psi'_{11} = .844$	$\psi_{11} = 1$	$\nu'_2 = \nu_2 - \nu_1\lambda'_{21} = 1.979 - (1.960)(1.152) = -.278$
$\nu'_1 = 0$	$\nu_1 = 1.960$	$\nu'_2 = \nu_2 - \nu_1\dfrac{\lambda_{21}}{\lambda_{11}} = 1.979 - 1.960\left(\dfrac{1.058}{.918}\right) = -.278$
$\nu'_2 = -.278$	$\nu_2 = 1.979$	$\nu_1 = E(y_1) = \alpha'_1 = 1.960$
$\alpha'_1 = 1.96$	$\alpha_1 = 0$	$\nu'_2 = E(y_2) - \lambda'_{21}\alpha'_1 = 1.979 - (1.152)(1.960) = -.278$
$E(y_1) = \bar{y}_1 = 1.960$		
$E(y_2) = \bar{y}_2 = 1.979$		

Mean structures. Inclusion of mean structures in SEM expands the measurement model to include a measurement intercept (Sörbom, 1974). Consider the measurement equation for a single indicator,

$$y_1 = \nu_1 + \lambda_{11}\eta_1 + \varepsilon_1 \tag{1.8}$$

where ν_1, the Greek "nu", is the measurement intercept for one measured variable. Figure 1.2 illustrates a four-indicator, one-factor measurement model with mean structures (for an accessible introduction, see Chapter 11 in Kline, 2010). This diagram is identical to that shown in Figure 1.1, except that notation for each intercept (ν_1 through ν_4) is added at the end of each factor loading arrow.

To estimate means in the model, the observed matrix to be analyzed must include mean information (i.e., fitting the moment matrix), either by including a vector of means when the variance–covariance matrix is input or by supplying the program with raw data from which the means can be obtained. A special adaptation of maximum likelihood estimation (mean-weighted maximum likelihood) is needed to fit models with mean structures (refer to Bentler & Yuan, 2000 and Hayduk, 1987, for an explanation of this rationale). Means are not typically estimated by default in most SEM programs and must be requested. An exception is when missing data estimation is requested, which will invoke mean estimates for the model. Programs usually include a set of defaults for mean structure analysis in order to identify the model, and users must add specifications to change these defaults.

Sufficient information is needed to identify the mean structure of the model, because measurement intercepts and factor means are estimated from the observed means. There cannot be more unknowns than equations. The formula for degrees of freedom expresses the relation between the number of observed values and unknown parameters, with degrees of freedom equal to 0 indicating the model is just identified and positive degrees of freedom indicating the model is overidentified. To calculate degrees of freedom, the usual formula is extended to

$$df = \left[\frac{J(J+1)}{2} + J\right] - q$$

Here, J is the number of observed variables and q is the number of parameters free to be estimated. Counting the number of free parameters and determining whether a model is identified is not always readily apparent (Rigdon, 1994). The number of observed means will be equal to the number of observed variables in any standard application, so J appears in two terms inside the brackets to count the number of unique variance–covariance elements and to count the number of means. For a single latent variable with four indicators, the number of variance–covariance elements in the obtained matrix is [4(4 + 1)]/2 = 10 and

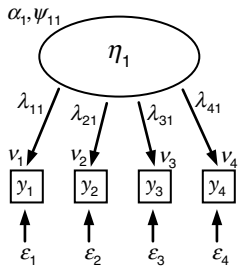

Figure 1.2 Four-Indicator Measurement Model with Mean Structure.

the number of observed means is 4, which equals 14 parameters available to be estimated. For the model illustrated in Figure 1.2, $q = 12$, because there are three loadings, one factor variance, four measurement residuals, three measurement intercepts, and one factor mean estimated.[3] The degrees for freedom for this example, therefore, equal 14 – 12 = 2. Notice that the inclusion of the means using one of the two common strategies for identification leads to degrees of freedom that are equal to a model without any mean structure. They are equal in this model and many commonly specified models, because the number of required mean estimates is equal to the number of observed means provided with the data.

Measurement Error and Means. Measurement error affects estimates of means differently from the way it affects variances. In classical test formula terms, measurement error has no impact on the true score, because its expected value $E(e)$ is zero.

$$E(X) = E(T) + E(e)$$
$$E(X) = E(T) + 0$$
$$E(X) = E(T)$$

Systematic factors, however, can have a positive or negative expected value, and, therefore, may affect the true score, an effect usually referred to as *measurement bias* if it is connected to an important social group.

For latent variables with multiple indicators, the expected value of a measured variable, $E(y_j)$, is a function of the measurement intercept, ν_j, and of the product of the loading, λ_{jk}, and the factor mean, α_k, for a particular indicator, j, and a particular factor, k.

$$E\left(y_j\right) = \nu_j + \lambda_{jk}\alpha_k$$

As with regression analysis, where the intercept is a function of the predictor mean and the unstandardized regression slope, $\beta_0 = \bar{Y} - \beta_1\bar{X}$, the measurement intercept for one measured variable in the latent variable model is simply the mean of the indicator variable minus the product of the unstandardized loading and the factor mean.

$$\nu_j = E\left(y_j\right) - \lambda_{jk}\alpha_k \tag{1.9}$$

In other words, the intercept is the mean of the indicator variable conditioned on or adjusted by the unstandardized loading and the factor mean.[4] Manipulation of Equation (1.9) shows the factor mean as a function of the observed mean, the measurement intercept, and the loading.

$$\alpha_k = \frac{E\left(y_j\right) - \nu_j}{\lambda_{jk}}$$

A loading estimate that is larger than 1.0 will lead to a factor mean that will be smaller than the observed mean for the indicator variable. A loading estimate that is smaller than 1.0 will lead to a factor mean estimate that will be larger than the mean of the indicator variable. It may seem strange that a single mean for the factor can be derived from this formula when the values for each indicator differ, but each measurement intercept takes on a unique value for each indicator making the equality true given its observed mean and indicator.

Factor Mean Scaling. As with factor variances, a scaling constraint is needed to identify the mean structure. In one approach, which I will call the *referent intercept identification* approach, the loading is set equal to 1 and the intercept is set equal to 0. This approach is

also commonly referred to as the "marker variable" approach. With this scaling, the mean of the factor is equal to the expected value of the indicator variable.[5]

$$E(y_j) = \alpha'_k \tag{1.10}$$

The factor mean is denoted as α'_j to signify that it is estimated under the referent intercept identification approach. Analogously to the classical test theory formula where the expected mean of the observed score equals the expected mean of the true score, the expected value of the observed score for the referent variable and the mean of the latent variable model are equal. By implication, SEM does not provide an advantage in estimating means by removing measurement error, because measurement error does not affect the mean estimate.

For all other indicators of the factor under the referent intercept identification approach, the intercept estimate will be a function of the observed mean of the indicator and a proportional weighting of the factor mean as shown in Equation (1.11).

$$\nu'_2 = E(y_2) - \lambda'_{21}\alpha'_1 \tag{1.11}$$

Because, under this identification constraint, the factor mean is equal to the observed mean of the referent variable (Equation 1.10), the other intercepts will be a function of the observed mean for the referent variable.

$$\nu'_2 = E(y_2) - \lambda'_{21}E(y_1)$$

The other common scaling constraint, which I will call the *factor mean identification* approach, sets the factor mean equal to 0, the factor variance equal to 1, and freely estimates the measurement intercepts and loadings. Equation (1.11) implies that if the factor mean is set to 0, the estimate of the intercept for each of the indicator variables will be equal to its observed mean.

$$\begin{aligned} \nu_1 &= E(y_1) - \lambda_{11}\alpha_1 = E(y_1) - \lambda_{11}(0) = E(y_1) \\ \nu_2 &= E(y_2) - \lambda_{21}\alpha_1 = E(y_2) - \lambda_{21}(0) = E(y_2) \\ &\vdots \\ \nu_i &= E(y_i) - \lambda_{jk}\alpha_j = E(y_i) - \lambda_{jk}(0) = E(y_i) \end{aligned} \tag{1.12}$$

Because the second term drops out regardless of the value of the loading, the identity relation between the intercept and the observed mean holds for each of the indicators (and holds even if we were to use a referent loading approach for identifying the factor variance). The factor mean identification approach does not explicitly provide any information about a factor mean for a single latent variable measured at one time point, but it can be useful for examining mean differences, a topic discussed further in Chapter 3.

We can show the relation between the two identification approaches for mean structures. Means and intercepts obtained when using one identification approach can be computed from the means and intercepts obtained when using the other identification approach. Drawing from Equations (1.10) and (1.12), the observed mean for the referent variable is equal to its intercept (ν_j) if the factor mean identification approach is used, and the observed mean is equal to the factor mean (α'_j) if the referent intercept identification approach is used.

$$E(y_1) = \nu_1 = \alpha'_1$$

An intercept for any non-referent indicator under referent intercept identification will be a function of the intercepts and its loadings obtained under factor mean identification.

$$\nu_2' = \nu_2 - \nu_1\lambda_{21}'$$

Similar to the link between the loading under the referent loading identification specification and the ratio of loadings under the unit factor variance specification (Equation 1.7), we also know that the intercept is a function of the ratio of factor loadings.

$$\nu_2' = \nu_2 - \nu_1 \frac{\lambda_{21}}{\lambda_{11}} \tag{1.13}$$

Thus, we see that factor means and intercepts are dependent on one another and that measurement intercepts are dependent on one another and a ratio of the factor loadings.

Example 1.2: Measurement Intercepts and Factor Means

The second section of Table 1.1 also illustrates how intercepts and means obtained using the two identification approaches can be derived from one another using the social exchanges data set for the same factor model discussed in Example 1.1.

Effects Coding Identification Approach

Constraints on a referent indicator or the factor values themselves are the traditional approaches to identifying factor variances and means and are, by far, the most widely used approaches. There are several inconveniences to these identification approaches for interpretation or implementation, however, particularly when it comes to longitudinal applications. Scaling of the factor mean or variance to one particular indicator could be problematic if that indicator is not representative of the other indictors. Differences in the factor mean, for example, will in fact represent differences in the mean of the indicator variable that serves as the referent. Similarly, a factor variance of 1 and factor mean of 0 are convenient loci, but they are values that are removed from the original metric of the variables. Moreover, setting the factor variances is not an option for longitudinal models for scaling endogenous factors to test for longitudinal measurement invariance.[6] Little and colleagues (2006) proposed the *effects coding identification* approach as an alternative scaling method that sets factor variance and mean using complex parameter constraints. Each will be discussed in turn.

Factor Variance Scaling. To scale the factor variance using the effects coding method, the value of one loading for a factor is constrained to be a function of the other loadings for the factor. Consider that the average of several loadings is the sum of the estimated values divided by the number of loadings on the factor. The average loading for a factor with three indicators, for example, would be $\bar{\lambda}_{j1} = (\lambda_{11} + \lambda_{21} + \lambda_{31}) / 3$. It follows that requiring one of the loadings to be equal to the number of loadings minus the remaining loadings will produce an estimate of the factor variance that is a function of the three loadings. For the three indicator example, we would require the first loading to be equal to

$$\lambda_{11} = 3 - \lambda_{21} - \lambda_{31}$$

When this requirement is made for the estimate of the first loading, the factor variance will be a weighted function of the covariance of the indicators. With a little alteration to

Equation (1.3), the factor variance can be shown to be a function of the covariance among pairs of observed variables and the factor loadings. Stated in terms of the first two loadings, for instance, the factor variance is equal to covariance between the two variables weighted by the product of the two loadings.

$$\psi_{11} = \frac{\mathrm{Cov}(y_1, y_2)}{\lambda_{11}\lambda_{21}} \tag{1.14}$$

In general, the factor variance will be equal to the average of the covariance between each pair of indicators divided by the product of their respective loadings for J observed indicator variables.

$$\psi_{kk} = \frac{\sum_{1}^{J}\left[\mathrm{Cov}\left(y_j, y_{j^\circ}\right) / \lambda_{jk}\lambda_{j^\circ k}\right]}{J} \tag{1.15}$$

The subscript j° is used to denote another variable or loading for the same factor. Under these constraints, the average of the loadings will be equal to 1. Consequently, the variance of the factor can then be considered an average of the covariances, putting the scaling of the variance in terms of the variances and covariances among the indicators.

Factor Mean Scaling. A similar strategy is used to scale the factor mean so that it will be a weighted average of the observed indicator means. For the three indicator example, the estimate of one of the measurement intercepts is constrained to be a function of the two remaining measurement intercepts.

$$\nu_1 = 0 - \nu_2 - \nu_3$$

These constraints lead to a factor mean that is a weighted average of the observed means, $\bar{y}_j$.

$$\alpha_k = \frac{\sum_{1}^{J}\bar{y}_j\lambda_{jk}}{J} \tag{1.16}$$

The advantage of effects coding identification is that the factor mean can be interpreted in terms of all of the observed indicators rather than in terms of the mean of only one observed indicator or an arbitrary 0 value.[7] As will be apparent in later chapters, specifying the factor mean as a weighted average has great utility for the interpretation of longitudinal structural models that have means as a concern. A disadvantage of the effects coding identification approach is that the SEM software must have special features that allow for complex equality constraints, though most, but not all, programs presently include these capabilities.[8] Another potential disadvantage is that the weighted average may weight some indicators more heavily than others if they have substantially different metrics or variances. A simple linear transformation of the observed values could be used to put any disparate items on a similar metric, however.

Comments

A solid understanding of latent variable variances and means is an invaluable foundation for judicious application of longitudinal structural equation models. Two important ideas concerning reliability and measurement error have been introduced thus far.

First, measurement error adds to the variance of the true score, which can bias estimated associations among variables (disattenuation). This issue is discussed at greater length in Chapter 4. Second, latent variable mean estimate does not benefit from the removal of measurement error, because random error does not have a biasing effect on the estimate of the mean. Statistical comparisons of means may benefit from the estimation of measurement error, however. This issue is discussed further in Chapter 3.

It is instructive to consider how latent variable variances are derived, because latent variable variances are fundamental building blocks of nearly all the longitudinal models discussed in this text. Model specification choices and appropriate interpretation of results will frequently depend on how the latent variable variance has been defined. For example, we saw that estimates of loadings and factor variances are inextricably connected, a fact that plays a role in some of the essential longitudinal measurement invariance concepts that will be introduced in Chapter 2. We also saw that scaling choices for latent variable variances are arbitrary in the sense that the solution obtained under one specification approach can be computed from another specification approach. This algebraic equivalence is the reason that significance tests and standardized solutions are unaltered by various approaches to factor variance identification. Although scaling choices seem inconsequential for simple, cross-sectional models, they are more consequential for many longitudinal models, warranting further considerations. Choices about the scaling of the latent variable means is generally more critical than the scaling of the latent variable variances in estimating longitudinal models where means are a principal interest, because statistical tests and conclusions from longitudinal models involving individual level change (e.g., second-order growth curve or latent difference models) or mean differences (ANOVA models) can be affected by scaling choices.

Latent Variables with Binary Indicators

SEM with binary variables may receive only cursory coverage or no coverage in introductory courses or texts, so I will give a brief introduction to some of the core concepts. It is important to review some of the special issues involved, because there are some unique considerations when binary indicators are used to estimate latent variables. I will also delve further into a few selected aspects that will later have special relevance to longitudinal models. Following the discussion of binary variables, a few additional considerations for ordinal variables will be addressed. Throughout this chapter, I will assume estimation of continuous latent variables. Categorical latent variables can be estimated using binary, ordinal, or continuous indicators in latent class models, which are discussed in Chapter 10.

Following the same general rationale that underlies the necessity for alternative estimation of regression models (for reviews of regression modeling with binary and ordinal variables, see Long, 1997; O'Connell, 2006), special considerations are needed when estimating structural equation models whenever binary variables are endogenous in the model. Binary or ordinal predictor variables in regression analysis and exogenous variables in SEM require no special treatment, however. There are two principal reasons why ordinary least squares for regression or maximum likelihood estimation for continuous variables in SEM should not be used whenever the dependent variable is binary. One reason is that a perfect relationship between a continuous predictor and the outcome probability (that $y = 1$ given x) is best fit to an S-shaped curve. The second reason is that the normal error distribution assumption underlying continuous regression approaches will be violated if the outcome is binary, leading to biases in the standard errors and significance tests. The same justification applies to structural equation models whenever the model includes a binary dependent variable for any of the paths or whenever latent variables include binary indicators.[9]

Because the latent variable measurement model regresses a binary indicator on the latent variable, a simple regression model is implied in which the outcome is binary. Binary and ordinal variables thus require special estimation procedures and a number of unique considerations for interpreting path coefficients, loadings, and measurement residual variances.

These special considerations aside, most of the general concepts related to measurement error and mean structures with continuous indicators apply similarly to binary indicators. As with continuous indicators, measurement error (along with unique systematic variance) is removed from the variance of the latent variable estimate, following the general logic of the formulas shown in Equations(1.2) and (1.4). With binary indicators, the means of the latent variable have a special interpretation tied to the proportions of the observed variables. Latent variable means also are unaffected by measurement error if indicators are binary, as is the case with continuous indicators. The following sections review the special issues related to estimation and interpretation of structural models with binary dependent variables.

Threshold Conceptualization

A convenient way of conceptualizing regression models with a binary and ordinal dependent variable is to posit an unobserved continuous distribution underlying the observed variable. To simplify matters as much as possible, we will consider only binary variables first and then return to ordinal variables later. The concept of the *threshold* is based on the notion that if there were to be an unobserved continuous variable, often termed y^*, that determines the observed binary value, then the continuous variable could be considered a propensity toward which the observed value would likely be equal to 1 if a certain value on the continuous variable has been exceeded. For example, an employee may only be promoted after a certain threshold of accomplishment has been exceeded. We may only have information about the promotion, but there may be some hypothetical or real continuous value of success that underlies the observed promotion. Although many binary variables can be regarded this way, an underlying continuous variable does not have to be theorized in order for the statistical model to apply. The threshold idea is merely a useful conceptual tool. The threshold concept is usually used to explain probit regression analysis and the generalized linear model, but can be applied analogously to logistic regression (Aitchison & Silvey 1957; McKelvey & Zavoina 1975; Nelder & Wedderburn, 1972). Figure 1.3 is a visual analogue for the correspondence between an observed variable y and the underlying unobserved variable y^*. Values below the threshold τ on y^* are observed as $y = 0$, and values above the threshold on y^* are observed as $y = 1$.

The y^* concept is also employed in the context of correlation analysis (Olsson, 1979), where y is conceptualized as a variable that crudely categorizes an otherwise continuous variable. Dichotomization of continuous variables results in an attenuation of the correlation coefficient in comparison to the value of the correlation that would be obtained if the variable had not been dichotomized (Cohen, 1983; Peters & Van Voorhis, 1940). In this sense, the categorized variable can be seen as containing a degree of inaccuracy or error in the measurement of the more precise continuous variable. Special correlation coefficients correct for this attenuation. Tetrachoric (for binary variables), polychoric (for ordinal variables), or polyserial (for binary or ordinal variables mixed with continuous variables) correlations (heretofore all are referred to collectively as "polychoric" correlations) take into account the loss of information when observed binary variables are used as representations of unobserved, continuous, and bivariate normally distributed y^* variables. Each is a special case of the general approach.

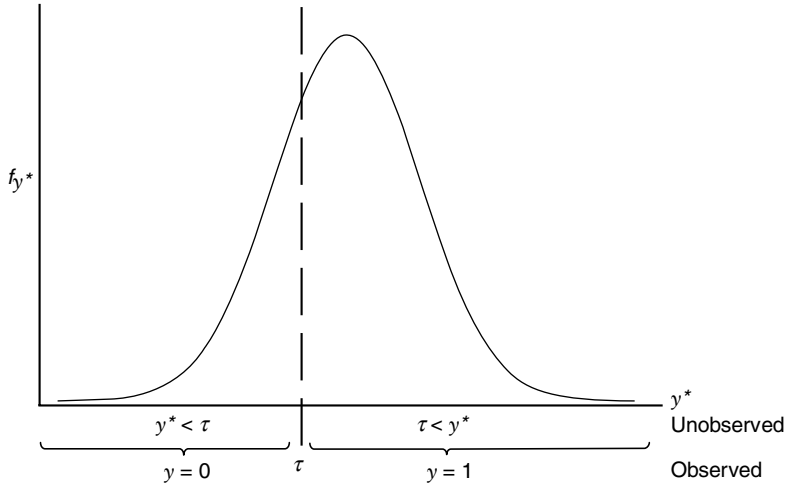

Figure 1.3 The Relation between the Unobserved Distribution of y^* and the Observed Values of y.

The mathematical connection employed with regression models, or *link function*, between y and y^* is nonlinear, because a perfect relationship between a continuous predictor and the probability that a binary variable is equal to 1 tends to have a slight nonlinear function. Two common link functions are used for regression models with a binary dependent variable – the logit function and the probit function. The concept of the link function allows us to retain the linear regression model on the right-hand side of the equation if we consider a more complicated function for y^*.

$$y^* = \beta_0 + \beta x + \varepsilon \tag{1.17}$$

Part of the convenience of the y^* formulation is that the right-hand side of the equation is the same as the common linear regression model. The logistic model put into this form can be said to be "linear" in the parameters. Throughout the text, I use the Greek β to denote an unstandardized regression parameter estimate rather than b, in part to emphasize the equivalence of regression estimates and path estimates. For similar reasons, the error or residual is given by ε rather than e. Logistic and the closely related probit regression require different link functions, discussed in more detail next.

Link Functions for Binary Variables

The logit link function used with logistic regression is a logarithmic function, whereas the inverse normal distribution link function used with probit regression is a more complicated mathematical function. Note that neither of these analysis approaches involves a transformation of the observed y scores. It is the predicted values that are transformed in order to optimally fit the observed binary values to a prediction line. In practice, results from probit and logit regression models typically produce very similar results and lead to identical statistical conclusions (Long, 1997).

Logit Link Function. The logit link function involves a fairly simple transformation of the predicted scores using the natural logarithm of the probabilities.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta x$$

In this equation, p is the probability that $y = 1$, and ln is the natural logarithm (base e). The left-hand side of the equation is often referred to as the *logit* of the probability. To return to the raw probability that $y = 1$, the exponential function, which is the complementary function to the natural logarithm, is employed. For a regression model with a single predictor, the probability that the observed y is equal to 1 taking into account, or conditioned on, x is

$$p \mid x = \frac{1}{1 + e^{\beta_0 + \beta x}} \tag{1.18}$$

On the left hand side of the equations, $p|x$ represents the conditional probability of y given x, $p|x = \mathrm{P}(y = 1|x)$, which provides useful information about the predicted scores for a particular value of x. Equation (1.18) is derived from the standard logistic *cumulative distribution function* (cdf), and, when stated in this form, it is a reference for the probability that a value is equal to a greater than a certain point.[10] By inserting a value for x and the obtained values for the regression coefficients, the predicted probability that $y = 1$ for this value of x can be computed. Transformation of the slope using the exponential function, e^{β}, gives the odds ratio. The exponential function uses the mathematical constant equal to approximately 2.718 as the base. An odds ratio is the odds that $y = 1$ compared with $y = 0$ for each increment in x. The same transformation of the intercept, e^{β_0}, gives the odds that $y = 1$ if x is equal to 0 (i.e., how many more times y is equal to 1 compared with how many times y is equal to 0, given that x is equal to 0). If x is binary, then e^{β_0} returns the proportion of cases in the $x = 0$, $y = 1$ group.

Probit Link Function. Probit regression estimates use an inverse normal link function to linearize the regression model and transform the predicted values of y to an underlying y^* value.

Instead of the logit transformation of $\ln\left[p/(1-p)\right]$, the probit transformation is a more complicated transformation based on the inverse of the cdf for the standard normal curve with mean of 0 and variance of 1.

$$y^* = \Phi^{-1}(p \mid x) = \beta_0 + \beta_1 x$$

The standard normal cdf, symbolized by Φ, can be used to obtain the proportion values in the distribution at or below the specified value.

$$\Phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}z^2\right) dz$$

Sometimes referred to as the "normal ogive," the normal cdf is the same function used to obtain cumulative probabilities from the z-table. Fortunately, hand computations are no longer necessary. But, just for the record, the value z is a standard normal distribution score, π is the mathematical constant equal to approximately 3.142, and exp is the power function using the constant e as the base.

The probit regression slope estimates represent the increment in y^* for each unit change in x. Because the transformation leads to interpretation of the y^* outcome in standardized value terms, the standardized solution is convenient because the standardized slope represents the change in y^* for each standard deviation change in x. The cdf can convert

unstandardized regression estimates into a predicted probability value, using a z-table or computer spreadsheet, by inserting a desired value of x and the coefficient estimates.

$$p \mid x = \Phi\left[-\left(\beta_0 + \beta_1 x\right)\right] \tag{1.19}$$

The normal cdf is usually defined as the proportion of the distribution equal to or less than the specified value, but subtracting the result from 1 or reversing the sign of the coefficients, as I have done in Equation (1.19), produces the value the probability that $y = 1$.

Estimation Methods

With SEM applications for binary or ordinal variables, it is not appropriate to use regular maximum likelihood (ML) estimation designed for continuous variables, and special estimation must be requested. There are two estimation approaches commonly used for binary variables, maximum likelihood estimation and weighted least squares (WLS).[11]

Maximum Likelihood Estimation. Maximum likelihood (ML) estimation for binary variables, sometimes referred to as "marginal maximum likelihood," is less widely available in SEM software programs and is less commonly employed by researchers than WLS[12] but binary ML has some interpretation advantages in that logistic regression path coefficients can be obtained then converted to odds ratios. The commonly employed link for ML is the logit link, giving logistic estimates, but a probit link is also possible. (Assume ML implies a logit link in this text unless otherwise indicated.) As a "full information" estimation approach, it has more often been employed to study testing and measurement under an item response theory (IRT) approach than for general applications of SEM. Binary ML estimation also is likely to have performance edge when standard missing data assumptions are not met, because all available information for each case is used in the likelihood estimation (see Chapter 13 for more on this topic). The ML approach uses an expectation maximization (EM) algorithm with numeric integration where polychoric correlations are analyzed instead of the raw covariance matrix. This estimation process can be computationally intensive with large models and can produce biased standard errors, unless remedial options are used. Robust estimates (Satorra & Bentler, 1994) for standard errors work well for sample sizes over approximately 250 and estimates are superior to the unadjusted full information ML estimates according to simulation studies (e.g., Yang-Wallentin, Jöreskog, & Luo, 2010). The robust ML method appears to perform similarly to the WLS with mean and variance adjustments (WLSMV; see below) method, with a potential advantage of the robust ML method when the latent variable distribution is skewed (DeMars, 2012).

Weighted Least Squares Estimators. Weighted least squares (WLS)-based approaches are more commonly used than binary ML, because they are widely available in SEM software programs. WLS is a general estimation approach also implemented in regression analysis. With regression analysis, cases in the data set are differentially weighted when fitting the regression line in order to address violations of distributional assumptions of the dependent variable or outliers. Also known in the SEM literature as asymptotic distribution free (ADF) or arbitrary generalized least squares (AGLS) estimation, SEM software programs use WLS to differentially weight residuals derived from fitting the implied and obtained covariance matrices. The asymptotic covariance matrix, which includes the variance and covariances of the parameter estimates (and, thus, information about standard errors), is the weight matrix for the version of WLS employed in the estimation of structural

equation models. WLS/ADF/AGLS estimation is rarely applied to raw data or the observed covariance matrix, however, because very large samples are typically required for accurate standard error estimates.

More modern approaches to WLS developed for binary and ordinal variables have much better small sample performance. These WLS-based methods use a multiple-step estimation involving polychoric correlations as input to create the asymptotic covariance matrix used for weighting in the WLS estimation (see Finney & DiStefano, 2013 for a review). This estimation strategy has several advantages, including better standard errors, chi-square statistics, and the availability of several standard fit indices. The method is sometimes referred to as a "limited information" estimator, because the full information table is not used when analyzing the polychoric correlations in the presence of missing data. WLS-based approaches may not work as well as full information methods when standard missing data assumptions are not met (more details can be found in Chapter 13). One version of this WLS estimation approach, diagonal weighted least squares (DWLS), uses a diagonal weight matrix instead of the inversion of a full weight matrix (Muthén, 1984; 1993), which appears to have the most statistical and computational efficiency (Muthén, du Toit, & Spisic, 1997). DWLS combined with robust (Satorra–Bentler) standard errors and a mean-and-variance-adjusted chi-square statistic generally performs well even for fairly small sample sizes (e.g., $N > 200$; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009). Examples throughout the text use one particular implementation of the robust DWLS method, which I will refer to as weighted least squares with mean and variance adjustments (WLSMV).[13] Given appropriate specifications (see below), this estimation approach produces probit estimates.

Factor Variance and Measurement Residual Variance

The essential elements of the measurement model when indicators are binary do not differ from the measurement model when indicators are continuous. What differs is that the measurement equation can be conceptualized as a logit or probit regression with the unobserved y^* serving as the outcome. For the simple single variable case, we can write

$$y_1^* = -\tau_1 + \lambda_{11}\eta_1 + \varepsilon_{11} \tag{1.20}$$

In the binary case, the negative of the threshold $-\tau_1$ replaces the intercept ν_1. The model differs from the continuous case, because association between the latent variable and the observed indicator is nonlinear and requires an alteration of the predicted value as in the logit or probit regression model. If binary ML is used, the link function for y^* is the logit link, and, if WLSMV is used (with one of the parameterization methods), the link function for y^* is the probit link. Latent variables will generally be defined by three or more indicators, so Equation (1.20) is generalized to the multivariate case in the SEM measurement model shown in Equation (1.1). As with continuous indicators, the factor variance must be identified by using a scaling constraint based on one of the three approaches (referent loading, factor variance, or effects coding).

Because y^* is an unobserved variable with an unknown distribution, the link between y and y^* cannot be made without identifying constraints on the residual variance. This is analogous to the need for an identifying constraint in defining the factor variance. In a general sense, the scaling constraint is arbitrary because y^* is an unknown variable whose variance is defined by the constraint. Although results will be similar under most conditions, ML and WLSMV results will not be identical. Each link function has corresponding

assumptions about the distribution, where the logistic distribution is assumed for ML estimation and a normal (probit) distribution is assumed for WLSMV estimation.

ML Logistic Estimates. Binary ML estimation if used with a logit link function (as is typical) produces path coefficients that are equivalent to logistic regression coefficients. Consequently, coefficients can be converted to odds ratios using the same logistic transformation formula, e^{β}, with the odds ratio representing the change in the odds that y is equal to 1 given one unit increment in x (or η if the predictor is a latent variable).

Binary ML estimation identifies the y^* distribution with the logistic distribution constraining the measurement error residual variance, θ_{jj}, while estimating the variance of y^*. The variance of y^* is given as

$$\mathrm{Var}(y^*) = \lambda_{jk}^2 \psi_{kk} + (\pi^2 / 3) \tag{1.21}$$

The last term on the right is a common scaling metric for the variance of the logistic distribution (Menard, 2010), and is approximately equal to $(3.142)^2 / 3 \approx 3.290$. As the equation makes apparent, the values of the loadings and the factor variance are dependent on how the variance of y^* is identified. The square root of the variance, $sd(y_j^*) = \sqrt{\mathrm{Var}(y^*)}$, can be used to obtain standardized loadings, λ_{jk}^*, with the following formula:

$$\lambda_{jk}^* = \frac{\sqrt{\psi_{kk}}}{sd(y_j^*)} \lambda_{jk} \tag{1.22}$$

Note that this equation exactly parallels the formula used for standardized coefficients in regression, $\beta^* = \beta(sd_x/sd_y)$, where β is the unstandardized coefficient. Because the latent variable η_k substitutes for x, the square root of the variance of the factor $\sqrt{\psi_{kk}}$ substitutes for the standard deviation of x. If factor variance identification is used with the factor set equal to 1 and all loadings are freely estimated, then the computation of the standardized loadings simplifies to $\lambda_{jk}^* = \lambda_{jk} / sd(y_j^*)$.

WLSMV Probit Estimates. The WLSMV method has two options for the scaling constraints on the y^* distribution. The *delta parameterization* (sometimes known as "marginal parameterization") constrains the residual variances, θ_{jj}, and estimates the residual variance of y^*. Because the assumed distribution of y^* is normal, however, the logistic scaling metric $\pi^2/3$ is replaced by 1.

$$\mathrm{Var}(y^*) = \lambda_{jk}^2 \psi_{kk} + 1 \tag{1.23}$$

Standardized loadings can be obtained according to Equation (1.22) as long as this modified value for the variance is used for the standard deviation of y^*.

In addition to the delta parameterization, identification can be achieved by constraining the variance of y^* and estimating measurement residual variance. The *theta parameterization* (or sometimes "conditional parameterization") standardizes the residual variance of the y^*, with $\mathrm{Var}(y^*) = 1$, which corresponds with the standard normal y^* distribution presumed in probit regression. The theta parameterization has some advantages because equality constraints on the measurement residual variance are possible, such as when measurement invariance of residuals is of interest, and correlated measurement residuals can be added to the model. The delta and theta parameterization are equivalent in the sense that the fit of the model, parameter significance, and standardized coefficients are unchanged by the parameterization (Finney & DiStefano, 2013).

Because of this relation to the probit model, predicted probabilities can be estimated for parameters obtained from theta parameterized WLSMV estimation by using the normal cdf as in Equation (1.19). For the measurement equation, values for the threshold, loading, and desired value of η_k can be inserted into the equation to obtain the estimated probability that the observed $y = 1$ at a specified value of η_k.

Factor Means and Measurement Thresholds

In the case of binary indicators, measurement intercepts and latent variable means provide information about proportions. In general, where measured variables are coded 0 and 1 (e.g., for "no" and "yes" responses), the average of all the responses gives the proportion endorsing a response option of 1, providing information about the proportion of the sample responding "yes," for instance. Because of this intuitive interpretation of means with this coding scheme, it rarely if ever makes sense to use coding schemes other than 0 and 1 for binary variables.

The relationship between the factor mean and the threshold in the binary case is the same as the relationship between the factor mean and the intercept in the continuous case. If the referent identification approach is used (setting the threshold for one indicator equal to 0) the factor mean will be equal to the value of the threshold for the referent indicator obtained using the factor mean identification (setting the factor mean equal to 0). The values are of opposite sign, because the threshold is estimated for the probability that $y = 0$, whereas the mean corresponds to the probability that $y = 1$, thus $-\tau = \alpha'$. Recovering observed proportions for binary indicators is more complicated than recovering the observed mean in the continuous indicator case, however, because of the scaling constraints involved in identifying the y^* distribution. A cdf conversion is needed to connect the factor mean or measurement threshold estimates back to observed proportions.

ML Logistic Estimates. Because the measurement model is a logistic regression when binary ML estimation is used, the conditional probability that y is equal to 1 can be obtained using the logistic transformation. But because the factor variance and loadings are dependent on the variance of y^*, the transformation is not as simple as the usual transformation given in Equation (1.18). If factor identification is used, where the mean and variance are 0 and 1, respectively, and the measurement intercepts and loadings are freely estimated, then the threshold gives the marginal proportion for the observed y. The approximate value of the probability that an observed indicator y_j equals 1 can be obtained by dividing by the threshold estimated standard deviation of y^* and adding an adjustment multiplier (the value 1.7) commonly used in conjunction with the standardized logistic cdf (Long, 1997).

$$P(y_j = 1) = \frac{1}{1 + e^{1.7(\tau^*)}} \tag{1.24}$$

The formula uses the standardized threshold, which takes into account the standard deviation of y^*, where $\tau^* = \tau / sd\left(y_j^*\right)$. The standard deviation of y^* is $sd\left(y_j^*\right) = \sqrt{\text{Var}\left(y^*\right)}$, with Var($y^*$) as defined in Equation (1.21). Alternatively, the observed probability can be obtained using the normal cdf with the standardized threshold τ^*, as in Equation (1.25) below. The standardized threshold may be printed with the standardized solution and computation may not be needed. When the referent threshold is constrained for identification instead of the factor mean, the marginal probability for the referent indicator can

be obtained by replacing the threshold in Equation (1.24) with the estimated factor mean, α_k. Under this identification strategy, the thresholds for the non-referent items no longer correspond to the marginal probabilities and must be adjusted using a ratio of the loadings as in the same way as continuous indicators.

WLSMV Probit Estimates. Because the WLSMV estimation has two possible parameterizations, the application of the normal cdf to the measurement intercepts or factor means to obtain proportions depends which parameterization is used in the estimation. Under the delta parameterization, the thresholds correspond to the standard normal cdf, which can be used to find the probability that the observed y is equal to 1. As with the continuous indicators, when factor mean identification is used, with α_k set to 0, the factor variance is set equal to 1, and all thresholds are freely estimated, threshold estimates for each indicator correspond to their respective marginal proportions.

$$P\left(y_j = 1\right) = \Phi\left(-\tau_j\right) \tag{1.25}$$

Transformation using the standard normal cdf is indicated by Φ and can be accomplished with most standard spreadsheet programs. If the factor mean is identified by setting the threshold to 0 for a referent indicator y_1, then the mean of the factor can be used to obtain the marginal probability for the referent indicator, $P\left(y_1 = 1\right) = \Phi\left(\alpha'_k\right)$. The probabilities for any non-referent indicator must be obtained after taking into account the ratio of the loading to the referent loading, however.

The theta parameterization makes a different assumption about the distribution of y^*, estimating its variance and setting the measurement residual variance equal to 1. The different variance for y^* necessitates a scaling adjustment if the standard normal cdf is used to obtain proportions. The scaling factor for each indicator, $\Delta_j = 1/\sqrt{\lambda_{jk}^2 \psi_{kk} + 1}$, gives the relation between the y^* under the delta parameterization, where it is freely estimated, and y^* under the theta parameterization, where it is constrained to be equal to 1 (Muthén & Asparouhov, 2002).

Multiplying the intercept estimate under factor mean identification then gives the marginal probability that the observed y will be equal to 1 for theta parameterization.

$$P\left(y_j = 1\right) = \Phi\left[\Delta_j\left(-\tau_j\right)\right]$$

Marginal probabilities are obtained for each indicator under the factor mean identification, and if the referent threshold identification is used (with the referent loading set equal to 1), the same equation can be used to find the marginal probability for the referent indicator by substituting α'_k for $-\tau_j$.

Effects Coding Identification. When the effects coding identification approach is used (Little et al., 2006), the factor mean corresponds to a weighted average of the observed proportions, which can be obtained using the logistic or normal cdf conversion. This identification method uses complex constraints on the intercept, so that the estimate of one intercept is a function of the remaining intercept estimates. For example, the following constraint would be used for a factor with three indicators:

$$\tau_1 = 0 - \tau_2 - \tau_3$$

Using the appropriate cdf transformation with the factor mean estimate, the weighted average of the observed proportions could be reproduced. The effects coding approach

differs from the proportion obtained from the referent identification and factor identification methods, but it is one that takes into account all of the indictors. It is worth noting that these complex constraints are applied to the linear coefficients from the model, and therefore represent log (ML) or inverse normal (WLSMV) constraints on the probabilities. Although these nonlinear constraints parallel constraints used in other contexts (e.g., longitudinal invariance, proportional hazards), it is advisable to keep in mind that the constraints represent a more complicated link to the observed data than is the case with continuous variables.

IRT Interpretation. Readers familiar with psychometrics may recognize that Equation (1.17) parallels the IRT equation used in educational testing and elsewhere. The IRT formulation describes the connection between true ability, θ, and observed scores, conceptually represented by $y^* = -b + a\theta$. (The theta in this equation should not be confused with the theta used for the elements of the measurement residual matrix). It is known as the two-parameter model; the parameter b is the difficulty (or location) parameter, which is the mean probability of correct response conditioned on ability, and a is the discrimination parameter, which describes the relation between ability and the response. Because 0 typically refers to an incorrect response, $-b$ is used instead of b, so that higher values of the difficulty parameter, b, will be associated with lower probability of a correct response. A link function can then be used to relate the observed variable, y, to the latent ability factor, θ, producing the item characteristic curve used throughout testing applications for general item analysis and item bias assessment.[14] If theta parameterization is used and the factor variance is standardized, the discrimination parameter, a, is equal to the loading, λ_{ik}, and the difficulty parameter, b, is equal to the negative of the threshold, $-\tau$ (Kamata & Bauer, 2008), eliminating any need for computations to obtain the IRT parameters and the item characteristic curve. Kamata and Bauer (2008) and DeMars (2012) are excellent sources for how these specifications affect the transformation of factor analysis estimates to IRT parameters.

Example 1.3: Factor Means and Intercepts with Binary Indicators

Three questions about unwanted advice from family and friends – "Give you unwanted advice"; "Question or doubt your decisions"; and "Interfere or meddle in your personal matters" – were originally measured on a 5-point scale but were dichotomized (no unwanted advice vs. any unwanted advice) for the purposes of illustration. Artificial dichotomization generally results in the loss of information and can potentially lead to incorrect conclusions (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002), so I do not recommend it as standard practice. A one-factor model was tested with the three binary indicators, and results are presented in Table 1.2. Conversions of binary MLR (ML with robust standard errors) and WLSMV produce marginal probabilities very close to the observed proportion (.923) of those responding affirmatively to the question "Give you unwanted advice."

Comments

I have discussed two common factor estimation methods used with binary variables and shown the connection to traditional regression models, such as logistic and probit analysis. Both robust ML and WLSMV perform well under most standard conditions, suggesting that there are no major statistical reasons to prefer one over the other. The referent, factor, and effects coding identification approaches are all possible with binary variables,

Table 1.2 Computation of Marginal Proportions from Threshold Estimates with Binary Indicators, $P(y_1 = 1) = .923$

	Referent identification	*Factor identification*	*Computations*
MLR	$\tau_1' = 0$ $\alpha_1' = 8.002$ $\lambda_{11}' = 1$ $\psi_{11}' = 29.219$	$\tau_1 = -8.043$ $\alpha_1 = 0$ $\lambda_{11} = 5.439$ $\psi_{11} = 1$	$sd\left(y_1^*\right) = \sqrt{\lambda_{11}^2\psi_{11} + \pi^2/3} = \sqrt{\left(5.439^2\right)(1) + 3.29} = 5.733$ $\tau_1^* = \tau_1 / sd\left(y_1^*\right) = -8.043/5.733 = -1.403$ $P\left(y_1 = 1\right) = \dfrac{1}{1 + e^{1.7(\tau^*)}} = .917$ $P\left(y_1 = 1\right) = \Phi\left[\left(-\tau_1^*\right)\right] = .920$
WLSMV (delta)	$\tau_1' = 0$ $\alpha_1' = 1.428$	$\tau_1 = -1.428$ $\alpha_1 = 0$	$P\left(y_1 = 1\right) = \Phi\left(\alpha_1'\right) = .923$ $P\left(y_1 = 1\right) = \Phi\left(-\tau_1\right) = .923$
WLSMV (theta)	$\tau_1' = 0$ $\alpha_1' = 10.361$ $\lambda_{11}' = 1$ $\psi_{11}' = 51.660$	$\tau_1 = -13.110$ $\alpha_1 = 0$ $\lambda_{11} = 9.127$ $\psi_{11} = 1$	$\Delta_1' = 1/\sqrt{\lambda_{11}^2\psi_{11} + 1} = \sqrt{1^2\left(51.660\right) + 1} = 1/7.256 = .138$ $\Delta_1 = 1/\sqrt{\lambda_{11}^2\psi_{11} + 1} = \sqrt{9.127^2\left(1\right) + 1} = 1/7.256 = .109$ $P\left(y_1 = 1\right) = \Phi\left[\Delta_1\left(-\tau_1\right)\right] = .923$ $P\left(y_1 = 1\right) = \Phi\left(\Delta_1\alpha_1'\right) = .923$

MLR = maximum likelihood with robust standard errors; WLSMV = weighted least squares with mean and variance adjustments.

and we can see that each can be derived from the other. Thus, the general principles of the derivations of the factor variances and factor means are similar when binary indicators are used to define a latent variable. Although the connection between observed proportions and factor means is more complicated with binary variables than the connection between observed means and factor means is with continuous variables, it is clear that factor means are just functions of the observed variable values. Though the choice of factor identification may seem arbitrary in the end, the interpretation of unstandardized effects in longitudinal structural models will depend on these choices. The fit of the model and standardized solutions obtained under referent and factor identification approaches will not differ, however. Just as with continuous variables, the effects depend on the factor identification method if the referent or factor identification approach is chosen instead of the effects coding approach. Effects coding also has the advantage of combining all indicators in the computation of the factor mean, whereas the other two approaches must base the factor mean on only one of the indicators.

Latent Variables with Ordinal Indicators

The y^* framework can be extended to variables with more than two ordinal categories, where $C - 1$ thresholds are required to link the observed variable with the underlying variable. The IRT concepts can also be applied to the ordinal case and application is referred to as the graded response model (GRM; Samejima, 1969). Either ordinal ML or WLSMV estimation with the same estimation process as outlined for binary variables may be applied to ordinal indicators. Indicators must have rank ordered values, such as "none of the time," some of the time," and "all of the time" or "unemployed," "part-time job," and "full-time job," rather than nominal categories, such as "Protestant," "Catholic,"

and "Jewish." Nominal categories require a multinomial estimation approach which differs in some respects. As with binary indicators, the ordinal ML method with logit link function produces ordinal logistic estimates, whereas the WLSMV method produces probit regression estimates when the theta parameterization is used. Probit path or loading parameters represent the change in y^* for each unit change in the x variable (either measured predictor or latent variable), and standardized estimates can be used to characterize the change in y^* for each standard deviation increase in x. Odds ratios may also be used with ML estimation to provide information about the average odds that the observed ordinal variable, y, increments by 1 unit (i.e., from 0 to 1, 1 to 2, etc.) for a unit increase in x.

With ordinal measured indicators, thresholds still provide information about observed category proportions, but the thresholds no longer correspond with the marginal probability of the observed variable. With more than two categories, there must be $C-1$ thresholds for each indicator, and a referent threshold for one indicator or the factor mean should be set equal to 0 for identification purposes. More formally, for any number of thresholds, $y = c$, if $\tau_c < y^* \leq \tau_{c+1}$, where c is a particular response category of observed variable y. Similar to the binary case, logit or probit transformations can be used to calculate the observed proportions from the thresholds. Use of the cumulative normal distribution function returns the proportion at or below the threshold, and, thus, the proportions reflect the cases for all of the lower categories ($y \leq c$).

Because of the additional thresholds that must be estimated, the extra parameters place limits on the number of parameters that can be estimated overall. The number of possible free parameters is equal to $[p(p-1)]/2 + (C-1)$, counting each threshold as one of the parameters to be estimated. As with binary variables, delta or theta scaling constraints may be used for identifying the y^* distribution for WLSMV estimates. The process for recovering the cumulative observed proportions is the same for ordinal variables as it is for binary variables. Standardization and logistic transformation (Equation [1.24]) or normal cdf (Equation [1.25]) can be used with ML estimates and the normal cdf is used (Equation [1.25]) for WLSMV estimates.

Other Variable Types

In addition to binary and ordinal dependent variables, special estimation is needed for other variable types, including nominal categorical variables, count variables, or variables with zero-inflated distributions. Though estimation algorithms for these variable types have begun to be added to SEM software programs (e.g., Muthén & Muthén, 1998–2012), they are not widely available to date. Because most of the general concepts introduced here for continuous, binary, and ordinal variables also can be applied to other variable types, I will not discuss them (for an introduction, see Liu & Powers, 2007; Muthén & Asparouhov, 2009). When the dependent variable represents a discrete event that may occur after the period of observations for some cases, it is considered censored and is best suited for survival analysis models, which are discussed in Chapter 12.

Comments

This review provides some critical background for gaining a more comprehensive understanding of longitudinal structural models. The fundamental concepts of latent variable variances and means are necessary for distinguishing among the hypothesis testing goals of various longitudinal models and forming appropriate conclusions from their results.

Factor identification is usually considered an arbitrary choice, yet it becomes a more important consideration of many longitudinal modeling issues, such as longitudinal measurement invariance which I will turn to in the next chapter.

Recommended Readings

There are several global discussions on the meaning and utility of latent variables that supply a very useful theoretical background for many of the topics discussed in this chapter (Bollen, 2002; Bollen & Hoyle 2012; Borsboom, Mellenbergh, & van Heerden, 2004). I recommend two chapters, one by Kline (2010: Chapter 11) and one by Hayduk (1987: Chapter 9), for lengthier discussions of mean structures, a topic that is often neglected in many SEM introductions. Little, Slegers, and Card (2006) describe effects coding identification, which has some advantages over other identification methods and is not commonly discussed in introductory texts. Long's (1997) text on regression with non-continuous variables, has a particular emphasis on probit regression and the threshold model and is an invaluable resource on the topic. Finney and DiStefano (2013) provide an accessible and comprehensive introduction to the issues related to testing structural equation models with binary and ordinal variables. Several papers have a somewhat more technical, but also more detailed, discussion of structural modeling with ordinal variables, giving important specifics on interpretation, threshold identification, y^* scaling, and estimation that are generally not covered elsewhere (Muthén, 1993; Muthén & Asparouhov 2002; Jöreskog & Moustaki, 2006). A useful overview of the translation of IRT modeling concepts into an SEM framework is provided by Kamata and Bauer (2008).

Notes

1 When systematic variation common to two or more indicators of a latent variable is not estimated in the model as one or more covariances among measurement residuals or a second latent variable, then the systematic variation will become part of the latent variable variance estimate. Specifying covariances among measurement residuals (or "correlated errors"), however, will remove the common systematic variance from the latent variable variance estimate. Note that the covariance among measurement residuals must be between systematic factors associated with each measured variable, because, by definition, measurement error is random and cannot be correlated with any other factors.

2 The formula for degrees of freedom available from any introductory SEM text is $\frac{J(J+1)}{2} - q$ here, J is the number of observed variables and q is the number of parameters estimated in the model.

3 Because constraining the factor variance and factor mean would also identify the model, counting J variances, J loadings, and J intercepts as free parameters will give the same total.

4 One can also say that the mean of y_j, or expected value of $E(y_j)$, is a total effect decomposed into direct effects, ν_j, and indirect effects, $\lambda_{jk}\alpha_k$ (Kline, 2010).

5 Note, however, that if the intercept is set equal to 0 but the loading is estimated freely (i.e., different items are used for the mean and variance scaling constraints), then there will no longer be a simple relationship between the observed mean and the factor mean, because the factor mean will be proportionately weighted by the loading for the referent variable that has its intercept set equal to 0. Thus, it will rarely if ever make sense to use different items for the loading referent and the intercept referent, because this one-to-one relationship of the factor mean to the referent variable's mean will be lost.

6 In Chapter 2, I discuss yet another scaling alternative that sets the factor mean and variance only for the first time point, the *single occasion identification* approach, which must be used in conjunction with longitudinal equality constraints in order to identify the model.

7 Meredith and Horn (2001) make the case that very different conclusions may be derived about mean changes depending on the scaling choices made for each factor, noting that use of different referent intercepts to identify factor means could even lead to different conclusions about whether the factor means increase or decrease over time. They recommend a similar solution in terms of basing factor means on a composite average of the indicator means (weighted or unweighted) using a two-step approach. Their Bartlett factor score approach has considerable resemblance to the effects coding approach proposed by Little and colleagues, whose approach is much simpler to implement when software features permit.

8 As of this writing, several major packages include a feature for setting complex constraints, including LISREL, Mx, Mplus, Amos (basic syntax), and R lavaan.

9 I largely ignore the circumstance in which there is a mix of binary and continuous dependent variables used in either the measurement or structural portions of the model. A mix of variable types is certainly acceptable, and, except for declaring the binary dependent variables in the model, no additional efforts are required for the analyst. SEM software programs handle the combined estimation automatically, though interpretation of parameter estimates needs to be tailored to the type of dependent variable.

10 An alternative, equivalent form of this equation is $p \mid x = \frac{e^{\beta_0 + \beta x}}{1 + e^{\beta_0 + \beta x}}$ which returns the probability equal to or less than the threshold value (i.e., that $y = 0$). With either version of the equation, its complementary interpretation can be obtained by reversing the sign of the coefficients. Both equations are based on the standard logistic distribution which assumes a mean of 0 and variance of $\pi^2/3$, which uses the mathematical constant *pi*, equal to approximately 3.142.

11 Presently, not all software programs include the same estimation options for binary and ordinal variables. The current version of Amos, for example does not include either a WLS approach or the ML approach, providing a Bayesian estimation approach instead. Although there appears to be increasing interest in Bayesian estimation for binary variables, it has not been widely adopted in software programs or by researchers as of yet. I therefore leave discussion of this method to other sources (e.g., Lee, 2007; Levy & Choi, 2013).

12 Of the major general SEM software programs, only LISREL and Mplus currently offer ML (full information) estimation for binary and ordinal dependent variables.

13 For software programs that implement a WLS estimator, not all implement the process in the same way. LISREL requires a two-step approach using the PRELIS preprocessor to construct polychoric correlations which then must be analyzed in a separate step in LISREL (Jöreskog & Sörbom, 1996). The WLSMV estimation in Mplus (1998–2012), on the other hand, is a robust DWLS method that generates the polychoric correlations for estimating the asymptotic covariance matrix and applies robust adjustments in one automatic process. EQS (Bentler & Wu, 2002) uses an alternative robust method described as a "partitioned maximum likelihood" approach, obtaining estimates in separate steps depending on the types of variables involved.

14 In a special case of the latent variable model with binary indicators, specifying all loadings equal to 1 and a standardized latent variable produces values for the Rasch model (Takane & de Leeuw, 1987).

References

Aitchison, J., & Silvey S. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, *44*, 131–140.

Bentler, P. M., & Wu, E. J. C. (2002). *EQS 6.1 for Windows user's guide*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Yuan, K.-H. (2000). On adding a mean structure to a covariance structure model. *Educational and Psychological Measurement*, *60*, 326–339.

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*, 605–634.

Bollen, K. A., & R. H. Hoyle. (2012). Latent variables in structural equation models. In Rick H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 56–67). New York: Guilford Publications.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249–253.

DeMars, C. (2012). A comparison of limited-information and full-information methods in mplus for estimating item response theory parameters for nonnormal populations. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 610–632.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age Publishing.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 625–641.

Hayduk, L. A. (1987). *Structural equations modeling with lisrel.* Baltimore, MD: Johns Hopkins University Press.

Jöreskog, K. G. (1973). Analysis of covariance structures. *Multivariate Analysis*, *3*, 263–285.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology*, Vol. II. New York: W. H. Freeman.

Jöreskog, K. G., & Moustaki, I. (2006). *Factor analysis of ordinal variables with full information maximum likelihood*. Retrieved from http://www.ssicentral.com/lisrel/techdocs/orfiml.pdf.

Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: User's reference guide.* Chicago, IL: Scientific Software International.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 136–153.

Keesling, J. W. (1972). *Maximum likelihood approaches to causal analysis* (Ph.D. thesis). University of Chicago.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach*. New York: Wiley.

Levy, R., & Choi, J. (2013). Bayesian structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 563–623). Charlotte, NC: Information Age Publishing.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Liu, H., & Powers, D. A. (2007). Growth curve models for zero-inflated count data: An application to smoking behavior. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 247–279.

Loehlin, J. (2004) *Latent Variable Models* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Lord, F. M., & Novick, M. R.(1968). *Statistical theories of mental test scores*. Reading, MA Addison-Wesley.

Maruyama, G. (1997). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40.

McKelvey, R. & Zavoina W. (1975). A statistical model for the analysis of ordinal level variables. *Journal of Mathematical Sociology*, *4*, 103–20.

Menard, S. (2010). *Logistic regression* (2nd ed.). Thousand Oaks, CA: Sage.

Meredith, W., & Horn, J. L. (2001). The role of factorial invariance in modeling growth and change. In. L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington DC: American Psychological Association.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. New York: CRC Press.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.
Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.
Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf.
Muthén, B., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC Press.
Muthén, B., du Toit, S. H., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Retrieved from http://www.statmodel.com/bmuthen/articles/Article_075.pdf.
Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.) Los Angeles: Muthén & Muthén.
Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370–384.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.
O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage.
Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443–460.
Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill.
Rigdon, E. E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling*, *1*, 274–278.
Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Iowa City, IA: Psychometric Society.
Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.
Takane, Y., and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.
Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In: A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences*. New York: Seminar.
Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*, 161–215.
Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 392–423.

Keywords
latent variables, mean structures, structural equation modeling, binary, ordinal

2 Longitudinal Measurement Invariance

Any longitudinal structural model should begin with a complete understanding of the measurement properties of the latent variables used in the analysis. The objective is to ensure that measurement properties of latent variables are stable over time, termed *measurement invariance*, to avoid mistaking changes in measurement properties for hypothesized changes in the construct. Millsap and Cham (2011) more formally define longitudinal invariance as the case in which the conditional distributions of observed values are unchanging given the same latent variable values over time. The most obvious source of changes in measurement properties is when item wording, response options, or measurement procedures are modified at one or more waves of data collection. Even if there are no changes in wording, response options, or procedures, however, measurement properties may still vary, often for unknown reasons. SEM provides the most versatile and precise tool for investigating longitudinal measurement invariance.

Discussions of measurement invariance issues have nearly always focused on group comparisons (e.g., Byrne, Shavelson, & Muthen, 1989; Meredith, 1964; Millsap, 2011; Vandenberg & Lance, 2000), and only a few authors have considered longitudinal measurement invariance specifically (Bontempo, Grouzet, & Hofer, 2012; Millsap & Cham, 2011; Widaman, Ferrer, & Conger, 2010). Although the two contexts have many general concepts and analysis strategies in common, it is valuable to consider many of the issues specific to longitudinal measurement invariance.

Nested Model Tests

To investigate measurement invariance, statistical tests comparing nested models are conducted to determine whether observed differences are greater than what would be expected due to chance (Fix, Hodges, & Lehmann, 1959). I first discuss some important issues related to these statistical tests before discussing procedures specific to investigating longitudinal measurement invariance.

Likelihood Ratio Test

The SEM approach makes possible precise statistical comparisons of parameter estimates across groups or over time through imposition of equality constraints. Comparisons should always be made using raw, unstandardized variables, because important variance information is lost if variables are standardized, potentially leading to incorrect conclusions (Cudeck, 1989). Invariance tests are conducted by comparisons of nested models with and without constraints of individual or sets of parameters. Nested models, at minimum, involve the same cases and the same measured variables in the models being compared. Two models are compared that have the same structure, with one model including some

restrictions on free parameters (i.e., more degrees of freedom). Bollen (1989) defines a nested model as "any model which requires that some function of its free parameter equals another free parameter or equals a constant is nested in the identical model that has no such restriction" (p. 291), but constraints also may include particular linear or nonlinear determinant functions between two parameters or imposing non-equivalent restraints between parameters. It may not always be readily apparent that two models are nested (Bentler & Satorra, 2010; Hershberger, 2006). A common example is that a one-factor confirmatory factor analysis (CFA) is nested within a two-factor CFA with the same items, because the two-factor model implies a correlation between the two latent variables equal to 1.

Nested models are generally compared with a chi-square difference test, or more formally called a *likelihood ratio test,* where $\Delta\chi^2 = \chi^2_{M0} - \chi^2_{M1}$. Model M0 is nested within M1, because additional restrictions on the parameter estimates have been made, leading to a model chi-square for M0 that will be equal to or greater than the less restricted model, M1. The difference in chi-square values is compared to the chi-square distribution using degrees of freedom equal to the difference in degrees of freedom from the two models, $\Delta df = df_{M0} - df_{M1}$. A significant result indicates that the parameters that are constrained to be equal do indeed differ, or, are "non-invariant." Other statistical tests are possible, such as Lagrange multiplier or Wald tests (Yoon & Millsap, 2007), but these are generally not used for a priori comparisons in practice.

The simple chi-square difference test is not appropriate unless data are multivariate normal and standard maximum likelihood (ML) estimation is used. With the Satorra–Bentler scaled chi-square for nonnormal data or the Yuan–Bentler scaled chi-square for nonnormal missing data (Yuan & Bentler, 2007), the chi-square difference test can be computed using the following formula (Satorra, 2000; Satorra & Bentler, 2001):

$$\Delta\chi^2_{SB} = \frac{\chi^2_{M0} - \chi^2_{M1}}{\left(df_{M0}scf_{M0} - df_{M1}scf_{M1}\right)/\, df_{M0} - df_{M1}}$$

The difference in chi-square values is scaled by the difference in degrees of freedom and the scaling correction factor (scf) for the two models, a weighting value based on multivariate kurtosis used in computing the Satorra–Bentler scaled chi-square. The scf is equal to the ratio of traditional ML chi-square to the Satorra–Bentler scale chi-square for the model, or $\text{scf} = \chi^2_{ML} / \chi^2_{SB}$.[1] Because the ML estimate is inflated proportionate to the amount of multivariate kurtosis, the ratio of the scf becomes larger with greater kurtosis. The weighted difference in chi-squares can be negative in some instances, and an alternative testing procedure can be used if this arises (Satorra & Bentler, 2010).

Nested tests for ordinal analysis methods are not widely available in software programs currently, and there has been limited simulation work comparing methods. One suggestion has been to use a weighted least squares (WLS) estimator just for comparison of model chi-square values using the simple difference for chi-square and degrees of freedom. An alternative is a more elaborate vanishing tetrad test (Hipp & Bollen, 2003). Asparouhov and Muthén (2006) have adapted the tests developed by Satorra (2000) and Satorra and Bentler (2001) that compute the estimated ratio of the weighted likelihoods of two models using weighted least squares with mean and variance adjustments (WLSMV) estimation for ordinal variables.[2]

Binary and ordinal ML estimates also may be used to compare model chi-square values in nested tests (e.g., Kim & Yoon, 2011; Woods, 2009). Item response theory (IRT) software, and more recently, some SEM programs, use likelihood ratio tests derived from the log likelihood deviance statistic using full ML approaches or use Pearson chi-square values. The ML fit statistics generally do well with binary or few ordinal categories but may

become less accurate fit estimates with more categories and more items (Maydeu-Olivares, Cai, & Hernández, 2011). Robust ML estimates for binary models are becoming increasingly available, and if they are used for nested tests, the weighted difference tests using the scaling correction factor should be used as discussed above.

Effect Size

In conducting chi-square difference tests for determining longitudinal invariance, it is important to distinguish between statistical significance and practical importance. There may be many circumstances where statistical power to detect differences is high, even for rather trivial departures from invariance. In such circumstances, researchers may choose to consider the assumption of invariance to be essentially met for practical purposes, concluding that any bias that occurs from violating invariance assumptions is of minor importance. Decisions about the magnitude of the violations will need to depend on the researcher's knowledge about the topic, the nature of the research question, and the standards of practice in the area of research, however.

The opposite circumstance is also a potential concern. There may be too little power to detect significance even though the magnitude of the difference is large. The likelihood ratio test for measurement invariance has the goal of finding in favor of the null hypothesis (Hancock, Stapleton, & Arnold-Berkovits, 2009), and the null hypothesis cannot be proven to be true. Evidence in support of invariance is only evidence that does not contradict it, leaving open the possibility that invariance may not be the true state of affairs. In practice, too much statistical power may be more of a problem than too little statistical power, because simulation work (French & Finch, 2006; Marsh, Balla, & McDonald, 1988; Saris & Stronkhorst, 1984) suggests that chi-square difference tests generally have sufficient power for the minimum sample sizes typically recommended for SEM (e.g., approximately 100 for continuous, normally distributed, and non-missing data).

Either problem requires some way to gauge whether a likelihood difference is small or large in magnitude (Brannick, 1995; Kelloway, 1995). Many researchers informally use the percentage change in the chi-square value as a rough approximation of the effect size of the difference. Consider an example in which the chi-square value for the model with no constraints is approximately 250 with 20 degrees of freedom and the chi-square for the model with equality constraints is approximately 260 with 21 degrees of freedom. The difference would be statistically significant at alpha equal to .05, but the percentage chi-square increase in the model with constraints would be fairly small, $(260/250 - 1) \times 100 = (1.04 - 1.00) \times 100 = 4\%$. Even if the difference is statistically significant, researchers may not consider 4% to be large in some circumstances. In other circumstances, 4% may be considered too high to be considered trivial. A problem with this approach is that it does not take into account the difference in degrees of freedom and so may not be ideal.

Another simple assessment of the magnitude of the difference of the fit of two nested models is to adapt the w effect size estimate used for contingency chi-square effect size computations.

$$w = \sqrt{\frac{\Delta\chi^2}{N(\Delta df)}}$$

The value of w is equal to *phi* (and Pearson's correlation coefficient) in the 2×2 contingency chi-square or Cramer's V in the $n \times m$ case. The advantage of this index is that it can be used to reference standard conventions for small ($w = .1$), medium ($w = .3$), and large ($w = .5$) effect sizes suggested by Cohen (1992).[3] For the hypothetical example above, if we

assume $N = 100$, w is equal to $\sqrt{10/100*1} = .32$, suggesting a medium effect size for the model difference, which is a moderate effect representing only approximately 9% of the variance. If we assume the sample size is 1,000, however, w is equal to $\sqrt{10/1000*1} = .1$, which is a small effect size representing only approximately 1% of the variance. The w approach has an advantage over the percent change in chi-square approach in that it takes into account sample size and degrees of freedom in its estimate. For tests of individual parameter differences, it is also possible to compute effect size estimates based on mean differences or variance differences (Nye & Drasgow, 2011).

A more precise method of gauging magnitude of differences in comparing two models is to use alternative fit indices, such as the Tucker–Lewis index (TLI; Tucker & Lewis, 1973), for the two models. The original conception of such fit indices was as relative or incremental fit indices that could be used to compare nested models to assess degree of difference in fit, not just to assess the overall fit of the model (i.e., comparison to the independence model with all variables uncorrelated). Small values could be considered trivial in magnitude (McGraw & Jöreskog, 1971; Tucker & Lewis, 1973). Use of alternative fit indices in this manner does not seem to be widespread among researchers, however. Lack of a single preferred index and recommendations for cutoffs for deciding upon a substantial effect size have likely contributed to the underuse of the magnitude of effect of the nested difference. To address these uncertainties, some authors (Cheung & Rensvold, 2002; Fan & Sivo, 2009) have investigated how a variety of difference-of-fit measures perform under various conditions. Results suggest that many indices have undesirable properties, such as values that were correlated with overall fit of the model, model complexity, or expected values that were not equal to 0. Cheung and Rensvold recommended three indices, but Fan and Sivo suggested that only McDonald's Centrality Index (Mc; McDonald, 1989) was minimally affected by model size when comparing means.

$$\mathrm{Mc} = \exp\left[-\frac{1}{2}\left(\frac{\chi^2 - df}{N-1}\right)\right]$$

The difference in fit between the two models, ΔMc, can then be calculated by subtracting the Mc for M0 from the Mc for M1. Note that, in contrast to the direction of difference taken for $\Delta\chi^2$, the ΔMc is computed by subtracting the more constrained model (M0) from the less constrained model (M1), as higher values of Mc indicate better fit and the less constrained model will have a value equal to or greater than that of the more constrained model. Although a cutoff for ΔMc has been suggested (ΔMc > .02 indicating a difference; Fan & Sivo, 2009), it is important to keep in mind that the objective of the ΔMc or other differences in alternative fit measures is to gauge the magnitude of effect rather than to determine statistical significance. The $\Delta\chi^2$ is generally sufficient for determining statistical significance.

Comments

These measures of magnitude of difference offer some additional information about invariance by moving beyond a simple assessment of statistical differences and, thus, may aid researchers in deciding whether significant differences are of practical importance or not. Assessing and reporting magnitude of effect of the difference is valuable, but any of the approaches discussed above should be taken as approximate values that supplement a statistical test of invariance using the likelihood ratio or other significance test. The w effect size approach is useful because the magnitude of the difference between models is put on a familiar scale used with other statistical tests. Although convenient to calculate,

w may be impacted by model size, nonnormality, or other data characteristics. The alternative fit index difference measure, such as ΔMc, provides a more precise comparison of model differences that is relatively unaffected by sample size and model size. Because the Mc measure is not available in all software programs, however, it will be less convenient for researchers if it must be computed by hand. The Comparative Fit Index (CFI; Bentler, 1990) is included in the output of more software programs but may be less suitable for comparisons of intercept and mean differences. Caution is warranted in using any of these measures when data are nonnormal or ordinal as there is little information presently about how such factors may impact their values.

Invariance Testing Strategies

The order in which sets of parameters (e.g., loadings, measurement residuals, measurement intercepts) should be compared has been the focus of many discussions of measurement invariance (e.g., Jöreskog, 1971; Taris, Bok, & Meijer, 1998; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000; Widaman & Reise, 1997). There has been a variety of opinions on the appropriate order of testing but no consensus among authors. Because these recommendations have been extensively reviewed elsewhere (see Vandenberg & Lance, 2000), I will make only a few brief observations before discussing some of the details of several specific types of invariance tests.

One commonality among authors is the recommendation that a series of omnibus tests be conducted that impose equality constraints on sets of parameters, such as comparing all loadings for a factor across groups or over time. Although the omnibus approach reduces the likelihood of Type I error because fewer tests are conducted than when comparing each individual parameter, there may be some instances in which the omnibus test is inconsistent with more specific tests (Byrne, Shavelson, & Múthen, 1989) or the researcher may have specific hypotheses about the invariance of individual or a subset of parameters.

Jöreskog (1971) suggested a sequence of invariance tests using confirmatory factor analysis that has influenced nearly all subsequent discussions of measurement invariance. The initial proposed step is a comparison of the entire variance–covariance matrix across groups. The purpose is to determine whether any further tests would be needed or the two samples should be combined for further analysis. Perhaps because this rationale does not translate as easily into longitudinal invariance tests or might be more cumbersome to conduct, most authors have not recommended a similar step when discussing longitudinal invariance (e.g., Millsap & Cham, 2011; Widaman, Ferrer, & Conger, 2010). Following a comparison of the variance–covariance matrices across groups, Jöreskog recommended a step that established that the same model is appropriate in each of the groups, usually referred to as *configural invariance*. Assuming configural invariance, he then recommended a progressive set of nested tests for specific groups of parameters, each tested while constraining the prior matrices to be equal: loadings, factor variances, and measurement residuals.

Meredith (1964, 1993) proposed a classification terminology for levels of measurement invariance, which is the most widely used. The terminology mirrors the classic testing concepts of parallel, tau-equivalent, and congeneric tests (Lord & Novick, 1968). *Weak factorial invariance* refers to the case when loadings are equal over time but intercepts, unique variances, latent means, and latent variances vary over time. *Strong factorial invariance* refers to the case when loadings and intercepts do not vary but unique variances, latent means, and latent variances vary over time.[4] *Strict factorial invariance* involves invariant loadings, intercepts, and measurement residuals. *Structural factorial invariance* involves invariant factor means, factor variances, loadings, intercepts, and measurement residuals. Figure 2.1 illustrates these concepts graphically, with grayed symbols depicting invariant parameters.

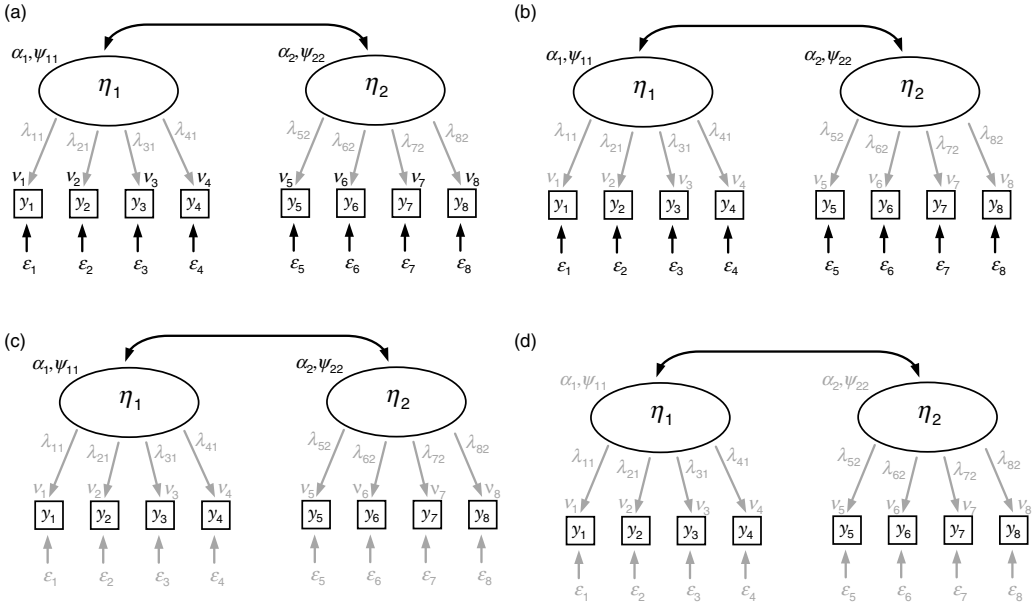

Figure 2.1 Graphic Depiction of Meredith's Factorial Invariance Definitions: (a) weak invariance; (b) strong invariance; (c) strict invariance; (d) structural invariance. Note: grayed lines and symbols represent parameters that are equal over time, where it is assumed that the equality is only between the longitudinal counterparts of each (e.g., $\lambda_{11} = \lambda_{52}$ or $\nu_1 = \nu_5$).

If any set of parameters, such as all loadings for a single factor, are not equal across groups or over time, specific invariance tests are conducted to identify which parameters (e.g., which particular loadings) are the sources of the invariance. The term *partial measurement invariance* is usually used for the circumstance in which some of the parameters of a set are invariant while others are not. With only a few exceptions (e.g., Byrne, Shavelson, &Muthén, 1989; Horn, McArdle, & Mason, 1983; Reise, Widaman, & Pugh, 1993), the topic of partial invariance tests has not been given a great deal of attention, most likely because authors have generally assumed that focused tests that contain individual parameters or subsets of parameters will be conducted by the researcher as needed to identify the sources of invariance.

Although Jöreskog's invariance testing strategy omitted intercepts and factor means, Sörbom (1974) expanded invariance testing to include mean structures. Most authors have subsequently included intercepts and means in their recommended strategy for invariance testing (e.g., Meredith, 1993; Widaman and Reise, 1997), though recommendations have varied on which stage of the process intercepts and means should be tested. Vandenberg and Lance (2000) suggested that measurement intercept invariance be tested only after testing the invariance of factor loadings. They suggested that factor variance and factor mean invariance should be tested last as part of the investigation of substantive hypothesized group differences. Millsap (2011) discusses invariance testing at length and emphasizes inclusion of mean structures at all stages. Nearly all authors suggest that invariance of measurement residuals is not required in order to proceed with substantive hypotheses about group differences in means or structural parameters. There has been little discussion of the larger issue of what researchers should do if measurement invariance assumptions are not met, a topic I will return to after discussing invariance tests for each set of parameters.

Configural Measurement Invariance

It is generally recommended that the researcher establish *configural invariance* – that the same latent variable structure exists in the multiple groups or over time – before proceeding with nested tests of parameter invariance (Horn et al., 1983; Widaman & Reise, 1997). Establishing configural invariance in the longitudinal case simply involves separate cross-sectional confirmatory factor models to establish that the measure has the same single- or multiple-factor structure at each time point (i.e., indicators load on the same factors at each wave), the model fits well at each time point, indicators load on the same factors, and loadings are all of acceptable magnitude. As with most structural model tests, careful consideration should be given to hypothesized alternative models during this stage, including examination of modification indices to investigate sources of lack of fit. When multiple factors are involved, one aspect of the factor structure that may vary over time is the covariance among factors. Because factor covariances are related to factor variances, however, tests of equality of covariances do not make sense without considering equality of factor variances also. Although configural invariance can be investigated using exploratory factor analysis (e.g., Reynolds & Harding, 1983), confirmatory factor analysis provides a more hypothesis-driven and precise testing approach in which alternative models can be compared empirically. One potential pitfall of testing separate cross-sectional models is that each analysis may be based on different sample sizes due to attrition, and this may lead to differences in power to detect incorrect models at each wave. Because general confirmatory factor analysis techniques are used to establish configural invariance, I leave details of this general process to more in-depth discussion provided by standard introductory SEM texts (e.g., Bollen, 1989; Brown, 2006; Kline, 2010; Maruyama, 1997).

Loadings and Factor Variances

Loadings

Tests of longitudinal and multigroup invariance of factor loadings, often referred to as "metric invariance" tests (e.g., Horn & McArdle, 1992; Steenkamp & Baumgartner, 1998; Thurstone, 1947), are a primary focus for most researchers. If factor loadings vary over time, it suggests that indicators have differential importance in defining the latent variable and this may lead to erroneous conclusions about change of the underlying construct. When loadings for a full set of indicators are tested for invariance, the identification approaches (i.e., referent indicator identification, factor identification, or effects coding identification) are statistically equivalent. However, because factor identification, which constrains factor variances to 1, also implies that factor variances are equal, tests conducted with factor identification will only be equal to tests conducted with referent or effects coding identification if equality constraints on factor variances are included in the test.[5]

Example 2.1: Loadings

To illustrate, I tested loadings and factor variances using three companionship items from the social exchanges data set ($N = 574$). Syntax and data sets used in the examples are available at the website for the book. All longitudinal examples in this chapter include estimated covariances between each item's repeated measurement residuals (e.g., θ_{14}, representing the covariance between ε_1 "good company" at Time 1 with ε_4 "good company" at Time 2), a point discussed below in the section "Measurement Residuals." An unconstrained model with all loadings freely estimated except for the first indicator, which had

a loading set equal to 1 for identification purposes, produced an ML chi-square value of 9.911, with 5 degrees of freedom ($p = .0778$). The CFI of .997 and Standardized Root Mean Square Residual (SRMR: Bentler, 1995) of .028 suggested that this model fit the data well overall, and all loadings were statistically significant. If only loadings were constrained to be equal, the model did not differ significantly from the unconstrained model, $\chi^2(7) = 12.077$, $p = .098$, $\Delta\chi^2(2) = 2.166$, ns, suggesting that longitudinal metric invariance was met with this scale. Note that the chi-square difference test would be unchanged if the latent variables were identified by using the effects coding approach (while constraining all loadings to be equal over time).

Factor Variances

Although some authors have suggested that tests of equality of factor variances are unnecessary or too strict, establishing that factor variances are invariant may be desirable if unequal factor variances might muddle conclusions from latent variable models of change or if the indicators from a latent variable will be used as a composite measure (I return to this point later under the section "Consequences of Noninvariance"). Longitudinal changes in factor variance may be misleading in cross-lagged panel models, for example (see Chapter 5). Tests of the equality of factor variances should not be conducted without also imposing equality constraints on all loadings for the factor. The reason is that without simultaneous constraints on factor loadings, constraining factor variances will only force the estimates of the loadings to change in compensation (Hancock et al., 2009). The dependence of factor variances on factor loading values is easily verified by inspection of Equation (1.4). Without all loadings constrained, it is not possible to evaluate factor variances independently.

An alternative identification approach is sometimes recommended with invariance testing (Reise et al., 1993). With this specification, which I will call the *single occasion identification* approach, the identifying constraint (either by setting the factor loading or a factor variance) is only imposed at the first time point, allowing the parameter to be "freely" estimated in each of the subsequent time points. As long as some equality constraints are imposed, the model may be identified. Under this identification approach, tests of factor invariance entail comparison of a model with and without constraining the factor variance at the second time point. Note that, because the single occasion identification approach with unconstrained factor variances is not identified without at least one scaling constraint on the loadings, tests of loading invariance are not possible with this identification approach. (Chapter 3 has further discussion of the single occasion identification approach).

With the referent loading identification approach, assigning a referent loading at each time point makes an implicit assumption about the invariance of the referent loadings. Adding equality constraints on the remaining loadings therefore leads to full set of constraints on the loadings. As long as a full set of loading constraints is used, it does not matter which indicator is used as a referent, because model fit is not impacted by the choice of referent. If a full set of loading constraints are not used, the test of equality of factor variances will not be the same for each choice of the referent, however (Steiger, 2002).

Example 2.2: Factor Variances

Longitudinal tests of equality of factor variances was illustrated using the three-item companionship measure from the social exchanges data set. An initial was set of tests were used to demonstrate that the choice of indicator impacts the invariance test of factor variances unless all of the factor loadings are constrained to be equal. When the first item ("good company") served as the referent but allowed the other two loadings to be freely estimated, a model constraining the factor variances to be equal over time resulted in a significantly

poorer model fit, $\chi^2(6) = 24.664$, $p < .001$, than the unconstrained model reported in Example 2.1, $\chi^2(5) = 9.991$, $p = .0778$. The significant difference, $\Delta\chi^2(1) = 14.753$, $df = 1$, $p < .001$, suggests the factor variance changed over time, although this change was modest in magnitude according to the two effect size measures, $w = .113$, $\Delta Mc = .011$.

As expected, when an alternative referent indicator ("recreational activities") was used, there was no change in the fit of the unconstrained model, $\chi^2(5) = 9.991$. The model specifying equal factor variances, however, had a different chi-square value, $\chi^2(6) = 17.595$. Although the decrement in fit was still significant, $\Delta\chi^2\ (1) = 7.684$, the likelihood ratio value was nearly half that obtained for the model with "good company" as the referent. This illustrates how the factor invariance hypotheses are not the same for the two identification approaches, because the latent variable variances are based on different observed variables. The factor variances at the two time points were $\psi_{11} = .834$ and $\psi_{22} = .488$ when "good company" was the referent yet $\psi_{11} = 1.115$ and $\psi_{22} = .804$ when "recreational activities" was the referent. Although the two comparisons had the same statistical conclusion in this example, findings may differ more substantially or may even lead to contradictory statistical conclusions in other examples. Likelihood ratio tests using effects coding identification (not illustrated) would differ from both results, because the factor variance is based on a weighted average of the three indicators. None of the identification approaches is correct or incorrect necessarily, but each approach represents a different definition of the latent variable variance and implies a different invariance hypothesis.

These examples illustrate that equality tests of factor variances should only be conducted when all factor loadings also are constrained to be equal over time. When all non-referent loadings are set equal in the constrained model, the chi-square is the same regardless of the referent. To demonstrate, longitudinal invariance of the companionship factor was tested again, this time by comparing a model with the factor variances constrained to be equal to a model with non-referent loadings constrained to be equal. The chi-square difference was the same regardless of which item was used as the referent, $\chi^2(8) = 37.553$, $p < .001$. This model differed significantly from the less restricted model, $\chi^2(7) = 12.077$, $p < .001$, $\Delta\chi^2(1) = 25.476$, $df = 1$, $p < .001$, suggesting that factor variances differed across the two time points. The w effect size measure was .211, a difference that was small to moderate in magnitude, and the ΔMc index was equal to .021, a value just over the cutoff suggested by Fan and Sivo (2009).

Specific Loading Tests

Metric invariance tests that focus on only a subset of loadings are complicated by the interdependence of factor loadings and factor variances. This issue, sometimes termed the *standardization problem* (Cheung & Rensvold, 1999; Johnson, Meade, & DuVernet, 2009), makes it challenging to identify the source of noninvariance when the omnibus test of loadings is rejected. When only a subset of loadings are tested for invariance, the choice of referent, or how the factor is "standardized," may lead to erroneous conclusions about which specific loadings differ. As shown in Equation (1.7), any two loadings of the same factor are proportionate to one another, and, as a consequence, a test of invariance of individual loadings will also involve a test of the invariance of loading proportions.

Consider a repeated measurement example shown in Figure 2.2 with two latent variables each consisting of four indicators. With factor variances set equal to 1, as with the factor identification approach, it follows from Equation (1.5) that a test of whether the second loading is invariant over time (i.e., $\lambda_{21} = \lambda_{42}$) is equivalent to a test that $\lambda'_{21}\sqrt{\psi'_{11}} = \lambda'_{42}\sqrt{\psi'_{22}}$ if the referent approach to identification was used (refer to Chapter 1 for discussion of the relationship between the two identification approaches). In essence, a test of the invariance of two loadings is not fully distinguishable from a test of invariance

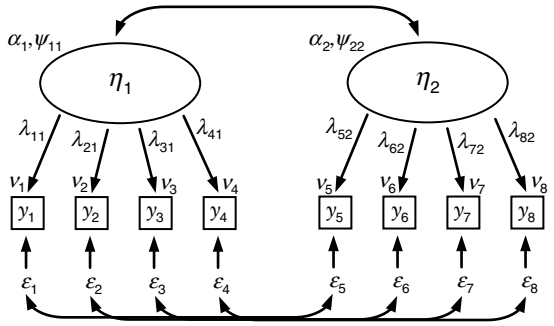

Figure 2.2. Two-Wave Four-Indicator Measurement Model.

of the two factor variances. Furthermore, Equation (1.7) implies that a test of the invariance of two non-referent loadings, such as $\lambda'_{21} = \lambda'_{42}$, under referent loading identification is not distinguishable from a test of the invariance of two referent loadings either.

$$\frac{\lambda'_{21}}{\lambda'_{11}} = \frac{\lambda'_{42}}{\lambda'_{32}}$$

Similar principles hold for effects coding identification. Even though the effects coding identification approach does not tie the factor variance to a specific loading, the factor variance is a linear combination of all of the loadings, resulting in the inability to distinguish loading ratios or loading-factor ratios.

The single occasion identification approach does not solve the standardization problem either. In the absence of equality constraints, it is unlikely that the model will be identified, making any comparisons with an unconstrained model as a baseline impossible. Setting a single identification constraint does not solve the standardization problem in any event (Raykov, Marcoulides, & Li, 2012), because loadings for the second factor are scaled to loadings for the first factor whenever longitudinal equality constraints on loadings or factor variances are specified.[6]

The dilemma presented by the standardization problem has provoked considerable discussion and a number of suggested solutions. Cheung and Rensvold (1999) proposed a series of tests they called *factor-ratio tests*. The strategy involves invariance tests of each of the $J(J-1)/2$ possible pairs of loadings, alternating through each possible indicator used as a referent. Yoon and Millsap (2007) also proposed a reasonable strategy that involves a post hoc search using modification indices. Cheung and Lau (2012) more recently proposed a simultaneous test of the invariance of all factor-ratios and suggested a bias-corrected bootstrap approach to significance, but this strategy requires software that can implement complex equality constraints and bootstrapping. Any of these approaches entail multiple statistical tests and may be subject to familywise error problems (Kim & Yoon, 2011). Simulation work by Stark and colleagues (Stark, Chernyshenko, & Drasgow, 2006) indicates that, with low sample sizes, a standard Bonferroni correction may lack statistical power, so a more powerful adjustment, such as the Sidák–Bonferroni (Sidák, 1967) or another adjustment (see Olejnik, Li, Supattathum, & Huberty, 1997 for a review) may be preferable if post hoc adjustments are used.

Example 2.3: Tests of Specific Factor Loadings

The invariance test of all loadings indicated that loadings were equal over time, and there would ordinarily be no imperative for further invariance tests. To illustrate the loading

ratio test proposed by Cheung and Lau (2012), however, I conducted a simultaneous test of longitudinal invariance of the full set of factor loading ratios, $\lambda_{52}/\lambda_{42} - \lambda_{21}/\lambda_{11} = 0$, $\lambda_{62}/\lambda_{42} - \lambda_{31}/\lambda_{11} = 0$, and $\lambda_{62}/\lambda_{52} - \lambda_{31}/\lambda_{21} = 0$. The test was conducted by defining a new parameter for each loading ratio and then testing each for significance using bias-corrected bootstrap confidence interval estimates. The results indicate that none of the loading-ratio differences were significant (all of the confidence intervals contained zero), .014 (–.139,.200), .128 (–.061,.313), .098 (–.056,.255).

Binary and Ordinal Indicators

Loading invariance tests conducted when indicators are binary or ordinal proceed in a similar fashion as when indicators are continuous. Nested tests can be conducted to compare a model with equality constraints placed on loadings over time to a model with no equality constraints imposed, assuming appropriate adjustments are made for robust ML and WLSMV chi-square difference tests (Asparouhov & Muthén, 2006). Delta or theta parameterizations with WLSMV lead to models with identical fit and tests of invariance for loadings and factor variances.

Proposed strategies for testing invariance with binary and ordinal indicators within the IRT framework differ somewhat from the approach used with continuous variables, because the focus is usually on evaluating tests for item difficulty and bias. The parameters of interest include *a*, the discrimination parameter and *b*, the difficulty parameter, the interpretation of which depends of the estimation approach. If WLSMV estimation with theta parameterization is used, the *a* parameter corresponds to item loadings and the *b* parameter corresponds to the intercept estimates (Kamata & Bauer, 2008). Other binary or ordinal estimation procedures require some transformation of the parameters for appropriate IRT values. The factor-ratio problem still applies with binary and ordinal indicators, but the approach to correctly identifying the invariant referent variable has differed with IRT applications compared with confirmatory factor analysis applications. The IRT approach has often involved multiple referents (or "anchors") with loadings for all but one indicator set to 1 (known as the "constrained baseline" or "all other" method), whereas the CFA approach has focused on alternating a single referent loading set to 1 across items (e.g., Kim & Yoon, 2011; Woods, 2009). The rationale for multiple referents is that the approach is less reliant on a single item that may not be invariant. Simulation work by Stark and colleagues (2006) suggests that use of the single referent may be better at correctly identifying invariant items to use as the identification referent, partly because the multiple referent approach produces difference tests based on an incorrect model in most instances. A variation on the multiple referent approach has recently been proposed (Lopez Rivas, Stark, & Chernyshenko, 2009) and has been shown to work well for identifying invariant referent items (Meade & Wright, 2012). It also may be possible to adopt the method proposed by Cheung and Lau (2012) using simultaneous multiple factor-ratio constraints applied in the context of either ML or WLSMV estimation for binary or ordinal variables, but this approach has not been investigated to date.

Example 2.4: Loading and Factor Invariance with Binary Indicators

As an example of factorial invariance tests with binary indicators, I explored whether the loadings and the factor variance for the unwanted advice latent variable from the social exchanges data set remained constant over a six-month period. WLSMV estimation was used with theta parameterization with referent identification that set the first loading equal to 1. A model with no equality constraints had a good fit to the data, $\chi^2(5) = 5.907$,

$p = .315$, CFI = .997. When longitudinal equality constraints were imposed on loadings and factor variances, the model fit did not significantly decrease, $\chi^2(8) = 9.390$, $p = .311$, $p < .001$, with $\Delta\chi^2(3) = 3.760$, $p < .289$, according to the weighted difference test. The results are consistent with the hypothesis that both factor variances and factor loadings were invariant. If the omnibus test of loading invariance had been significant, a simultaneous test of the equality of the loading ratios could have been conducted to investigate which specific loadings differed.

Including Means in Invariance Tests of Loadings and Factor Variance

Finally, a number of authors recommend inclusion of mean estimates when testing for invariance of loadings or factor variances, but inclusion of mean structures does not affect the invariance tests of loadings or variances provided that no equality constraints are imposed on the intercepts or factor means. As long as the mean structure is identified using any of the aforementioned standard scaling approaches, model fit and degrees of freedom are unaffected by the inclusion of means in the estimation. If the researcher is not interested in mean invariance, then inclusion of means in other invariance tests is not necessary. The converse is not the case, however. Intercept or factor means are impacted by the value of factor loadings, so it does not make sense to investigate mean invariance without considering the context of loading invariance.

Measurement Intercepts and Factor Means

Many longitudinal structural equation models will involve mean structures, and, for this reason, researchers may be interested in testing longitudinal invariance of intercepts or factor means. Measurement intercepts, rather than factor means, are usually the primary focus of invariance testing, because the usual goal is to establish that measurement properties do not change over time prior to investigating substantive hypotheses about changes or differences in the mean of the construct. Intercepts are partly a function of factor loadings, and this logically implies that tests of measurement intercept invariance will not be appropriate unless loadings are held invariant in the model. More generally, overall model fit is more a function of parameters related to the variance–covariance matrix than parameters related to the mean structure. As the size of the model increases, say with more observed variables, the number of elements in the observed variance–covariance matrix increases rapidly and the number of observed means only increases linearly. The consequence is that the overall model fit is more affected by any misspecification of factor variances, loadings, and measurement residuals than by misspecification of the mean structure (Fan & Sivo, 2009). This imbalance in the contribution of the variance–covariance elements and the mean structure elements to the overall chi-square may have an impact on any magnitude of effect estimates.

Measurement Intercepts

Tests of intercept invariance can be conducted with any of the three identification approaches are algebraically equivalent results, but it should be kept in mind that the factor mean identification approach, in which factor means are set to 0, implies equivalent factor means. In the case of the referent intercept identification approach, the choice of referent is arbitrary and results will be identical as long as all intercepts for a factor are tested together (assuming invariant loadings). Regardless of the identification approach to intercepts, it is important that the approach used for identifying the factor variance

(i.e., constraints on loadings or factor variance) match. In the case of referent loading approach, for instance, the same referent indicator should be used for the intercept and the loading to improve interpretability.

Example 2.5: Measurement Intercepts

To illustrate a test of measurement intercept invariance, I again used the social exchange companionship items. A model with loadings constrained equal but unconstrained mean structure served as the baseline comparison model, M0. This model, which used a referent intercept identification approach with $\nu_1 = 0$, had the same fit as the constrained model without mean structure, $\chi^2(7) = 12.077$, $p = .098$. If the effects coding approach was used for identification instead, the fit would be the same. A model with equality constraints on the intercepts for the second and third indicators, did not have a significantly different fit, $\chi^2(9) = 12.399$, $p = .192$, $\Delta\chi^2(2) = .322$, ns, supporting invariance of the measurement intercepts.

Factor Means

Comparisons of factor means belongs less in the category of measurement hypothesis tests than in the category of substantive hypothesis tests, a focus of ANOVA models discussed in Chapter 3. Nonetheless, the invariance test for factor means deserves some brief mention here. Because factor means are a function of measurement intercepts (or, rather, the observed means of the indicators), tests of factor means are not independent of intercept tests. Recall also that under the referent intercept identification approach the factor mean is equal to the observed mean of the referent indicator (Equation [1.10]). Thus the factor mean invariance test is really a test of the equality of the observed means for the referent indicator variables at each time point.

Example 2.6: Factor Means

To demonstrate, factor means at the two time points were compared for the companionship factor. A model with invariant factor means for the companionship factor, constraining the measurement intercepts to be equal over time, had a chi-square value of 14.439 and degrees of freedom equal to 10. This value was not significantly larger than the model constraining just the intercepts and freely estimating the factor means, $\Delta\chi^2(1) = 2.040$, ns. Notice that, based on Equation (1.10), the factor mean comparison would be equivalent to a test of the difference in the mean of the first item. The connection of this analysis to a repeated measures t-test or ANOVA will be discussed in greater detail in Chapter 3.

Specific Measurement Intercept Tests

As with loading and factor variance estimates, measurement intercepts can be expressed as a function of other measurement intercepts, loadings, and the factor mean (Equations [1.9] and [1.13]). Factor means are a direct function of observed indicator means, with no quantities estimated (e.g., loadings, measurement residuals), so equality constraints on a subset of measurement intercepts will force changes in the values of any unconstrained intercepts or the factor mean. Invariance tests for a subset of measurement intercepts therefore cannot easily isolate particular intercepts that are unequal. With the referent intercept approach to identification, there is no way to know if the referent itself is

invariant. Although the effects coding identification approach does not require a referent, a similar compensatory phenomenon also will obscure which indicator is invariant. Because the effects coding approach defines the factor means as a weighted function of the observed means, equality constraints on a subset of indicators would alter the values of the factor means and the unconstrained intercepts. Scaling constraints at only one time point are possible, but the approach does not avoid the dependency issue (Raykov et al., 2012), and, under this identification approach, an unconstrained model cannot serve as the baseline comparison because it will not be identified. The solutions for identifying particular non-invariant intercepts therefore follow the strategies recommended for the factor-variance dependence problem (e.g., Cheung & Lau, 2012; Lopez Rivas et al., 2009; Yoon & Millsap, 2007).

Example 2.7: Specific Measurement Intercepts

The omnibus test of measurement intercepts in the companionship example suggested that all intercepts were invariant, and, in practice, no further testing would be necessary. Cheung and Lau's multiple constraint method (Cheung & Lau, 2012) was demonstrated, however, simply for didactic reasons. Results from the analysis with all possible ratio constraints indicated that the model fit the data well, $\chi^2(6) = 14.138$, $p = .028$, CFI = .995, SRMR = .053. The bias-corrected bootstrap confidence limits suggested that one of the intercepts, τ_6 vs. τ_3 ("recreational activities"), was significant because zero was not included in the interval: –.150 (–.537, .173), –.398 (–.769, –.033), –.233 (–.596, .118). The significant result is inconsistent with the omnibus test in Example 2.5 that suggested none of the intercepts were significantly different, but the 95% upper limit (–.033) was near 0 and the p-value for the significance test was near the .05 alpha level ($p = .041$). Overall, this series of tests suggests that the measurement intercepts did not differ substantially and that the factor means did differ between the two time points.

Binary and Ordinal Variables

In the case of binary or ordinal variables, tests of intercept invariance involve constraints on thresholds, and, thus, a test of the longitudinal invariance of proportions. Identification of the model requires that at least one threshold per factor be set equal to 0, that the factor mean be set equal to 0, or that effects coding constraints be used. The single occasion identification approach is also possible, with identifying constraints at only one time point as long as thresholds are held equal over time. If there are more than two response categories, there will be $C - 1$ constraints that can be placed on thresholds for all non-referent indicators. For the referent indicator, one threshold is set equal to 0, but the remaining thresholds for the referent indicator may be tested for invariance.

The link between the observed variable y and the theoretical, underlying continuous variable y^* represents an arbitrary scaling determination that is not relevant for continuous variables. With ordinal variables, linking observed y scores to an underlying y^* requires an assignment of thresholds to cutpoints on the logistic or the normal distribution that determines the metric of the y^* variable. For a single, cross-sectional observation, this determination can be arbitrarily assigned by the observed proportions without any particular harm. With longitudinal data, however, arbitrarily or freely estimating the thresholds runs the risk of using a different underlying scaling for the y^* distribution for each time point. This could be likened to using the Fahrenheit temperature scale for one time point and the Celsius temperature scale for another time point to assess change in temperature over time. A one-unit increase in the Fahrenheit scale is not equivalent to a

one-unit increase in Celsius. If the distance between points on the scale differs over time, any assessments of change will be illogical.

In the context of ordinal variable categories, we need to make sure that each 1% difference in proportions between category 1 and category 2 is represented by the same units of standard deviation in their underlying y^* distributions at both time points. If the scaling metric is consistent over time, then the ratio of the difference between the thresholds for any two categories, designated below by the subscripts c and c', of the same repeated measurement should be equal to the ratio of the standard deviations of the two y* deviations, taking into account sampling variability (Mehta, Flay, & Neale, 2004).

$$\frac{\tau_{1c'} - \tau_{1c}}{\tau_{2c'} - \tau_{2c}} = \frac{sd_{y_1*}}{sd_{y_2*}}$$

In this equation, subscripts 1 and 2 represent two repeated observations of the same ordinal variable, say at Time 1 and Time 2. If the scaling is not comparable over time, this equality will not hold. Unfortunately, with only two or three categories, the assumption of comparable scaling cannot be tested and we must proceed as if the underlying scales are comparable. With three or more categories, this assumption can be tested for a longitudinal design (see Mehta et al., 2004 for further details).

Example 2.8: Thresholds for Binary Indicators

Thresholds and factor means for the binary model of unwanted advice were compared in a series of nested tests to illustrate threshold comparisons using WLSMV estimation with theta parameterization. The model using the referent intercept identification approach with equal loadings, equal variances, freely estimated measurement intercepts, and freely estimated factor means served as the baseline comparison model, $\chi^2(8) = 9.930$, $p = .311$. Note that the model fit and degrees of freedom are identical to the same model tested with no mean structure. A model with equality constraints on the loadings and adding equality constraints on the intercepts had a fit that was marginally poorer than the baseline model, however, $\chi^2(9) = 10.604$, $p = .357$, $\Delta\chi^2(2) = 1.170$, ns, based on the weighted difference test. For the purposes of illustration, threshold ratio tests were used to compare specific thresholds, but, because the omnibus test suggested no significant differences, this test would be unnecessary in practice. Indeed, none of the thresholds was significant, as the 95% bootstrap confidence interval included zero in each case, .257 (–.374, 14.236), .191(–.268, .772), .047 (–.461, .528). Thus, in the unwanted advice model, it appears that intercepts were constant across the two waves.

Measurement Residuals

Tests of loading or intercept invariance may be followed by an investigation of whether measurement residuals change. These tests would nearly always proceed while holding loading and factor variances equal over time. Many authors suggest that it is not critical to establish longitudinal invariance for measurement residuals, because use of latent variables in the analysis partitions out measurement residual variance. Inequality of residuals over time would therefore have no impact on the equality of the loadings or the factor variances over time as their values are a remainder function of the loadings and factor variances.

If a composite measure derived from the items in the latent variable will be used in subsequent analyses, however, changes in reliability or scale variance could impact inferences about change in some circumstances (Raykov, 2001). Changes in scale variances over time may have consequences for standard error estimates, significance tests, and power where a composite measure of the indicators is used (Raju, Lafitte, & Byrne, 2002). The potential impact of changes in reliability, measurement variance, or standard errors on specific longitudinal analyses using composite measures will become more apparent in later chapters. Not only do composite scores incorporate measurement error into the variance of the score, but they also are typically based on the assumption that their contributions to the composite score are all equal.

The identification approach does not affect invariance tests of measurement residuals. Model fit and parameter estimates for each measurement residual will be equivalent whether factor variances are set to 1, a referent loading is set to 1, or effects coding constraints are used. Moreover, as with tests of equality of loadings and factor variances, inclusion of mean structures will have no impact on tests of equality of measurement residuals over time. Imposing equality constraints on measurement residuals may be problematic, however, if true differences exist. Equality constraints on residuals may affect loadings and factor variances, depending on the other equality constraints in the model and the true parameter values.

When indicators are binary or ordinal, equality constraints on measurement residuals cannot be imposed with all estimation methods. For ML and WLSMV with delta parameterization, the identification of the y^* distribution places restrictions on the estimation of the measurement residuals and constraints cannot be used with these parameters (Muthén & Asparouhov, 2002). Constraints on measurement residuals can only be made using WLSMV estimation with theta parameterization, although Muthén and Asparouhov (2002) caution that interpretation may be problematic at least in some circumstances. This estimator also requires that nested chi-square tests use the appropriate weighting approach (Asparouhov & Muthén, 2006).

Example 2.9: Measurement Residuals

A likelihood ratio test of measurement residual invariance was conducted for the companionship items. The less restricted comparison model was the model in which factor variances and loadings were set equal over time but measurement residuals were freely estimated, $\chi^2(8) = 37.553$, $p < .001$, CFI = .983, SRMR = .117. The invariance model set all residual variances for the repeated measurement of each item equal over time ($\theta_{11} = \theta_{44}$, $\theta_{22} = \theta_{55,}$ $\theta_{33} = \theta_{66}$) in addition to constraining the loadings and the factor variances in the model. Covariances among each pair of residuals were estimated. The result was a poorer fit to the data than the comparison model, $\chi^2(11) = 78.779$, $p < .001$, CFI = .961, SRMR = .097, $\Delta\chi^2(3) = 41.226$, $p < .001$, although the magnitude of the difference was small to moderate, $w = .155$ and $\Delta Mc = .032$. This result suggests that the variance and the reliability of the measure differ at Wave 1 and Wave 2, and differences may be large enough to introduce biases in analyses using a composite index of the items. Use of latent variables, however, could allow measurement residuals to differ across time and would avoid these biases.

Correlated Measurement Residuals

A topic often neglected because of the usual focus on multigroup analysis is the role of correlated measurement residuals (sometimes "correlated errors" or "autocorrelated

residuals") in invariance testing. Allowing covariances among measurement residuals for indicators that are repeated over time (e.g., including $\text{Cov}(\varepsilon_1, \varepsilon_5) = \theta_{15}$ and $\text{Cov}(\varepsilon_2, \varepsilon_6) = \theta_{26}$ in the model shown in Figure 2.2) is common in longitudinal structural equation modeling, however. Measurement residual covariances represent any remaining association among items once the covariance of the factors has been taken into account, whether the factor covariances are modeled through unidirectional predictive paths or bidirectional correlations. In other words, covariances among measurement residuals involve longitudinal associations between unique sources of variance for each indicator that are not due to variance from the common factor. The covariation among measurement residuals concerns systematic variance not measurement error, as measurement error by definition is random and cannot be correlated with other phenomena. Covariation among measurement residuals may be a function of methodological artifacts, such as the fact that the same social desirability for the item affects responses at each time point, or substantive sources of variance, such as consistent responses to the sleep item on a depression measure due to an ongoing health condition. Specifying correlated measurement residuals in a model may have some impacts (although often minor) on loadings, factor variances, or measurement residual estimates. Given that longitudinal correlations among measurement residuals are likely in most cases, not including them yields an incorrect model. It is therefore advisable to test all aspects of measurement invariance while the theoretically expected covariances among measurement residuals are estimated.

Multiple Time Points

Although the examples discussed in this chapter are based on measurements at two time points, the issues related to invariance testing extend readily to more than two time points. It is important to examine invariance across all time points, and nested tests can be used to test equality constraints across more than two waves. Establishing measurement invariance at two waves does not necessarily indicate that measurements will be invariant across all waves of a multiple-wave design. With more time points, the likelihood test of whether parameters differ over time will typically be larger than it would be if there were fewer time points, because there are more opportunities for inconsistency in measurement procedures or other idiosyncratic events to occur. It is also acceptable to compare pairs of time points, but this increases the number of tests necessary to investigate invariance across all waves with potential implications for familywise error.

Second-Order Factor Models

Second-order factor models add another layer of complexity to invariance tests. Similar principles hold for tests of measurement invariance for these models as it does for tests of first-order models, but tests also are necessary for the second level of the hierarchical structure. The initial step should be to establish configural invariance over time to make sure that the second-order model is appropriate at each time point. First-order parameters can then be tested in the same manner as we have considered above, either by testing each subfactor separately or together in the hypothesized second-order model. The most important contrast with tests of first-order models is that second-order factor loadings, their intercepts, and the residuals/disturbances associated with each subfactor must be considered. Invariance of second-order factor loadings suggests that the association between a first-order and the superordinate second-order factor remains constant over time, but second-order loadings will have the same interdependence and same dependence on the variance of the second-order factor, a situation parallel to the interdependence for the simpler first-order

factor model. Residuals associated with the first-order latent variables can also be tested for invariance but may be considered of lesser importance if the second-order latent variable will be used in subsequent analyses. Because other authors consider second-order models at length (e.g., Chen, Sousa, & West, 2005; Ferrar, Balluerka, & Widaman, 2008; Sayer & Cumsille, 2001) and because Chapter 7 will consider second-order models in the context of growth curve models, I will not discuss them further here.

Consequences of Noninvariance

Frequently omitted from the ongoing dialogue about statistical details and appropriate strategies related to invariance testing is any in-depth discussion of the consequences for data analyses when measurement properties are not invariant (cf. Millsap & Kwok, 2004; Schmitt, Golubovich, & Leong, 2011). It is indisputable that, with traditional analyses such as ANOVA and regression analysis, varying measurement properties may lead to ambiguous or misleading conclusions about change. Because structural equation models analyze constructs differently, by partitioning variance due to the common factor from specific variance, latent variable models should be better equipped to cope with lack of measurement invariance. The primary focus of most discussions has been on what to do when there is evidence of only partial measurement invariance (e.g., Byrne et al., 1989). The most common practice is to identify the particular items that may be the source of noninvariance and eliminate them.

Configural Invariance

Where configural invariance is not met (i.e., the same model structure does not apply over time), the researcher may need to reconsider the factor structure, the indicators, or the theoretical constructs. Theory may need to be consulted to understand whether a change in factor structure over time would be expected given the context. In educational research, for instance, mathematical ability (e.g., reasoning vs. problem solving) may be expected to be differentiated only as development of these skills occurs over time. Edwards and Wirth (2009, 2012) explore methodology for modeling change when configural invariance is not expected theoretically.

Weak and Strong Invariance

Weak and strong invariance involve equality in loadings and measurement intercepts. For these types of invariance, it is helpful to appreciate that specific violations of invariance assumptions have specific consequences. Based on the understanding of parameter estimates and invariance tests developed in this chapter thus far, we can identify some likely consequences when measurement invariance assumptions are violated. When intercepts are not invariant, we can expect more serious impacts on factor means than on factor variances. When loadings are not invariant, we can expect factor variances and factor means to be impacted, because intercepts are weighted by their respective loadings in calculating the factor means.

Imposing equality constraints when parameters are not invariant runs the risk of introducing biases. Factor variances are partly a function of loadings, and longitudinal equality constraints on the loadings lead to the same relative weighting of the observed indicators over time when calculating the factor variances. If the true relationship between the indicator and the factor differs over time but equality constraints on the loadings are imposed, changes in just one observed indicator may be responsible for changes in the

factor variance, potentially impacting conclusions drawn from the model. Similar reasoning can be applied to the effect of changes in measurement intercepts on factor means when equality constraints are incorrectly imposed.

The consequences of incorrect equality constraints for parameter estimates will depend on the factor identification approach. Under referent identification, the choice of referent for the intercept or the loading is critical, because the factor variance and mean are scaled by the observed indicator used as the referent. If the mean and the variance of the observed variable used for the referent do not change, then the mean and variance of the latent variable will not change. This reasoning suggests that equality constraints will have little impact on the mean and factor variance, even if incorrectly imposed, as long as the referent variable is invariant.

This general point can be illustrated by a small simulation, the results of which are reported in Table 2.1. The model involved a four-indicator factor measured at two time points (as illustrated in Figure 2.2). In the table, the rows are grouped by four conditions in which values were varied for one item: (a) the value of the item mean, the item variance, and the factor loading did not change over time; (b) the observed indicator mean changed over time; (c) the observed indicator variance changed over time; (d) the loading (magnitude of the association between the indicator and the factor) changed over time. Changes in each value were large in magnitude, with the value at the second time point 60–80% larger than the first time point.

Results in the leftmost column in Table 2.1 illustrate that when no equality constraints were imposed, the parameter estimates of the factor means and variances were unaffected by any of these manipulations. As seen in the second column, when the altered values did not involve a referent indicator, changes in the item loading or the item variance had small effects on the factor variance or the factor mean. Changes in the indicator mean appeared to impact the factor mean and factor variance more, but these biases were relatively small overall (approximately 2.6% on average). In contrast, when the referent indicator value changed over time, more important differences were found in the mean and variance estimates for any of the alterations.

The implication of this simulation is that there appears to be greater drawbacks to imposing equality constraints than not imposing equality constraints under these conditions. The difficulty in practice is in determining that the parameters for the referent item are indeed invariant. The simulation only involved a limited set of conditions. I only altered one parameter at a time, for instance, and if several non-referent items are involved, biases could be larger (Byrne et al., 1989; Schmitt et al., 2011). For cases that do not involve a high proportion of varying parameters, however, the simulation results suggest that statistical tests involving the mean or the variance of the latent variable will be little affected by measurement invariance when loadings and intercepts are allowed to vary over time.

These results only apply to the referent identification approach, which links factor means and variances to a particular indicator. The consequences of incorrectly imposing equality constraints can be expected to be more substantial for other identification approaches, mirroring more what was seen in the simulation condition that altered values for the referent indicator. Factor identification (either single- or multiple-occasion) and effects coding identification approaches will distribute longitudinal changes in any one of the observed variables to loadings or intercept estimates for the other indicator items (as evidenced by Equations [1.7] and [1.13]). For the identification approaches not based on a single referent variable, factor variances and factor means are defined by all of the indicators. As a result, mean, variance, and loading changes in a subset of indicators may be more likely to bias effects on results in longitudinal models than when referent identification is used.

Table 2.1. Simulation Results Demonstrating Effects of Noninvariance on Latent Means and Variances.

Population values	*No equality constraints*	*Equality constraints*	
		Non-referent item unequal	*Referent item unequal*
Item mean, item variance, loading equal			
χ^2	167.389		
df	19		
α_1	3.006		
α_2	2.999		
ψ_{11}	2.082		
ψ_{22}	2.278		
Item mean changes			
χ^2	167.389	201.188	191.122
df	19	22	22
α_1	3.006	2.995	4.042
α_2	2.999	3.069	4.093
ψ_{11}	2.082	2.143	2.364
ψ_{22}	2.278	1.625	1.947
Item variance changes			
χ^2	147.128	149.969	149.969
df	19	22	22
α_1	3.006	3.006	3.081
α_2	2.999	2.999	3.074
ψ_{11}	2.099	2.188	2.281
ψ_{22}	2.298	2.192	2.297
Loading changes			
χ^2	176.442	193.061	193.061
df	19	22	22
α_1	3.006	3.006	2.594
α_2	2.999	2.999	2.588
ψ_{11}	2.089	2.325	1.728
ψ_{22}	2.296	1.993	1.473

Note: $N = 250$ with 500 replications for each cell. A single item was altered for each condition. Means were altered from 3.0 to 5.0, variances were altered from 4 to 8. Within-factor covariances were modified from 2.25 to 1.25, resulting in a change of the standardized loading from approximately .750 to approximately .400. Constraints were imposed on the intercepts but not the loadings when the item mean was altered, and constraints were imposed on the loadings but not the intercepts when the item variance or loading was altered. The autocorrelation between η_1 and η_2 for the unconstrained and unaltered model was approximately .400. Average SRMR for the unconstrained model was .067. No correlated uniquenesses were estimated although correlations among items over time were included in the covariance input.

Strict Invariance

Strict invariance involves equal measurement residual variance over time. Most authors seem to agree that the question of whether measurement residuals are invariant is not all that central. This usual tolerance for variability in measurement residuals over time is predicated on a presumption that a latent variable be used in subsequent analyses. But when residuals change over time, there will be changes in the variance of the scale score and its reliability, either of which may be problematic whenever the measure is used as a composite score. Although strict invariance may not be used when latent variables are analyzed, items used in a composite measure need to meet the standard of strict invariance to avoid mistaken conclusions due to changes in reliability.

Comments

What starts as a simple goal – to establish support for longitudinal measurement invariance so that substantive hypotheses can be tested with confidence that conclusions will not be contaminated by measurement artifact – turns out to be a much more complicated endeavor than it would seem at first blush. This complexity makes it difficult to make any simple recommendations about invariance testing, but a few thoughts based on the explorations in this chapter may provide some useful guidance.

Recommended strategies for the order of omnibus tests of invariance, such as tests of factor variance, all factor loadings, or all measurement residuals, have varied considerably. The identification problem for factor variances makes it difficult to isolate loadings and factor variances unless longitudinal invariance of all loadings can be established first. Although invariance of factor variances is often considered unnecessary, their integral role in many longitudinal analyses suggests that, in many instances, invariance tests of factor variance should follow invariance tests of loadings, if not just for the sake of more completely understanding the data. It is only sensible to test the equality of factor variances after loadings have been established as invariant, and, similarly, it is only sensible to test for the equality of factor means after intercepts have been established as invariant.

Identification of specific sources of noninvariance of loadings or intercepts requires special care because the interdependence of common factor loadings, factor variances, and intercepts requires an invariant referent. Both the referent and the factor identification approaches lead to implicit invariance assumptions that may obscure the true sources of invariance unless caution is used (i.e., the "standardization" problem). The best solution is to take advantage of proposed remedies that account for ratios of factor loadings or intercepts to accurately identify the referent indicator (e.g., Cheung & Lau, 2012; Lopez Rivas et al., 2009; Yoon & Millsap, 2007).

If an invariant referent indicator can be identified, latent variable models may be fairly robust to changing measurement properties if loadings and intercepts are allowed to vary over time. Longitudinal models that depend heavily on factor variances (cross-lagged panel models, autoregressive models, mediational models) or latent means (e.g., latent variable ANOVA, growth curve, or latent difference scores) may be robust to the extent that the parameters for the referent variable are not invariant. Although this robustness is a potential advantage of longitudinal structural equation models, it comes at the cost that the latent variable is scaled only to one item. The alternative is usually the deletion of items that are not invariant, which may sacrifice some elements of construct validity that may not have to be abandoned.

An issue too infrequently considered is the degree of noninvariance that is of practical importance. Sample sizes do not have to be that large to have sufficient power to detect rather trivial departures from invariance. If the magnitude of noninvariance is small, there may be relatively minor biases to subsequent analyses even if equality constraints are incorrectly imposed. Supplementing likelihood ratio tests with effect size estimates can help address this uncertainty, but this is seldom seen in practice. More detailed simulation work could serve as a guide to conventions regarding the magnitude of invariance violations that will introduce a tolerable level of bias.

The type of measurement invariance required depends on the specific analyses that will be used to investigate substantive hypotheses. Strict or structural invariance that include equal measurement residuals may be required if indicators are to be combined into composite scale scores (e.g., summing or averaging items) for subsequent analyses. Even strict invariance, in which loadings, intercepts, and measurement residuals are equal over time

but factor variances and factor means are not, may not be adequate for some types of analyses. Although changes in factor means would rarely be considered measurement artifact, changes in factor variances can affect results from any model with lagged effects such as cross-lagged panel models or any models that incorporate autoregressive paths. It can be argued that changes in the factor variance would presumably reflect changes in true score variance, which does not involve measurement artifact, but we should be mindful of the fact that changes in the latent variable variance over time will affect correlations and regression estimates. Changes in factor variances over time may be less problematic if substantive hypotheses are focused on mean changes, as is the case when latent growth curve models, latent difference models, or latent variable ANOVA models are involved. As we shall soon see, many longitudinal models may be a complex mélange of mean and variance information that make neglecting issues of measurement invariance risky business. Even if definitive answers to questions about measurement invariance cannot be found, the process of investigating measurement invariance can provide valuable information about the characteristics of the measures and the data.

Recommended readings

Given that invariance testing is a convoluted and often controversial subject, I recommend that the reader consider other viewpoints on invariance. One provocative paper is an insightful critique of measurement invariance tests by Hancock and colleagues (Hancock, Stapleton, & Arnold-Berkovitz, 2009). The papers by Meredith (1993) and Vandenberg and Lance (2000) are good general overviews of many of the concepts, definitions, and practical issues. Although these and other writings have focused almost exclusively on multigroup invariance comparisons, many issues are applicable to the longitudinal case. The chapters by Bontempo and colleagues (Bontempo, Grouzet, & Hofer, 2012) and Millsap and Champ (2011) review longitudinal invariance issues specifically. The text by Millsap (2011) is perhaps the most extensive general treatment on invariance issues to date, although it is more technical. Little and colleagues (Little, Card, Slegers, & Ledford, 2007) and Kline (2013) illustrate multigroup invariance tests with the effects coding identification approach. Invariance testing with binary and ordinal variables has received much less attention, but there are some exceptions that explore the issues specific to non-continuous variables with multigroup invariance tests (Millsap, 2011; Millsap & Yun-Tein, 2004) or longitudinal invariance tests (Bontempo et al., 2012). I particularly recommend some careful consideration of the elusive nature of testing for specific sources of invariance. Several authors delve deeply into the sticky details involved in this pursuit (Cheung & Rensvold; 1999; Steiger, 2002).

Notes

1 Software programs do not use the same ML chi-square value or SB chi-square value universally, and the scaling correction factor can be incorrectly obtained by using this ratio. Mplus and EQS output will produce the correct scf ratio from the two values, but output from LISREL should be recomputed (see Bryant & Satorra, 2012, for further details).
2 The Asparouhov–Muthén method is implemented in the Mplus software.
3 I chose this effect size index because of its broad applicability, its familiar effect size metric, and its ease of computation, but there are other alternatives that can be compared to conventional standards of effect size that might be adapted for this purpose (see Rosenthal, 1994, for a review).
4 The terms "scalar" and "full score" invariance have been used to refer to the case in which loadings and intercepts are invariant (De Beuckelaer & Swinnen, 2011; Van de Vijver & Leung, 1997, p. 144).

5 Although such an instance would be rare in practice, setting a scaling constraint at both time points but using a different value (e.g., $\psi_{11} = 2$ and $\psi_{22} = 1$) is not recommended in any instance in which invariance is tested, because it will lead to model estimates that are not equivalent to the models using other parameterizations (Steiger, 2002).

6 It is not always intuitive which models are equivalent under this approach. For example, to impose equality constraints on the factor variances, one might use the referent loading identification approach only at the first time point (e.g., $\lambda_{11} = 1$). The equivalent model using the factor identification approach only at the first time point would also need to constrain one of the loadings to be equal over time. Constraining any one of the loadings to be equal over time produces an equivalent fit due to the interdependence of the factor loadings and factor variances. Steiger (2002) provides a particularly lucid explanation of how the two common identification approaches may sometimes seem equivalent but do not always produce identical model fit or parameter estimates if only a subset of equality constraints are specified, referring to such phenomena as *constraint interactions*.

References

Asparouhov, T., & Muthen, B. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics*. Retrieved from http://statmodel2.com/download/webnotes/webnote10.pdf.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods*, *15*, 111–123.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bontempo, D. E., Grouzet, F. M. E., & Hofer, S. M. (2012). Measurement issues in the analysis of within-person change. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social sciences*. New York: Routledge.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, *16*, 201–213.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal, 19*, 372–398.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466.

Chen, F. F., Sousa, K. H., & West. S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 471–492.

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15*, 167–198.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1–27.

Cheung, G W., & Rensvold, R. B (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling: A Multidisciplinary Journal*, *9*, 233–255.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317–327

De Beuckelaer, A., & Swinnen, G. (2011). Biased latent variable mean comparisons due to measurement non-invariance: A simulation study. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Methods and applications in cross-cultural analysis* (pp. 117–148). New York: Taylor & Francis.

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*, 74–96.

Edwards, M. C., & Wirth, R. J. (2012). Valid measurement without factorial invariance: A longitudinal example. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 289–311). Thousand Oaks, CA: Sage.

Fan, X., & Sivo, S. (2009). Using Δgoodness-of-fit indices in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 54–69.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology*, *4*, 22–36.

Fix, E., Hodges, J. L., & Lehmann, E. L. (1959). The restricted chi-square test. In U. Grenander (Ed.), *Probability and statistics: The Harald Cramér Volume Edition* (pp. 92–97). New York: Wiley.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 378–402.

Hancock, G. R., Stapleton, L. M., & Arnold-Berkovits, I. (2009). The tenuousness of invariance tests within multisample covariance and mean structure models. In T. Teo & M. S. Khine (Eds.), *Structural equation modeling: Concepts and applications in educational research* (pp. 137–174). Rotterdam, Netherlands: Sense Publishers.

Hershberger, S. L. (2006). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 13–41). Greenwich, CT: Information Age.

Hipp, J. R., & Bollen, K. A. (2003). Model fit in structural equation models with censored, ordinal, and dichotomous variables: Testing vanishing tetrads. *Sociological Methodology*, *33*, 267–305.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117–144.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist*, *1*, 179–188.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 642–657.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*, 136–153.

Kelloway, E. K. (1995), Structural equation modelling in perspective. *Journal of Organizational Behavior*, *16*, 215–224

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 212–228.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation. *Educational Research and Evaluation*, 19, 204–222.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, & N. A Card (Eds.). *Modeling contextual effects in longitudinal studies* (pp. 121–148). New York: Routledge.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters upon DIF detection using the free-baseline likelihood ratio test. *Applied Psychological Measurement*, *33*, 251–265.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*, 391–410.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of IRT and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 333–356.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97–103.

McGraw, B. & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. *British Journal of Mathematical and Statistical Psychology*, *24*, 154–168.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*, 1016–1031.

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: latent growth curves with ordinal outcomes. *Psychological Methods*, *9*, 301–333.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*, 177–185.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Pyschometrika*, *58*, 525–543.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E. & Cham, H. (2011). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, and N. A. Card (Eds.) *Handbook of developmental research methods*. New York: Guilford.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479–515.

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf.

Muthén, B., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC Press.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*, 966–980.

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics*, *22*, 389–406.

Raju, N. S., Lafitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529.

Raykov, T. (2001). Studying change in scale reliability for repeated multiple measurements via covariance structure modeling. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 217–229). Lincolnwood, IL: Scientific Software International.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin 114*, *552–566*.

Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. *Educational and Psychological Measurement*, *43*, 723–728.

Raykov, T., Marcoulides, G. A., & Li, C. H. (2012). Measurement invariance for latent constructs in multiple populations. A critical view and refocus. *Educational and Psychological Measurement, 72*, 954–974.

Saris, W. E., & A Stronkhorst, H. (1984). *Causal modeling in nonexperimental research*. Amsterdam: Sociometric Research Foundation, SAS Institute, Inc.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In D. D. H. Heijmans, D. S. G. Pollock, & A. Satorra (Eds.), *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker* (pp. 233–247). Dordrecht, The Netherlands: Kluwer Academic.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*, 243–248.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp.179–200). Washington, DC: American Psychological Association.

Schmitt, N., Golubovich, J., & Leong, F. T. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: an illustrative example using Big Five and RIASEC measures. *Assessment*, *18*, 412–427.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210–222.

Sidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626–633.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting DIF with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in Cross-National Consumer Research. *Journal of Consumer Research*, *25*, 78–90.

Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods*, *7*, 210–227.

Taris, T. W., Bok, I. A.,& Meijer, Z. Y. (1998). Assessing stability and change of psychometric properties of multi-item concepts across different situations: A general approach. *Journal of Psychology*, *132*, 301–316

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park, CA: Sage.

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*, 4–70.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*, 10–18.

Woods, Carol M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*,*44*, 1–27.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *143*, 435–463.

Yuan, K.-H., & Bentler, P. M. (2006). Asymptotic robustness of standard errors in multilevel structural equation models. *Journal of Multivariate Analysis, 97*, 1121–1141.

Keywords
measurement invariance, longitudinal, nested models, likelihood ratio test

3 Structural Models for Comparing Dependent Means and Proportions

In this chapter, I introduce a range of SEM models that can be used for comparing means at just a few time points and show how these models can be related to traditional repeated measures analysis of variance (ANOVA). Repeated measures ANOVA tests for differences among two or more dependent means. Because we know that standard regression models represent a special case of SEM and that ANOVA is a special case of regression analysis (e.g., Cohen, Cohen, West, & Aiken, 2013; Graham, 2008), it makes sense that structural models can be used to compare means for repeated measures. This chapter generalizes the same structural equation models used to compare means when observed variables are continuous to compare proportions when the observed variables are binary or ordinal data. The analyses discussed focus on comparing means and proportions over just a few time points and are designed for *time-structured data*, in which data are collected at set regular intervals for all individuals.

A number of authors have discussed SEM approaches to ANOVA, although most have focused on the between-subjects ANOVA case using multigroup or multiple-indicator multiple-cause (MIMIC) models (e.g., Aiken, Stein, & Bentler, 1994; Hancock, 2003; Thompson & Green, 2006; Yuan & Bentler, 2006). Fewer authors have discussed repeated measures ANOVA (Rovine & Liu, 2012; Rovine & Molenaar, 2003; Voelkle, 2007), perhaps because there are a variety of more sophisticated longitudinal models available, such as latent growth curve models.

Why would we need to use a sophisticated analysis approach to test simple mean difference when ANOVA should generally do the job more easily? And why would we test such simple hypotheses about change when more advanced methods are available? There are several answers to these inevitable questions. The simple models discussed in this chapter provide an introduction to some of the fundamental concepts that underlie more sophisticated modeling strategies, such as latent growth curve models, latent difference models, and time series models. A richer understanding of these more complex models can be gained by understanding how SEM relates to some conventional analyses, such as repeated measures ANOVA or MANOVA. The modeling strategies discussed in this chapter can be easily applied to binary and ordinal variables, demonstrating that SEM is a single, general framework that encompasses conventional analyses such as chi-square and ANOVA. There are also potential statistical advantages by estimating latent variables, including larger standardized effect sizes, invariance tests, convenient missing data handling, and the possibility of incorporating mean comparisons into larger structural models.

Repeated Measures Analysis of Two Observed Continuous Means

Single Variable Difference Score Model

We begin by observing that the test of the differences between two related means, which can be conducted using a paired *t*-test ("dependent" or "repeated measures" *t*-test), is simply a test of whether or not the population mean difference equals zero. The *t*-test formula involves the average difference between two scores, $\bar{y}_{2-1} = \left[\sum\left(y_2 - y_1\right)\right] / N$, divided by its standard error estimate, $t = \bar{y}_{2-1} / SE_{\bar{y}_{2-1}}$. (The index *i* denoting individual scores is omitted throughout to simplify notation.) It follows then, that a *t*-test investigating whether a single mean is significantly different from zero is equivalent to the repeated measures *t*-test if the mean in the single-mean *t*-test is the average of difference scores. We also know that $t^2 = F$, which suggests that a repeated measures ANOVA is a test of difference between two means. Further, a special ordinary least squares (OLS) regression model could be specified that is also equivalent, if the dependent variable is used as a difference score and the model is tested as an intercept-only model, $y_{2-1} = \beta_0 + \varepsilon$. The intercept β_0 is a constant from the unstandardized solution and ε is the error term, which has a unique value for each case. The intercept provides an estimate of the expected value of the difference score or, in other words, the average difference score. The *t*-test of the coefficient for this model, β_0 / SE_{β_0}, is then equivalent to the matched pairs *t*-test value and $\sqrt{F}$ from the repeated measures ANOVA.

If we express this regression model as a structural model, then an equivalent model is depicted in Figure 3.1a, where a mean structure is estimated for the difference score, $y_{2-1} = \alpha_1 + \zeta_1$. The intercept α_1 is a constant, and the disturbance ζ_1 has a value for each individual. The model, which I will refer to as the *single variable difference score model*, tests the same hypothesis as the paired *t*-test (Rovine & Molenaar, 2003; Rovine & Liu, 2012).[1] The mean of the difference score variable, α_1, and its significance test is a Wald ratio (*z*-test or sometimes *t*-test) of the mean to its standard error estimate.[2] Assuming multivariate normality, ML estimate of the mean will be equivalent to the OLS estimate used in the conventional *t*-test and *F*-test. This model will be just identified with a chi-square equal to 0, but a test of a model that constrains the mean to 0 will be a test of the null hypothesis of no difference between the means in the population. The chi-square for the constrained model will be asymptotically equal to the *F*-value from the ANOVA, because the *t*-distribution equals the *z*-distribution and the *F*-distribution equals the chi-square distribution for one degree of freedom as $N \to \infty$. In practice, this implies that all of these statistical tests will be nearly identical under specified conditions when the sample size is equal to or greater than 120.

Equality Constraint Model

Another simple approach to testing the difference between two related means in SEM is to specify a model with two measured variables, freely estimating their means in one model and comparing the fit of the model to the fit of a second model with the factor means constrained to be equal. Figure 3.1b illustrates the *mean equality constraint model*. The difference in chi-square values from the two models will be asymptotically equal to the *F*-value from an ANOVA and t^2 from the matched pairs *t*-test.

Contrast Coding Model

An alternative to comparing two related means, which I will call the *contrast coding model*, exploits the measurement portion of the structural equation model to capture the

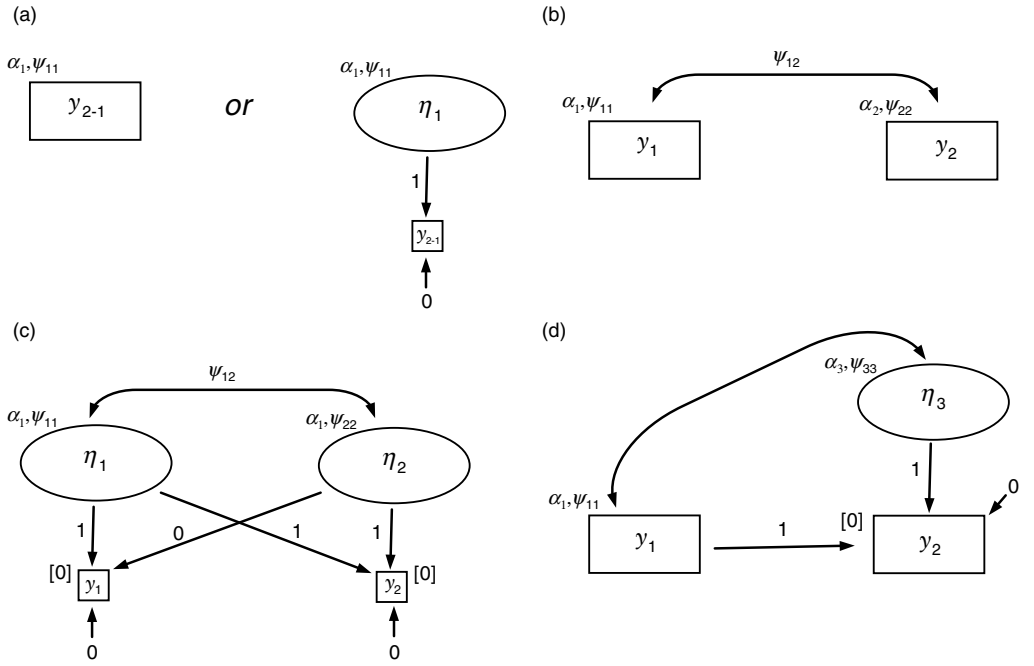

Figure 3.1 Four Approaches to Testing for the Difference between Two Observed Means: (a) single observed difference score model; (b) equality constraint model; (c) contrast coding model; (d) latent difference score model. Note: Ellipses represent latent variables and rectangles represent measured variables. Greek symbols outside of ellipses represent parameters that are estimated in the model, replaced by values when the parameter is set. Values presented in square brackets represent means or intercepts that have been set to the specified value.

differences between means. This model is illustrated in Figure 3.1c. A similar model was outlined previously for random effects models for dyadic data and is based on the parameterization of latent growth curves (Newsom, 2002; see also Voelkle, 2007).[3]

The logic of the contrast coding model is based on the use of a dummy-coded variable to explore mean differences. Let us first consider how we would obtain a test of the difference between two repeated measures using OLS regression. One method would be to compute a difference score, $y_2 - y_1$, to be used as the dependent variable in a special regression model that has only an intercept (i.e., no predictors). This type of regression model is allowable in some software programs but not all. Alternatively, the difference between y_2 and y_1 also can be estimated with a regression model if we reorganize the data such that there are two records for each case, one representing the Time 1 observation and one representing the Time 2 observation. This data structure is known as a "person × period" or "long" format. A dummy variable, with Time 1 equal to 0 and Time 2 equal to 1, is then created for use in the model as a predictor. The slope from the regression equals the average difference between the two means or the mean of the difference score. The intercept is equal to the value of y when $x=0$, which is equal to the mean on y for the first group. The statistical test for this regression analysis will tend to be underestimated, however, because the same individuals provide responses twice. If the correlation between the measures at the two time points is different from zero, as would often be the case in practice, this

analysis would violate the independence of observations (or errors) assumption and would underestimate the standard error used in the significance test.

Specification. We can conceptually reproduce the dummy variable regression with a structural model to obtain the difference between two observed scores using two measurement model equations and substituting latent variables for the regression coefficients. Because this analysis does not require restructuring the data, it can be conducted without violating the independence of error assumption. By constraining factor loadings and estimating factor means, the roles of the two parameters are switched so that the loading replaces the x dummy variable and the latent variables replace the regression coefficients.

$$E(y_1) = \eta_1(\lambda_{11}) + \eta_2(\lambda_{12})$$
$$E(y_2) = \eta_1(\lambda_{21}) + \eta_2(\lambda_{22})$$

In these equations, intercept and contrast factors are defined by a loading pattern with Time 1 and Time 2 observed variables (y_1 and y_2) serving as indicators of both latent variables. Both loadings on the first factor are set equal to 1, and the loadings on the second factor set equal to 1 and 0 for the two time points, respectively. Inserting the values for the loadings yields the following two equations.

$$E(y_1) = \eta_1(1) + \eta_2(0)$$
$$E(y_2) = \eta_1(1) + \eta_2(1)$$

When we substitute and rearrange the terms, it becomes clear that the average of the first factor is equal to the mean of the first observation (i.e., the value of y when x equals 0), and the average of the second factor is equal to the difference between the two observations.

$$E(y_1) = E(\eta_1)$$
$$E(y_2) = E(y_1) + E(\eta_2)$$

Rearranging the terms, gives:

$$E(\eta_2) = E(y_2) - E(y_1) = E(y_2 - y_1)$$

Thus, the mean of the first factor equals the mean of the first observed variable (Time 1), $E(\eta_1) = \alpha_1 = E(y_1)$, and the mean of the second factor equals the average difference score (or the difference between the means of the two observed variables), $E(\eta_2) = \alpha_2 = E(y_2 - y_1)$. I will call η_1 the *intercept factor* and η_2 the *contrast factor*. The model can be expanded easily to compare more time points, conduct trend analyses, and incorporate mean comparisons into larger structural models.

There are several details of model specification that should be mentioned. The measurement intercepts are set equal to 0, so that the means of the two factors can be estimated. Identifying the model in this way transfers the means of the observed variables to the latent means, so that the loadings provide special interpretations as the intercept and difference. To identify the latent factors, constraints need to be placed on measurement residual variances or factor variances. One approach is to set the measurement residuals at the two time points equal to 0. This allows the intercept factor variance, the contrast factor variance, and their covariance to be estimated. An alternative approach, which may be convenient for some applications, is to estimate the measurement residuals but constrain the variance of the intercept factor or the contrast factor to be equal to 0. This parameterization parallels

the multilevel regression or growth curve model, and it is possible to conceptualize the model as a random effects repeated measures ANOVA where differences vary across cases (Newsom, 2002).[4] A multilevel regression model with just two time points (or two cases per group) would also require constraints on one or more of the random effects to identify the model, a practice that is not uncommon in this analysis tradition (e.g., Stoolmiller, 1994). The multilevel parameterization allows for tests of homogeneity of variance by constraining the measurement residuals at the two time points to be equal.

Relation to ANOVA. Modification of this model to using effect codes (–1, +1) for the loadings for the contrast factor leads to an estimate for α_1 that represents the grand mean (i.e., average of the Time 1 and Time 2 observed means) and an estimate of α_2 that represents the difference of each observed mean from the grand mean. The effect codes imply that α_2 will be half the value obtained with a dummy coding scheme (0, 1).

With measurement residuals estimated, the measurement portion of the contrast coding model can be stated for an individual case as

$$y_{it} = \nu_t + \lambda_{t1}\eta_1 + \lambda_{t2}\eta_2 + \varepsilon_{it} \tag{3.1}$$

The index i is added to emphasize values for an individual case in the data set, and the subscript t refers to a particular time point. For now, assume there is only one observed variable at each time point. To identify the model, the measurement intercept, ν_t, is set equal to 0 and can be dropped from the equation. Because the values for λ_{t1} are set equal to 1, this parameter also can be dropped from the model. The expected value of y_{it} is then

$$\begin{aligned} E(y_{ti}) &= E(\eta_1) + E(\eta_2) \\ E(y_{ti}) &= \alpha_1 + \alpha_2 \end{aligned} \tag{3.2}$$

where, under the effect coding scheme (–1, +1), α_1 is the grand mean and α_2 is the average difference of each mean from the grand mean.

Using common notation for ANOVA (e.g., Winer, 1971), the statistical model can be stated as

$$\begin{aligned} Y_{ti} &= \mu + \pi_i + \tau_t + \varepsilon_{ti} \\ E(Y_{ti}) &= \mu + \tau_t \end{aligned} \tag{3.3}$$

where μ represents the grand mean, π_i represents the average of scores for each case (i.e., averaged across time points), τ_t represents the deviation of the mean at each time point from the grand mean, and ε_{ti} represents the error. In the ANOVA model, the expected value of π_i is assumed to be zero in the population, so it drops out of Equation (3.3). Thus, Equation (3.2) for the structural equation model is the same as Equation (3.3) for the ANOVA model.

There are several other parallels between the contrast coding model and the traditional ANOVA model worth noting. When effect codes are used, the variance of α_1 provides information about the variability scores across individuals averaged over time points. In other words, the factor variance, $\mathrm{Var}(\eta_1) = \psi_{11}$, is equal to the mean square for subject (MS_s). Following path tracing rules, decomposition of the covariance between two observed variables is

$$\mathrm{Cov}(y_1, y_2) = \lambda_{11}\lambda_{21}\mathrm{Var}(\eta_1) + \lambda_{12}\lambda_{22}\mathrm{Var}(\eta_2)$$

Substituting values specified in the contrast coding model with effect codes, we have

$$\text{Cov}(y_1, y_2) = 1 \cdot 1 \cdot \text{Var}(\eta_1) + \left[(-1) \cdot (1) \cdot (0)\right]$$
$$\text{Cov}(y_1, y_2) = \text{Var}(\eta_1)$$

Thus, the variance of the intercept factor is equal to the observed covariance to the two repeated measures.

The parameter estimates from the contrast coding model provide another useful quantity from repeated measures ANOVA. It can be shown that the mean square error (MS_{error}) is a function of the average of the observed variances and the covariance between the repeated measures (Winer, 1971).

$$MS_{\text{error}} = \frac{\text{Var}(y_1) + \text{Var}(y_2)}{2} - \text{Cov}(y_1, y_2)$$

From path tracing rules, we know that the residual variance in the measurement error is a function of the observed score variance, the square of the loading, and variance of the respective factor k.

$$\text{Var}(\varepsilon_{ti}) = \text{Var}(y_{ti}) - \lambda_{tk}^2 \psi_{kk}$$

The subscript t is used to indicate a particular time point and the subscript k stands for the kth factor. If the measurement residual variances for two time points, y_1 and y_2, are estimated and set equal, the equation becomes

$$\text{Var}(\varepsilon) = \frac{\text{Var}(y_1) + \text{Var}(y_2)}{2} - \lambda_{tk}^2 \text{Var}(\eta_k)$$

To simplify the notation, the i subscript is eliminated from y_1 and y_2, and the i and t subscripts are eliminated from ε. And when we substitute that $\text{Cov}(y_1, y_2) = \text{Var}(\eta_1)$ and $\lambda_{11}^2 = 1$ from the contrast coding model into the equation, we see that the estimate of the residual variance from the contrast coding model is equal to MS_{error}.

$$\text{Var}(\varepsilon) = \frac{\text{Var}(y_1) + \text{Var}(y_2)}{2} - \text{Cov}(y_1, y_2) = MS_{\text{error}}$$

Latent Difference Score Model

The structural model parameter specifications can be manipulated in another way to test the difference between two means. Recall that the mean difference between two time points can be estimated with a regression model using a pre-computed difference score as the dependent in an intercept only model. Equivalently, we can obtain the difference with a simple model of y_2 regressed on y_1, $y_2 = \alpha_2 + \beta_{21} y_1 + \zeta_2$, where the path coefficient, β_{21}, is set equal to 1. Inserting 1 into the model equation gives $y_2 = \alpha_2 + 1y_1 + \zeta_2$, which is the same as $y_2 - y_1 = \alpha_2 + \zeta_2$. This model, tested with any SEM software if the mean structure is estimated, has an intercept α_2 equal to the difference between the two means and has disturbance variance $\text{Var}(\zeta_2) = \psi_{22}$ equal to the variance of the difference scores.

This specification is a simple form of the more general *latent difference score model* (McArdle, 2001) that can be used with many time points. The measurement equation and the structural equation, taken together, reproduce the difference score model shown in Figure 3.1d.

$$y_2 = \alpha_2 + (\beta_{21})y_1 + (\lambda_{23})\eta_3 + \zeta_2$$

To specify the model, the path coefficient from y_1 to y_2 is set equal to 1, the structural intercept, α_2, is set equal to 0, and the disturbance, ζ_2, is set equal to 0. If we plug in these values and rearrange terms, we see that the latent variable η_3 represents the difference score.

$$\begin{aligned} y_2 &= (0) + (1)y_1 + (1)\eta_3 + (0) \\ \eta_3 &= y_2 - y_1 \end{aligned} \tag{3.4}$$

The estimate of the mean for the latent variable, α_3, will be equal to the average difference, and the variance of the latent variable will represent the variability of difference scores. Latent difference models are discussed at greater length in Chapter 9.

Example 3.1: Comparing Means for Two Observed Variables

Single Variable Difference Score Model. I illustrate the several SEM approaches to testing the difference between two related means using as the observed variable derived from an average of five items from the positive affect measure (self-rated feelings of "happy," "enjoying yourself," "satisfied," "joyful," and "pleased") in the social exchanges data set (see the Example Data Sets at the beginning of the book for a description) at Waves 1 and 2 ($N=574$). Syntax and data sets used in the examples are available at the website for the book. The means for the composite measure at Wave 1 and Wave 2 were 3.034 and 2.925. The mean difference was –.109, showing a decline in positive affect over the six-month period between interviews. This was a significant decline according to a matched pair *t*-test, $t(573)=-3.90$, $p < .001$, and a repeated measures ANOVA, $F(1,573)=15.24$. Note that the square of the *t*-value is equal to the *F*-value within rounding error, $(-3.90)^2=15.21$.

A single-variable structural equation model of the difference score, $y_2 - y_1$, with mean structure (see Figure 3.1a) produced a mean estimate that was equal to the average difference –.109. Dividing this estimate by its standard error (i.e., the Wald test) gave a *z*-value of –3.907, $p < .001$, that was equal to the matched pair *t*-test and the square root of the *F*-test within rounding error. A model constraining the mean to zero, had a significant chi-square, $\chi^2(1)=15.063$, $p < .001$, that was nearly identical to the *F*-test value and to the square of the Wald test. The Wald and the likelihood ratio tests statistics are asymptotically equivalent under ideal conditions, but may differ if there is a misspecified mean structure or nonnormal data.

Equality Constraint Model. An equivalent structural model was estimated using a nested model comparison of the composite variables for positive effect at Wave 1 and Wave 2 (see Figure 3.1b). With no other constraints (e.g., equal variances), the model freely estimating the means is just identified, so there is no fit information. The fit of a second model constraining the two means to be equal then represents a test of the difference between the two means. The chi-square for the constrained model was significant, $\chi^2(1)=15.063$, $p < .001$. This value matches the *F*-value from the ANOVA and the chi-square test from the difference score model.

Contrast Coding Model. To test the same hypothesis, the contrast coding model depicted in Figure 3.1c used the positive affect composite score measured at Time 1 and Time 2 as indicators. In this model, the mean of the intercept factor, η_1, estimates the observed mean at Time 1, and the mean of the contrast coding factor, η_2, estimates the difference between

the two means. Both loadings for η_1 were set equal to 1, and loadings for η_2 set to 0 and 1 for Time 1 and Time 2, respectively. For identification, the measurement intercepts and the measurement residuals were set equal to 0. The factor means, factor variances, and factor covariances were estimated.

The model was just identified. Results indicated that the mean of the first factor, α_1, was equal to 3.034, which is also the mean of the positive affect measure at Time 1. The estimate for the mean of the difference factor, α_2, was –.109, a match to the mean difference obtained from the other analyses. The z-value was –3.907, $p < .001$, for α_2, corresponding closely with the test statistics obtained with the other models. A second model, constraining the mean of the difference factor, α_2, to be equal to 0, led to a significant increase in the model chi-square, $\chi^2(1)=15.063$, $p < .001$, a value identical to the chi-square with the equality constraint model and the other analyses.

Latent Difference Score Model. The latent difference model (McArdle, 2001) provides an alternative test of the same hypothesis, with specifications following Figure 3.1d. This model is just identified. The mean of the first factor, α_1, was 3.034, and the intercept of the difference factor, α_2, was –.109, $z=-3.907$, $p < .001$. The square of the z-value was very similar to the F-value from the ANOVA and chi-square difference found with previous models, $z^2=(-3.907)^2=15.265$. The chi-square for model restricting the mean of the difference factor was equal to 15.064, which also closely matched the results from the other methods.

Comments

The structural modeling approach to repeated measures ANOVA is flexible, with several equivalent specification approaches that can be used to compare the difference between two means. These models are far simpler than what most researchers will use in practice, but they demonstrate the underlying equivalence of SEM and ANOVA. Moreover, this section has begun to show how the ANOVA models relate to other longitudinal models. Comparisons with the equality constraint model make use of the same strategy used for measurement invariance tests, and the contrast coding and difference score models are simple forms of the latent growth curve model and latent difference score models used with more repeated measurements. Within a broader modeling perspective that includes categorical variables, we next examine how these same ANOVA specifications also encompass traditional analyses for binary repeated measures.

Comparing Two Observed Binary Variables

Longitudinal structural equation models also may be used to investigate change in binary variables. Chapter 1 reviews SEM with binary variables, including estimation methods, unobserved y^* concept, thresholds, and model specification details. Statistical tests with binary variables involve comparisons of frequencies or proportions rather than means. With a 0 and 1 coding of the binary variable, the mean is equal to the proportion, so ANOVA structural equation models used to compare means with continuous variables can be applied to compare proportions with binary variables. The models covered in this chapter make use of continuous latent variables, but it is possible to conceive of binary or categorical latent variables and to investigate changes in proportions through latent class or latent profile analysis – referred to as *latent transition analysis*. These models are covered in Chapter 10.

Among conventional tests, McNemar's chi-square is the most common repeated measures test comparing two dependent (e.g., repeated measures) frequencies or proportions.

McNemar's chi-square is more generally described as a test of the *marginal homogeneity hypothesis* (Agresti, 2013). The term refers to the comparison of the marginal frequencies for consistent "yes" (or, alternatively, consistent "no") responses over time in a 2×2 repeated measures table. The Pearson chi-square test and its related variants are designed for a standard 2×2 contingency table and are not appropriate for a repeated measures design, because the observations are not independent. Although structural equation models can be constructed to test marginal homogeneity, there are fewer specifications options than available for continuous variables. With binary variables, the analysis of difference scores is not equivalent to the marginal homogeneity test. The subtraction of one binary variable (0, 1) from another binary variable leads to three possible values, –1, 0, and +1. Although an analysis of the three-category variable could be conducted using ordinal regression methods, I will not discuss this approach here.

Equality Constraint Model

A repeated measures comparison of two proportions can be conducted that follows the equality constraint model for comparison of two means described earlier. This method tests for marginal homogeneity using a likelihood ratio test comparing a model with constrained thresholds to a model with freely estimated thresholds. Thresholds rather than measurement intercepts are estimated for binary observed variables, so the test of the equality of proportions involves constraints on these thresholds. The appropriate test of the dependent proportions hypothesis requires estimation of the correlation between the two binary measures. For two time points, the freely estimated model is just identified, and the fit of the constrained model provides a test of the equality of the marginal proportions.

Because the correlation between the two variables is required for the appropriate model, the binary estimation method matters for this specification. Covariances of binary measured variables cannot be estimated (Muthén & Asparouhov, 2002), so the correlation cannot be included in the model. This limitation applies even if latent variables with single indicators are used to specify the model as well. Without including the correlation, the comparison is for two independent frequencies, which tests the same hypothesis as the traditional Pearson chi-square. The appropriate repeated measures hypothesis can be investigated using WLSMV with theta parameterization, however, because scaling constraints on the y^* distribution allow error variances to be estimated. The model chi-square will be an appropriate test of the hypothesis that the two proportions are equal over time, although the result will not be identical to McNemar's chi-square. Observed proportions can be recovered from the means of the factors using the appropriate cdf, which is the normal (probit) distribution when WLSMV estimation is used.

Contrast Coding Model

The contrast coding model is an alternative method of comparing marginal proportions. The model is specified following the general form as in Figure 3.1c. Full ML estimation for binary variables provides an appropriate repeated measures test of equal proportions that corresponds with conventional analyses. WLSMV estimation may also be used for this model, but the results do not correspond to the conventional McNemar's test. The factor mean estimates are identified by setting the measurement thresholds to be equal to 0. Because constraints cannot be placed on measurement residuals with this estimation approach, the model was specified with the multilevel regression parameterization with the contrast factor set equal to 0. Because the model is just identified with this

specification, the overall fit of a model with mean of the contrast factor set equal to 0 gives a test of equality of the marginal proportions. If binary ML is used, the Pearson chi-square for the fit of the model is equal to McNemar's chi-square.

Alternatively, the significance test of the contrast factor mean also indicates whether the marginal proportions differ. The *z*-test, or Wald test, is generally not equal to the likelihood ratio test, but it does match another known analysis. Conditional logistic regression is a less commonly used method for comparing marginal proportions that yields results that differ somewhat from McNemar's test (for a description, see Newsom, 2012). The conditional logistic regression analysis is designed to compare the proportions of discordant cells from a 2×2 repeated measures table. The discordant cells are the yes-no and no-yes cells, and comparing them tests the *axial symmetry hypothesis*. The difference between two consistent marginal proportions (e.g., yes-yes) is algebraically equivalent to the difference between the two discordant cells so the marginal homogeneity and axial symmetry hypotheses are equivalent for binary variables (Agresti, 2013). The significance test of the contrast factor mean gives the same result as the significance of the test of the coefficient in the conditional logistic analysis, where the square of the Wald ratio from the contrast coding model, $z^2 = \left(\alpha_2 / SE_{\alpha_2}\right)^2$ is equal to the Wald chi-square from the conditional logistic analysis.

Example 3.2: Two Binary Observed Variables

A simple example illustrating the relation between the contrast variable model with two binary observed measures and the McNemar test shows the general equivalence of the structural modeling approach to conventional tests for binary variables. Using a question from the social exchanges data set, I investigated changes in the respondent's report of whether or not a major health event occurred recently, with a yes/no question asked six months after the initial interview ($N = 574$). On the first occasion, only 42 (7.3%) reported a major health event, whereas, on the second occasion, 140 (24.4%) reported an event. The single degree of freedom McNemar's chi-square value was 68.600, $p < .001$, indicating a significant increase in the proportion reporting an event.

A contrast coding model that followed the model depicted in Figure 3.1c, but specified with variance of the contrast factor set to 0, was tested to illustrate. Binary ML estimation was used for greater comparability with conventional analyses. No fit information was available for this model, because it was just identified. The mean of the contrast factor was 1.735, SE = .237, $z = 7.329$, $p < .001$. The mean does not have a simple intuitive interpretation. It represents the natural log of the average difference, known as the subject-specific effect, which is not the same as the natural log of the difference between two proportions, known as the population-average effect. The square of the Wald ratio, $z^2 = (7.329)^2 = 53.714$, is not equal to the McNemar test but does match the results from a conditional logistic analysis within rounding error. A subsequent model that set the contrast factor mean equal to 0 had a significant chi-square value, $\chi^2(1) = 68.596$, $p < .001$, also indicating a difference in proportions. Because the comparison model is just identified, this value is the same as a likelihood ratio test and is nearly identical to the McNemar test.

Comments

The equivalence of the structural equation ANOVA models and conventional analyses for binary matched pairs, such as McNemar's chi-square, suggest a single powerful and flexible modeling framework that is general enough to be extended to non-continuous data

types. Next we see how the same modeling strategy can be extended to ordinal variables with only a few additional considerations.

Comparing Two Observed Ordinal Variables

Similar tests for comparing repeated proportions across two time points can be conducted when the observed variables are ordinal. Implementation of these models is a relatively straightforward extension, especially given that the same estimators, ML or WLSMV, can be employed. The usual hypothesis of interest concerns whether responses are consistent over time, such as whether those responding "never" at Time 1 are likely to respond "never" at Time 2 and those responding "sometimes" at Time 1 also respond "sometimes" at Time 2. For a repeated measures table that is larger than 2×2, however, the marginal homogeneity hypothesis is no longer equivalent to the axial symmetry hypothesis. Instead, the axial symmetry hypothesis becomes a joint hypothesis of marginal homogeneity and quasi-symmetry (Agresti, 2013). The marginal homogeneity (MH) test, therefore, must be computed from the difference of the test for the symmetry hypothesis (S) and the test of a quasi-symmetry hypothesis (QS), $MH = S - QS$. I introduce the general extension of the ANOVA models to ordinal variables here, but it is unavoidable that many important ordinal analysis issues must be left unaddressed.

Equality Constraint Model

One method of comparing proportions from two ordinal variables is to compare the fit of a model with measurement thresholds that are freely estimated to a model with measurement thresholds values that have been constrained over time. As with binary variables, the estimation options are limited, because correlations among the observed variables are needed. Neither ordinal ML nor WLSMV with delta parameterization can be used to test for marginal homogeneity with the proportional equality constraint model. An appropriate test can be obtained using WLSMV with theta parameterization by comparing a model with freely estimated thresholds to a model with longitudinal equality constraints on the thresholds.

One variant on the specification of the equality constraint is possible when variables involve three or more ordinal categories. Using WLSMV with theta parameterization, a single-indicator latent variable can be constructed for each time point, with each loading set equal to 1 and each measurement residual set equal to 0. The factor mean may be identified by setting the first threshold equal to 0 and estimating the remaining thresholds. Recall from Chapter 1 that such a constraint links the factor mean estimate to the proportion of the first ordinal category. Thus, a comparison of factor means is a comparison of the equivalence of the proportions in the first category, such as those responding "never" for a scale with "never," "sometimes," and "always" response options. With only two categories, this is not a problem, because the second category is determined. Comparing the factor means compares both proportions simultaneously, where the proportion responding "no" determines the proportion responding "yes" With three or more ordinal categories, however, a test of the equivalence of factor means can only compare proportions in the first category if the first threshold serves as the referent. A likelihood ratio test compares a model holding all thresholds equal to a model that allows all remaining $C - 1$ thresholds to vary over the two time points. The difference in chi-square values for the more restricted and the less restricted models provides a test of marginal proportions. Results of this model will closely correspond to other tests that are used for marginal homogeneity, such as the loglinear test.[5]

Contrast Coding Model

The contrast coding model can also be used for testing proportional differences over two time points. Following the general model specification depicted in Figure 3.1c, nested models using ML or WLSMV can be used to compare the chi-square estimates for a model with the difference factor mean constrained to zero or to a model in which the difference factor mean and $C-1$ thresholds are freely estimated. For identification, one threshold from each factor is set equal to 0. Results from the comparison yield a test of marginal homogeneity identical to results from the equality constraint model and closely matching traditional tests of marginal homogeneity. The cdf transformation of the mean of the intercept factor provides the observed proportion of the first response category at Time 1. The Wald test of the contrast factor mean (i.e., that the difference factor is equal to zero) is a test of the difference in proportions of cases in the first ordinal category. The appropriate cdf transformation – logistic for ML estimations and normal for WLSMV estimation – can be used for converting the factor mean estimate to the observed proportion.

Latent Difference Score Model

The latent difference score model is an equivalent method of testing the differences between two measured ordinal variables. The model is specified similarly to the model depicted in Figure 3.1d, but defining two single-indicator latent variables, η_1 and η_2, that replace y_1 and y_2. Similar constraints to those described for the equality constraint and the contrast coding models can be placed on the thresholds. One threshold from each factor is set equal to 0 for identification, with the remaining thresholds held equal over time. The model is otherwise specified in a similar manner to the continuous model, with the disturbance for the η_2 set equal to 0 and the causal effect of the difference factor η_3 on η_2 set equal to 1. A model with the difference factor mean constrained to be equal to 0 is compared to a model with the difference factor mean freely estimated for a test of marginal homogeneity.

Example 3.3: Two Ordinal Observed Variables

To illustrate the test of marginal homogeneity for two ordinal variables over time, I examined changes in response proportions over six months for one item ("you felt sad") from the depression measure in the social exchanges data set. Responses for the question "how often did you feel this way during the past week?" were made on a 4-point ordinal scale ranging from 0 "none of the time" to 3 "most of the time." Each of the following examples used WLSMV estimation with theta parameterization. The equality constraint model followed the general structure illustrated in Figure 3.1b but used two single-indicator latent variables instead of y_1 and y_2, each defined by setting the loading for its respective ordinal sadness variable equal to 1 and the measurement residual equal to 0. (An alternative specification placing equality constraints on all three thresholds could be used with any software program that allows estimation of correlations among ordinal variables.) For identification, the first threshold was set to 0. Constraining the factor and the remaining thresholds to be equal over time provides a test the marginal proportions are equal in all of the categories. The model with equality constraints had a nonsignificant chi-square, $\chi^2(3)=1.081$, $p=.782$. Another model allowing the factor means and the thresholds to vary over time (but keeping the identification constraint for the first threshold) was tested to estimate the factor means at each time point. The values obtained for α_1 and α_2 were –.306 and –.334, respectively. Using the normal cdf transformation, the estimated proportions precisely match the observed proportions of .620 and .631 for participants responding "never" to the sadness question at the two time points.

The marginal homogeneity hypothesis was also investigated using the contrast coding model (Figure 3.1c) and the latent difference score model (Figure 3.1d). The contrast coding model specified an intercept factor and a contrast coding factor. The intercept factor set loadings for the Time 1 and Time 2 measurements of sadness equal to 1, whereas the contrast coding factor set the loading for the two measurements to 0 and 1. The first threshold at each time point was set equal to 0 for identification. The marginal homogeneity test was conducted by setting the remaining two thresholds to be equal over time and the mean of the contrast factor equal to 0. The fit for the constrained model suggested that the marginal proportions for the sadness rating categories did not differ, $\chi^2(3)=1.081$, $p=.782$. The result was the same as that obtained with the equality constraint model and is close to the traditional marginal homogeneity test (Agresti, 2013) conducted by subtracting the chi-square for the quasi-symmetry model from the chi-square from the symmetry model, $\chi^2_{MH} = \chi^2_S - \chi^2_{QS} = 1.959 - 0.986 = 0.974$, which would not be a significant difference.

Allowing the second and third thresholds to be estimated freely at each time point and estimating the contrast factor mean produced a just identified model. The intercept factor mean from this model was –.412. The normal cdf transformation of this value returns a value of .660, approximately equal to the proportion responding "never" to the sadness question at Time 1. The estimate of the contrast factor mean (as well as the difference factor mean) was –.037. Using .5 minus the cdf transformed value returns the proportion estimate of .015, a value similar to the difference between the proportion responding "never" at Time 1 and Time 2.

The latent difference score model gives identical results to the contrast coding model. With thresholds freely estimated, the model is just identified. The mean of the difference factor was –.037, which equals the value obtained with the contrast coding model. With the factor mean for the difference latent variable, α_3, constrained to be equal to 0, the model had the same fit as the contrast coding model, $\chi^2(3)=1.081$, $p=.782$.

Comments

Analyses of ordinal data for repeated measures have often presented researchers with challenges, including identifying the best test, finding software features that will implement them, and understanding their underlying rationale. We now see how the same ANOVA model specifications used for continuous or binary variables can be conveniently applied to ordinal variables as well. Though space does not allow for a discussion of other variable types, there is no reason why the same models cannot be applied to other kinds of data, such as counts or variables with zero-inflated distributions, whenever SEM software programs allow.

Repeated Measures Analysis of Three or More Observed Means

I return to models with continuous variables to discuss ANOVA models for comparing means at several time points.

Equality Constraint Model

The comparison of two means within the structural modeling framework is easily extended to three or more time points via likelihood nested model comparisons (Figure 3.2a). With the equality constraint model, T repeated measures are included. An omnibus test can be obtained by imposing an equality constraint on all means, or specific contrasts can

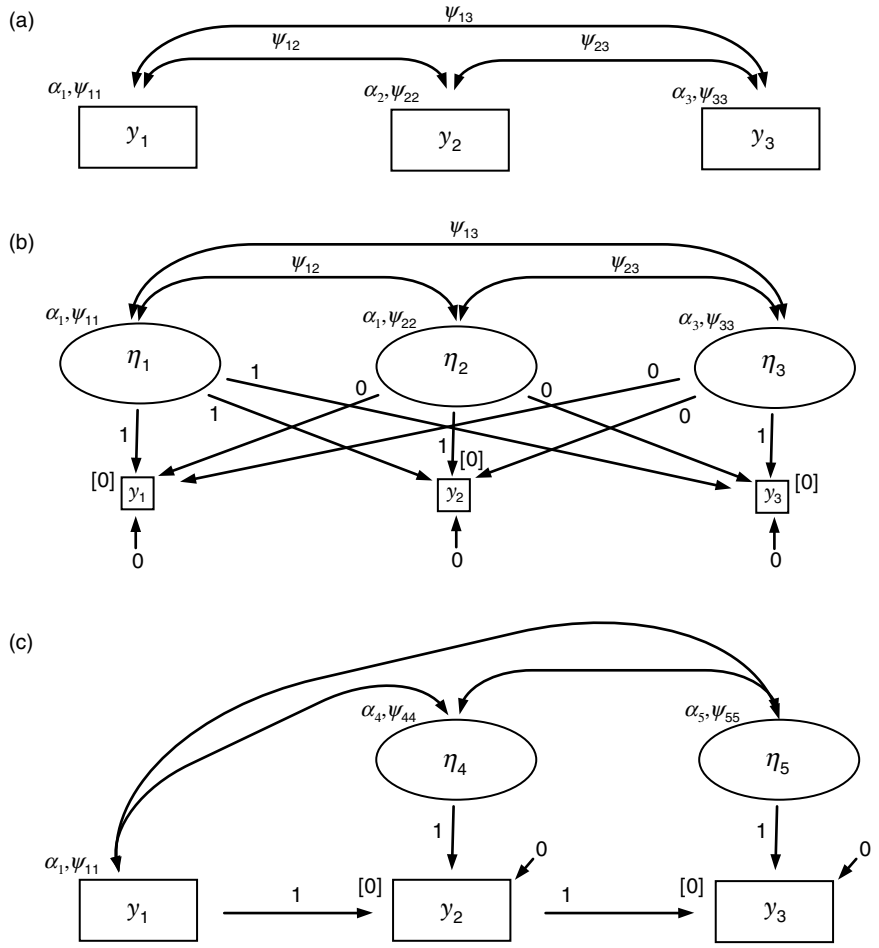

Figure 3.2. Three Approaches to Testing for the Difference between Three Observed Means: (a) equality constraint model; (b) contrast coding model; (c) latent difference score model.

be tested by imposing equality constraints on only a subset of means. For the omnibus test, the chi-square difference will have $T-1$ degrees of freedom. If correlations among observed variables are estimated, the unconstrained model will be just identified, and a test of significance of the difference among the means is obtained from the model chi-square constraining all means to be equal.

When more than two means are compared, the value of the resulting likelihood ratio will no longer closely correspond to the *F*-value from the repeated measures ANOVA because of different governing assumptions of the two analysis approaches. With three or more time points, repeated measures ANOVA is estimated under several assumptions, such as sphericity and compound symmetry. The chi-square and likelihood ratio test used in SEM are sensitive to multivariate normality as well as correct model specification (West, Taylor, & Wu, 2012; Yuan & Bentler, 2006), further complicating the comparison. The model chi-square from constraining all means to be equal more closely relates to the *F*-value from the multivariate analysis of variance (MANOVA; discussed below), which

makes fewer assumptions than repeated measures ANOVA. The equality constraint model can be modified to specify various error structures, such as requiring that variances be equal over time. Assumptions are discussed further below.

Contrast Coding Model

The contrast coding model can be expanded as an additional strategy for testing the equivalence of multiple means (Figure 3.2b). Although it is not initially apparent, this model resembles the repeated measures ANOVA model in several respects. Because an omnibus test of equivalence several population means, H_0: $\mu_1=\mu_2=\ldots=\mu_T$, is equivalent to separate tests of all possible paired differences between consecutive means, H_0: $\mu_2-\mu_1=0$ through $\mu_T-\mu_{T-1}=0$ (Stevens, 2012), a set of contrasts can be captured by $T-1$ differences. This equivalence is true, because one of the three possible differences contains information that is redundant with the other two differences. For example, if it is known that $\bar{y}_1=\bar{y}_2$ and $\bar{y}_1=\bar{y}_3$, then $\bar{y}_2$ must be equal to $\bar{y}_3$. Thus, a repeated measures ANOVA for three-wave design can be tested by examining just two differences, say $\bar{y}_2-\bar{y}_1$ and $\bar{y}_3-\bar{y}_1$, in the contrast coding model. The following specification of loadings for two contrast coding factors, η_2 and η_3, obtains the necessary contrasts.

$$y_1=\eta_1(1)+\eta_2(0)+\eta_3(0)+\varepsilon_1$$
$$y_2=\eta_1(1)+\eta_2(1)+\eta_3(0)+\varepsilon_2$$
$$y_3=\eta_1(1)+\eta_2(0)+\eta_3(1)+\varepsilon_3$$

Because η_1 has loadings all assigned to 1, its mean, α_1, will be an estimate of the average of the observed variable for the first wave, $\bar{y}_1$. The other two latent variables, η_2 and η_3 are contrast factors with loading vectors containing contrast codes of 0, 1, 0 and 0, 0, 1. Their means, α_1 and α_2, will represent differences in observed variable means $\bar{y}_2-\bar{y}_1$ and $\bar{y}_3-\bar{y}_1$. Taken together, the model can assess the differences among the three means, as shown in the following equations.

$$E(y_1)=E(\eta_1)$$
$$E(y_2)=E(\eta_1)+E(\eta_2)$$
$$E(y_3)=E(\eta_1)+E(\eta_3)$$

By substituting in y_1 for η_1, we have

$$E(y_2)=E(y_1)+E(\eta_2)$$
$$E(y_3)=E(y_1)+E(\eta_3)$$

Rearranging terms then shows that factors η_2 and η_3 represent the two sets of difference scores.

$$E(\eta_2)=E(y_2-y_1)$$
$$E(\eta_3)=E(y_3-y_1)$$

The expected values of the estimated means for the two factors, α_2 and α_3, equal the average difference between scores from the two waves. An omnibus test of differences among the three waves can be obtained by the fit of a model with both difference factor means constrained to be equal to 0.

Choosing to examine these two particular mean differences is, of course, arbitrary and we could just as well construct contrast variables with other coefficients, as long as they constitute a complementary set of orthogonal contrasts that will together capture the three mean differences. Researchers may be interested most often in the differences between means at adjacent time points (e.g., whether $\mu_2 = \mu_1$ and $\mu_3 = \mu_2$ in the population), but, for some applications, such as a study with a pretest–posttest design with a subsequent follow-up, comparisons of the pretest to the posttest and the follow-up may be of most interest (i.e., whether $\mu_2 = \mu_1$ and $\mu_3 = \mu_1$ in the population). This contrast coding model can be adapted for other comparisons as well, including comparing the average of two waves to a third wave as with the Helmert contrast (e.g., 2, –1,–1 and 0, –1, 1), forward differencing (–1, 1, 0 and 0, –1, 1), or trend analysis (e.g., –1, 0, 1 and 1, 0, 1).

The traditional *F*-test can be derived in several ways. A nested test can be conducted which compares a model with the means of the two contrast coding factors constrained to a model in which the factor means of the two contrast coding variables are free to be estimated. Whereas the nested model comparison provides an omnibus hypothesis test, the ratio of the mean of each contrast coding to its standard error, α_i / SE_{α_i}, provides a Wald test of contrast of each pair of means. This result will match the paired *t*-test and the a priori repeated measures comparison from ANOVA, where $\left(\alpha_i / SE_{\alpha_i}\right)^2 = F$ with equivalent significance tests for large *N*.

Alternatively, the sum of squares for the repeated measures effect can be computed from the results to obtain the *F*-value if the differences among all means are known. With the coding scheme described above, in which the means of the two contrast coding factors represent $\bar{y}_2 - \bar{y}_1$ and $\bar{y}_3 - \bar{y}_1$, the third mean difference, $\bar{y}_3 - \bar{y}_2$, is simply derived from the other two, $(\bar{y}_3 - \bar{y}_1) - (\bar{y}_2 - \bar{y}_1) = \bar{y}_3 - \bar{y}_1 - \bar{y}_2 + \bar{y}_1 = \bar{y}_3 - \bar{y}_2$. The sum of squares for the mean effect for the within-subjects Factor A in a repeated measures analysis of variances is then easily obtained from the paired differences among all the means (Winer, 1971, p. 265).

$$SS_A = N\left[\frac{(\bar{y}_2 - \bar{y}_1)^2 + (\bar{y}_3 - \bar{y}_2)^2 + (\bar{y}_3 - \bar{y}_1)^2}{3}\right]$$

Because the contrast factors from the structural modeling approach give information about paired differences, their means can be used to calculate the sum of squares for the repeated measures effect. For three time points, the two contrast factor means, α_2 and α_3, can be substituted into the formula above to obtain the sum of squares,

$$SS_A = N\left[\frac{\left(\alpha_2^2 + (\alpha_3 - \alpha_2)^2 + \alpha_3^2\right)}{3}\right] \tag{3.5}$$

The mean square MS_A is computed by dividing the sum of squares by degrees of freedom, $SS_A/(a-1)$, with *a* representing the number of repeated measurements (see Voelkle, 2007, for an alternative approach).

The ANOVA is the ratio of the mean square effect to the mean square error, $F = MS_A / MS_{\text{error}}$. When the measurement residuals are estimated and set equal in the contrast coding model, the measurement residual variance, $\theta_{(jj)}$ estimate is equal to the MS_{error}. Thus, a simple computation using values from the model output provides the traditional repeated measures ANOVA. The ANOVA tests will be best approximated when multivariate normality assumptions of SEM have been met or when the Satorra–Bentler scaled chi-square estimate of model fit is used.

Latent Difference Score Model

The specification of the latent difference score model for three or more time points (Figure 3.2c) follows a fairly simple extension of the model for two waves. $T-1$ difference score factors, η_k, are specified at each time point except the first, with the path between the factor and the observed variable at the corresponding time point set equal to 1. Each observed variable, y_t, is predicted by the observed variable at the prior time point, y_{t-1}, with the path set equal to 1. The disturbance ζ_t associated with each endogenous variable is set equal to 0. Correlations among exogenous variables are estimated. The difference factor means could be used to compute the sum of squares in the same manner as shown for the contrast coding model in Equation (3.5). An omnibus test can also be obtained by setting the difference factor means equal to 0, with the fit representing a test of the null hypothesis that all means are equal.[6]

ANOVA Assumptions

There are several statistical assumptions with repeated measures analysis of variance that can be explored with the SEM models describe above. I only mention some of the underlying statistical assumptions as many are standard for OLS-based analyses (e.g., normally distributed errors). For repeated measures ANOVA, error variances are assumed to be equal at each time point. In the context of the contrast coding model, this implies that the measurement residual variances are equal. In modeling notation, $\theta=\theta_{11}=\theta_{22}=\ldots=\theta_{(jj)}$. In addition, the diagonals of the variance–covariance matrix $(k \times k)$ of measurement residuals are assumed to be equal, so that $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = \mathrm{Cov}(\varepsilon_1, \varepsilon_3) = \cdots = \mathrm{Cov}(\varepsilon_j, \varepsilon_{j\circ})$, where $j \neq j\circ$. Together these stipulations constitute the *compound symmetry* assumption. The contrast coding model specification provides one method of testing these assumptions. Likelihood ratio tests comparing models with and without constraints on the measurement residual variances and covariances provide a test of compound symmetry. In practice, most researchers consider the compound symmetry assumption to be too strict and focus on the *sphericity* assumption, which states that the variances of all paired difference scores must be equal. This assumption is also easily assessed with the contrast coding model using constraints on the contrast factor variances. As long as the model is identified, any of these assumptions can be relaxed, which allows considerable flexibility compared with conventional ANOVA.

MANOVA

The structural modeling approach most closely resembles the MANOVA repeated measures analysis, because it makes fewer assumptions about the variance–covariance structure. MANOVA works well with moderate and larger sample sizes (Algina & Kesselman, 1997), and, given sufficient sample sizes, tends to have greater power. The MANOVA procedure from some conventional software programs reports a multivariate *F*-value that differs from the univariate *F*-value. The degrees of freedom for the two types of statistical tests are the same, but their values may differ depending on the degree to which assumptions, such as the sphericity assumption, are violated. If variances and covariances are freely estimated in a model constraining all of the means to be equal, the χ^2 will be approximately equal to $(df)(F)$, where df is equal to the numerator degrees of freedom (i.e., $T-1$). This approximation follows the normalizing function (e.g., Sankaran, 1963) that accounts for the discrepancy between the two statistics when $df > 1$. The discrepancy occurs because the expected value of chi-square is equal to df, whereas the expected value of F is equal to 1.

Multivariate *F*-values are sometimes printed in the output for MANOVA procedures in some conventional software, but all programs print several other multivariate statistics, such as Wilks's lambda. Wilks's lambda is perhaps the most widely used MANOVA test statistic, and it is informative to relate the chi-square test from the SEM approach to this conventional statistic. Because the sampling distribution associated with Wilks's lambda is complex, a conversion to Bartlett's chi-square provides an exact estimate of significance using the *F*-distribution for a pure within-subjects design (Tatsuoka, 1988).[7]

$$\chi^2_{\text{Bartlett's}} = -\left[N - 1 - (T+1)/2\right] \ln \Lambda_{\text{Wilks's}}$$

ln is the natural log, or $\log_e$. From this formula, we can derive an approximation of Wilks's lambda, $\Lambda_{\text{Wilks's}}$, from the model chi-square by substituting its value for the chi-square for Bartlett's chi-square. With some algebraic manipulation, we can work backwards to derive an approximation of Wilks's lambda.

$$\Lambda_{\text{Wilks's}} \approx \exp\left[\frac{\Delta\chi^2}{-\left[N - 1 - (T+1)/2\right]}\right] \tag{3.6}$$

The approximate equivalence illustrates the link between the SEM ANOVA model and conventional MANOVA, enabling the computation of several other multivariate statistics from Wilks's lambda.

Trend Analysis

The codes for the contrast coding model can be modified to specify trend contrasts. For a three-wave design, codes of –1, 0, +1, and 1, 0, 1 provide tests of linear and quadratic trends, for instance. This can be implemented by substituting these codes for the dummy or effect coding schemes described above.

$$y_1 = \eta_1(1) + \eta_2(-1) + \eta_3(1)$$
$$y_2 = \eta_1(1) + \eta_2(0) + \eta_3(0)$$
$$y_3 = \eta_1(1) + \eta_2(1) + \eta_3(1)$$

Or, in other words, the second and third columns of the loading matrix, $\boldsymbol{\Lambda}$, for y_t regressed on η_k represent the linear and quadratic contrasts, respectively.

Of course, the linear trend test in the three-wave case is also equivalent to the contrast of the first and third means. With additional waves, higher-order polynomial trends can be tested. Polynomial contrast codes for higher-orders can be obtained from many texts on ANOVA (e.g., Keppel, 1991).

Example 3.4: Mean Comparisons of Observed Means for Three Waves

Comparison of three observed repeated measures was demonstrated using the positive affect measures over the first three waves from the social exchanges data set ($N = 574$). Each measurement, separated by six months, was the average of the five items used in Example 3.1. The observed means of the composite measure were 3.034, 2.925, 2.825, suggesting a small decline in positive affect over time. The repeated measures ANOVA comparing the three means was significant with the univariate and multivariate tests, univariate $F(2,572) = 29.45$, multivariate $F(2,572) = 30.14$, Wilks's lambda = .905, indicating that the differences among the positive affect means was unlikely to be due to chance.

The equality constraint model can be used to compare nested models with one model freely estimating means at each time point and one model constraining the means to be

equal over time (Figure 3.2a). The model with means freely estimated is just identified, so, in this case, the chi-square for model with means constrained to be equal provides the test of the hypothesis, $\chi^2(2)=57.507$, $p<.001$. Results indicated that there was a significant difference among the means. Note that the chi-square value from this model was approximately equal to the univariate and multivariate *F*-value after the normalizing transformation is taken into account, $\chi^2/df=57.507/2=28.754$. Computation of Wilks's lambda using the chi-square for the model, 57.507, in Equation (3.6) was .904, also similar to that obtained with conventional software.

The contrast coding model, depicted in Figure 3.2b, provides another test of the equivalence of the positive affect means across the three waves. This model specified a latent intercept factor with all loadings set equal to 1 and two contrast factors with dummy coded loadings of 0, 1, 0, and 0, 0, 1. To identify the factors, the measurement intercepts and residuals were set equal to 0 at each of the three time points. Factor, means, variances, and covariances were estimated. The mean of the intercept factor represents the mean at the first wave, the mean of the first contrast factor represents the mean difference between Time 1 and Time 2 ($\bar{y}_2-\bar{y}_1$), and the mean of the second contrast factor represents the mean difference between Time 1 and Time 3 ($\bar{y}_3-\bar{y}_1$). This set of mean comparisons was arbitrary and other comparisons could have been chosen instead.

The model as specified was just identified. Constraining the two contrast factor means to be equal to 0 provided an overall test of whether the three means were equivalent. The chi-square of the mean-constrained model was significant, $\chi^2(2)=57.507$, $p<.001$, and was identical to the factor mean constraint analysis. The means for the two contrast factors were –.109, $p<.001$, and –.209, $p<.001$, representing the difference between the observed means, $\bar{y}_2-\bar{y}_1$ and $\bar{y}_3-\bar{y}_1$, respectively.[8]

The coding scheme for this model can be modified to obtain a test of linear and quadratic trends, where the loadings for η_2 are set to –1, 0, 1 and the loadings for η_3 are set to 1, 0, 1 for a test of the two respective trends. The model is just identified, and when the contrast factor means were set equal to 0 for the omnibus means test, chi-square was identical to that obtained with the dummy coded example, $\chi^2(2)=57.507$. Mean of the intercept factor, α_1, was 2.925 and matches the positive affect mean at Time 2. The significance test for α_2 and α_3 represent tests of the two trends. The *z*-value for the mean of the linear trend factor (α_2) was –7.776, $p<.001$, which had a squared value, $-7.776^2=60.466$, similar to the *F*-value for the ANOVA linear trend test, $F(1,573)=60.36$, $p<.001$. The *z*-value for the mean of the quadratic trend variable (α_3) was .183 ($.183^2=.033$), also approximately equal to the ANOVA quadratic trend test, $F(1,573)=.03$.

A latent difference score model was tested to demonstrate an additional method of testing difference across the three waves (Figure 3.2c). The fit of the model with both the difference factor means constrained to be equal to 0 was significant, $\chi^2(2)=57.507$, $p<.001$, providing an omnibus test of mean differences. As expected, this value was commensurate with the chi-square values obtained with the other models. The just identified model freely estimated the difference factor means, giving the average intercept, first difference, and second difference of 3.034, –.109, and –.100, respectively. These values match the observed means and mean differences.

Repeated Measures with More than Two Observed Binary or Ordinal Variables

The models described above for binary and ordinal variables extend readily to more than two waves. The nested tests of means from three or more correlated factors with a single binary or ordinal indicator is specified identically to the model test of two waves, except

that means for three or more waves are constrained to be equal, providing an omnibus test of mean differences. As with models for continuous variables measured at three or more time points, subsets of means can be constrained for planned contrasts. The same results can be obtained via the contrast coding specification using $T-1$ contrast factors. The equality constraint, contrast coding, or latent difference score models test the same hypotheses as conventional statistical tests, such as Cochran–Mantel–Haenszel or loglinear models. The two loadings of the contrast coding model also can be modified to test for linear, quadratic, or higher-level trends, depending on the number of time points.

Repeated Measures Analysis of Two or More Latent Variables

The basic concept for mean comparisons can be applied to latent variables with multiple indicators by estimating factor means. Figure 3.3 illustrates three ANOVA models for latent variables with multiple indicators at two time points, but each could be extended to three or more time points. The details of the model specifications for each type of model are the same as for models with observed variables, except that a latent variable is specified for each occasion and factor means are compared instead of observed means. Because expected values are not affected by measurement error, latent means simply reflect some combination of their observed indicators, depending on which specification is used to identify the factor mean.

There are three methods of factor mean identification that are relevant for ANOVA models with latent variables – referent intercept identification, single occasion identification, and effects coding identification. The interpretation of the factor means depends on the identification approach, so it will be worthwhile to give some careful thought about how the latent variable mean is defined under each. The following sections thus closely examine identification strategies to provide a more in-depth understanding of the interpretation of the factor means. A general review of factor identification can be found in Chapter 1.

Choice of identification approach is arbitrary in terms of model fit, however, and likelihood ratio tests of mean differences will be equivalent across different identification methods. This equivalence holds only when means are compared in tandem with longitudinal equality constraints on loadings and measurement intercepts. Initial tests of measurement invariance assumptions are usually recommended before proceeding with further modeling (Vandenberg & Lance, 2000; for further details, see Chapter 2), and I will assume a suitable level of invariance has been established for the examples presented in this chapter. If measurement invariance holds in the population, then variation of the measurement parameters across occasions in the sample should have a trivial impact. If measurement invariance does not hold in the population, use of equality constraints in the model may lead to biased estimates and should not be used.

Referent Intercept Identification

With the referent loading and measurement intercept identification of each factor, we can extend any of the ANOVA models discussed thus far to compare factor means over time when each factor has multiple indicators. Figures 3.3a–3.3c parallel the models for single indicators depicted in Figures 3.1.b–3.1d. For each model depicted in the figures, referent identification is shown, setting one intercept and one loading for each factor. In usual practice, the remaining loadings for each indicator would be set equal over time.

Likelihood ratio tests can be used to investigate the assumption of equal variances or covariances across time points. With the equality constraint model, depicted in Figure 3.3a, a test of the compound symmetry assumption could be obtained if there are three or more

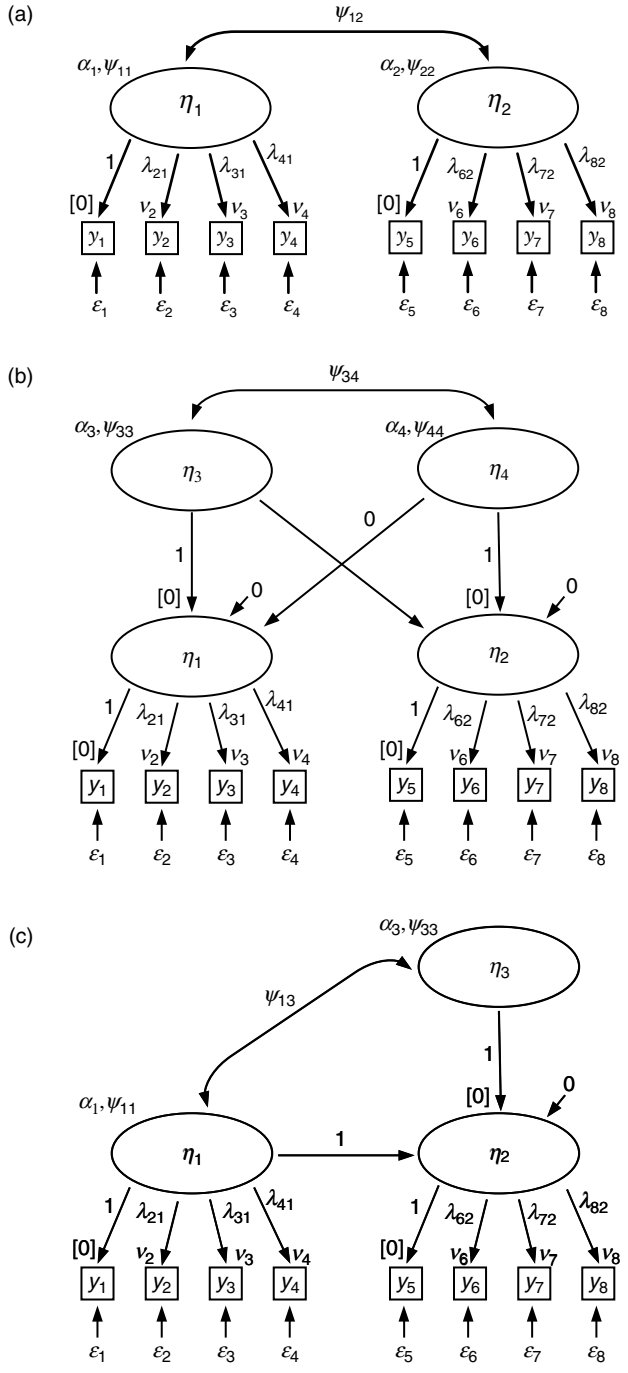

Figure 3.3 Three Approaches to Testing for the Difference between Two Latent Means: (a) equality constraint model; (b) contrast coding model; (c) latent difference model. Note: Greek symbols represent parameters that are estimated in the model, replaced by values when the parameter is set. Values presented in square brackets represent means or intercepts.

time points by imposing longitudinal equality constraints on factor variances and factor covariances over time. With the contrast coding model, depicted in Figure 3.3b, disturbance variances for first-order factors represent residual variances conditioned on the intercept and contrast factors and could be compared similarly. In the latent difference score model, however, placing equality constraints on the difference score factors (η_2 and η_3 in Figure 3.3c) would not be equivalent to either of these tests. Because these factors represent variability of the paired differences, comparing them is akin to assessing the sphericity assumption.

Single Occasion Identification

An increasingly common alternative to referent identification for multigroup comparisons is to set the factor mean of one group equal to 0 while freely estimating the factor means in the other groups (e.g., Byrne, 1994; Widaman & Reise, 1997). With this constraint, the mean of the factor in the second group is equal to the group difference between the factor means. The specification can be applied to some longitudinal models, constraining the mean at one occasion (e.g., the first time point) to be equal to 0 while estimating the other factor means (e.g., Widaman, Ferrer, & Conger, 2010). Single occasion identification provides a method of standardizing the factor to have unit variance and mean 0 when it is infeasible to standardize the factor at multiple occasions. Two factor means, for instance, cannot be compared if they are set to 0 for identification purposes. A model setting the factor mean and variance at only one occasion is only identified if longitudinal equality constraints on all measurement intercepts and factor loadings are used, however. With two waves, this identification approach leads to an estimate of the factor mean at the second time point that equals the difference in factor means, as long as the covariance between the factors is freely estimated.[9] The test of the significance of the second factor mean (as compared with 0) provides a statistical test of the difference between the two factor means.

Why Single Occasion Identification Works. But how do we interpret the factor means and, hence, the factor mean difference in this case if there is no obvious scale set for the factor mean estimate at the first time point? To understand what the latent mean difference represents under this model specification, consider a simple model in which we estimate the factor mean at a single time point by setting all loadings equal to 1 and all intercepts equal to 0. We know that, assuming measurement residuals have a mean of zero, the relation between the observed means of the indicators and the factor mean for a single variable is the following (e.g., Bollen, 1989):

$$E\left(y_j\right) = \nu_j + \lambda_{jk}\alpha_k$$

This can be restated in the more general matrix form, which is convenient to explain the logic if more than one latent variable is involved.

$$\bar{\mathbf{Y}} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}$$

where $\bar{\mathbf{Y}}$ is a $J \times 1$ vector of observed means for J observed variables, ν is a $J \times 1$ vector of measurement intercepts, $\boldsymbol{\Lambda}$ is $J \times K$ matrix of all of the loadings, and α is a $K \times 1$ vector of factor means (Appendix A further explains the notation). The factor mean(s) can then be shown to be a function of the factor loadings, the observed means, and the intercept (Mulaik, 2009).

$$\boldsymbol{\alpha} = \left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\left(\bar{\mathbf{Y}} - \boldsymbol{\nu}\right) \tag{3.7}$$

Suppose we have a latent variable with four indicators has all intercept values set equal to 0 and all loadings are set equal to 1. $\boldsymbol{\Lambda}$ then becomes a vector of ones, $\left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}$ is $1/J_k$, and $\left(\boldsymbol{\mu}_y - \boldsymbol{\nu}\right)$ reduces to $\boldsymbol{\mu}_y$. So, we have

$$\boldsymbol{\alpha} = \left(1/J\right)\mathbf{1}'\boldsymbol{\mu}_y$$

which is equivalent to $\alpha_k = \Sigma\mu_j / J$, or the average of the four observed indicator means. It follows that the factor mean in this special case model is an equally weighted average of the means of each of the observed indicators.

Using $\mu_{j,t-1}$ and μ_{jt} to refer to the means for two consecutively repeated measurements, the difference between two factor means with J_k items per factor would be

$$\alpha_t - \alpha_{t-1} = \frac{\sum_{j=1}^{J_{kt}} \mu_{jt}}{J_k} - \frac{\sum_{j=1}^{J_{k,t-1}} \mu_{j,t-1}}{J_{k,t-1}} \tag{3.8}$$

which is simply the difference between the averages of the item means measured at the two time points. Suppose now that the first factor mean is set equal to 0. Then α_1 drops out of Equation (3.8), and the mean of the second factor, α_2, will be equal to the difference between the averages of the observed means at the two time points.

If the loadings were to be freely estimated, the means of the factors would be weighted by the magnitude of the loadings, so that indicators with higher loadings would be weighted more heavily in the estimate of the factor mean.

$$\alpha_k = \frac{\sum\left(\mu_j / \lambda_{jk}\right)}{J}$$

With freely estimated loadings, the difference between two factor means would simply be the difference between the weighted averages of the observed means at each time point, each taking into account their factor loadings. This point is simply didactic, because we cannot identify individual factor means without either setting its value or one of the measurement intercept values equal to 0.[10]

The Role of Equality Constraints. In practice, the interpretation of the factor mean with single occasion identification is more subtle, because longitudinal equality constraints on all loadings and measurement intercepts are employed. Equation (3.7) still holds, of course, but we must take into account the estimate of the measurement intercept for each pair of items. With equality constraints, the intercept is the average of the two observed means for each repeatedly measured variable [e.g., $\nu_1 = \nu_5 = \left(\mu_1 + \mu_2\right)/2$, for two four-indicator factors]. The simple algebraic expression for the factor mean in this case parallels Equation (3.7), with (*j*) and (*jk*) subscripts indicating that parameters are set equal across waves and J_k giving the number of indicators for the factor.

$$\alpha_k = \frac{\sum\left[\left(\mu_j - \nu_{(j)}\right)\right] / \lambda_{(jk)}}{J_k} \tag{3.9}$$

We can use this expression to show the computation of the difference of two factor means, assuming intercepts and loadings differ across time, with the subscript *jt* to refer to the repeated measurement of *jt–1* at the subsequent time point.

$$\alpha_t - \alpha_{t-1} = \frac{\sum\left[\left(\mu_{jt} - \nu_{jt}\right) / \lambda_{jkt}\right]}{J_k} - \frac{\sum\left[\left(\mu_{jt-1} - \nu_{jt-1}\right) / \lambda_{jkt-1}\right]}{J_{k,t-1}}$$

If the intercepts and the loadings are equal across time so that all $\lambda_{jkt} = \lambda_{(jk)}$ and all $\nu_{jt} = \nu_{(j)}$, then a little algebraic manipulation demonstrates a simpler expression of the factor mean difference when measurement intercepts and loadings are held constant across the two time points.

$$\begin{aligned}
\alpha_t - \alpha_{t-1} &= \frac{\sum\left[\left(\mu_{jt} - \nu_{(j)}\right) / \lambda_{(jk)}\right]}{J_k} - \frac{\sum\left[\left(\mu_{jt-1} - \nu_{(j)}\right) / \lambda_{(jk)}\right]}{J_k} \\
&= E\left[\left(\mu_{jt} - \nu_{(j)}\right) / \lambda_{(jk)}\right] - E\left[\left(\mu_{jt-1} - \nu_{(j)}\right) / \lambda_{(jk)}\right] \\
&= E\left[\frac{\left(\mu_{jt} - \nu_{(j)}\right) - \left(\mu_{jt-1} - \nu_{(j)}\right)}{\lambda_{(jk)}}\right] \\
&= E\left[\frac{\mu_{jt} - \mu_{jt-1}}{\lambda_{(jk)}}\right]
\end{aligned}$$

In the above equations, expectations are taken over the items rather than cases. Thus, the average difference between each pair of observed means divided by their respective factor loadings, $E\left[(\mu_{jt} - \mu_{jt-1}) / \lambda_{(jk)}\right]$, gives the difference between factor means. If the first factor mean, α_1, is set to 0 for identification, then the estimate for the second factor mean, α_2, becomes the difference between the factor means.[11] Consequently, the test of the significance of the factor mean at Time 2 using single occasion identification (i.e., H_0: $\alpha_2 = 0$), provides yet another method of testing the mean difference over time.

More than Two Time Points. Single occasion identification can be generalized beyond the two time points discussed here, with similar implications for interpretation of the mean differences. Instead of factor means that depend on loadings and intercepts averaged across two time points, however, averages are taken across the several time points. The factor means for each time point other than the referent time point, (e.g., α_2 through α_T), represent a difference from the referent factor mean. Tests of differences among several factor means are tested by constraining factor means for all factors, except the referent factor, equal to 0. The chi-square of the constrained model is compared to the chi-square of the model with the factor means freely estimated.

Effects Coding Identification

Effects coding identification (Little, Slegers, & Card, 2006) uses complex equality constraints on factor loadings and measurement intercepts. Instead of setting parameters to particular values as with referent and single occasion identification, effects coding identifies the factor by imposing a constraint on each parameter from a factor so that the loading or the intercept must be estimated in relation to loading or the intercept for the other items for the factor. The constraints produce an estimate of the factor variance that is

equal to a weighted average of the observed variances of the indicators and an estimate for the factor mean that is equal to a weighted average of the observed means of the indicators (refer to Equations [1.15]and [1.16]). The factor mean is

$$\alpha_t = \frac{\sum\left(\mu_j - \nu_j\right) / \lambda_{jk}}{J}$$

for a single occasion. When equality constraints are imposed over time, the observed means are weighted by common estimates of the intercepts and loadings.

$$\alpha_t = \frac{\sum\left(\mu_j - \nu_{(j)}\right) / \lambda_{(jk)}}{J}$$

Effects coding identification thus provides a similar weighted interpretation of the factor mean and variance to that obtained with single occasion identification. With equality constraints, they have the same fit and equivalent likelihood ratio tests of factor mean differences. The advantage is that factor means can be estimated at each wave without reference to the baseline mean.

Example 3.5 Repeated Measures Analysis of Two latent Variables

The examples below illustrate mean comparison tests conducted by constraining factor means to be equal. The referent coding model and latent difference model can also be used to compare factor means using either referent or effects coding identification. Software examples of each model are available on the website for the book.

Comparison to Conventional ANOVA. I first illustrate a test of the difference between two repeated latent factor means using an equal weighting of the positive affect items used in the previous examples (N = 574). The purpose of this analysis was simply to demonstrate that a latent variable model specified with certain constraints will produce results equivalent to those obtained from traditional repeated measures ANOVA of an equally weighted composite index. This model specification would generally not be used in practice, however. Five individual items served as indicators of a single factor at each of the two time points. Each of the loadings was set equal to 1; all intercepts were set equal to 0; all measurement residuals for each factor were set equal to one another; both factor variances were constrained to be equal; and both factor means were freely estimated. This model did not have an ideal fit as a result of the constraints on the loadings, $\chi^2(59)=152.206$, $p < .001$, CFI=.952, SRMR=.097. The estimated factor means were 3.034 and 2.925 for Time 1 and Time 2, respectively. These values are equal to the observed means for the composite average measure for each time point. A model restricting the factor means to be equal over time had a larger chi-square value, $\chi^2(60)=167.279$, which was a significantly poorer fit than when the factor means were allowed to differ from one another, $\Delta\chi^2(1)=15.073$, $p < .001$. This chi-square difference is a close match to the result obtained in Example 3.1 using ANOVA and structural modeling comparisons of the composite means. The contrast coding and latent difference score models could also be used to make this comparison, and would produce identical results when using equally weighted loadings and intercepts. The results from this example show that when factor means are compared, even when measurement residuals are estimated to remove measurement error, the results are equal to those obtained with a composite measure or traditional ANOVA. Measurement error does not affect the mean difference.

Referent Identification. A subsequent model was tested to illustrate referent intercept identification. A model comparison of means for the positive affect factor used the "happy" item as the referent indicator, setting all other loadings and measurement intercepts equal over time. This model produced factor mean estimates of 3.051 and 2.938 for the two time points. The factor means were similar but not identical to the observed means for the referent item, because equality constraints on loadings and intercepts were used. When equality constraints were removed, the factor means at the two time points, 3.045 and 2.945, were identical to the observed means for the "happy" item. A likelihood ratio test comparing the fit of models with and without longitudinal equality constraints of factor means indicated a significant difference, $\chi^2(44)=118.719$, $\chi^2(42)=103.653$, $\Delta\chi^2(1)=15.066$, $p < .001$. Had loadings and intercepts been exactly invariant, these results would have been identical to the results from the previous model. Thus, the discrepancy between the mean estimates and the nested tests for the comparisons conducted with and without the equality constraints illustrates the potential impact on the analyses if the equality constraints are imposed when invariance assumption is not correct. To the extent that the parameters differ in fairly trivial magnitudes, equality constraints should have little practical effect. To the extent that the parameters differ more substantially, but equality constraints are imposed, biases in the estimate of the factor mean comparisons will be introduced.

Single Occasion Identification. I next conducted the analysis using single occasion identification in which the factor mean at the first time point is set equal to 0 and the factor mean at the second or subsequent time points are freely estimated. To test differences in the positive affect mean, the mean of the factor at Time 1 was set equal to 0 while the variance of the factor was set equal to 1. Loadings and measurement intercepts for all pairs of repeated measured variables were constrained to be equal over time. The model chi-square, $\chi^2(42)=103.653$, was the same value as in the previous model using referent indicator identification. The mean of the second factor was significantly different from 0, $\alpha_2=-.192$, $z=3.868$, $p < .001$, indicating there was a difference between the two factor means. The likelihood ratio test, obtained from constraining the second factor mean to be equal to 0, provides a test of mean differences, $\chi^2(43)=118.719$, $\Delta\chi^2(1)=15.066$, $p < .001$, was equal to the likelihood ratio test obtained from the model using referent identification. The Wald test of the second factor mean obtained with the single occasion model also tests this difference, $z=3.868$, $p < .001$. Note that squaring this value is approximately equal to the likelihood ratio chi-square, $z^2=(3.868)^2=14.961$. Although the mean difference from this model is larger than the difference in the means obtained in the referent identification model ($2.938-3.051=-.113$), the likelihood ratio tests are identical in the two models because the factor means must be evaluated relative to their variances, which are not the same under the two identification approaches. The magnitudes of the differences in the means are equal, taking into account the factor variances in each case.

Effects Coding Identification. To illustrate effects coding identification, a nested test of factor mean equivalence was conducted. A complex equality constraint was imposed by requiring that the first loading be equal to $5-\lambda_{(2)}-\lambda_{(3)}-\lambda_{(4)}-\lambda_{(5)}$ and the first intercept be equal to $0-\nu_{(2)}-\nu_{(3)}-\nu_{(4)}-\nu_{(5)}$. The parenthesis subscripts indicate parameters that have been set equal over time. These constraints produce factor mean and factor variances values that are weighted averages of the observed means and variances for the indicators, respectively. Subscripts in parentheses indicate that a common value for each parameter was estimated for each of the two measurement time points. For the model in which the factor means were allowed to differ, the mean for positive affect factor at Time 1 was equal to

3.034 and the mean at Time 2 was equal to 2.925. Factor means were then compared via a nested test with their values constrained to be equal in one model but not in the other. The model chi-square and chi-square difference values were the same as those obtained with the referent indicator and single occasion models above, $\chi^2(42)=103.653$ vs. $\chi^2(43)=118.719$, $\Delta\chi^2(1)=15.066$, $p < .001$.

Comments

These examples illustrate that ANOVA models can be estimated using latent variables at each time point. Although results using each of the ANOVA models shown in Figure 3.3 were not reported, each of these modeling approaches may also be used with latent variables. If the indicators are equally weighted in specifying the factor mean, then the results are highly comparable to the conventional ANOVA with an equally weighted composite. This is because measurement error does not affect mean estimates. When loadings are estimated, the factor mean represents a weighted average of the observed variables. Without longitudinal equality constraints on loadings and intercepts, this weighted average is defined considerably differently depending on the identification method. With longitudinal equality constraints, the factor means under each identification approach represent algebraically equivalent weighted averages of the indicators. Although a weighted average of the indicators may have added validity compared with an equally weighted average, values derived from a weighted index will tend to be highly correlated with an equally weighted index, particularly when observed means and loadings do not vary substantially across indicators for the factor (Fava & Velicer, 1992). There are a number of potential advantages of incorporating latent variables, however, including measurement equality constraints, relaxation of repeated measures assumptions, correlated measurement residuals, and extensions to more complex models. Hancock (2003) also shows that latent mean comparisons may have a power advantage over traditional ANOVA, because the standardized effect size tends to be larger.

As evident from the examples, model fit and likelihood ratio tests of factor means will not differ by identification approach as long as longitudinal equality constraints on measurement parameters are included in the model. This makes the identification approach an arbitrary choice. Nonetheless, there may be some intuitive appeal to the factor mean interpretation under single occasion and the effects coding identification compared to referent identification, because the factor mean can be viewed as a weighted composite of observed variables. Factor means under referent identification, in contrast, are closely tied to the referent variable.

Repeated Measures Analysis of Latent Variables with Binary and Ordinal Indicators

Comparison of three or more latent variables with binary or ordinal indicators follows the same procedures described for comparison of three or more means with continuous indicators (as in Figures 3.3a–3.3c). Factor means or differences in factor means, however, must be transformed to approximate the observed proportions. Similar options exist for identification. With referent identification, in which the first threshold of one indicator is set equal to 0, the factor means represent proportions of the item used as referent. The other methods will produce a weighted average of the observed proportions. As with continuous indicators, longitudinal constraints on loadings and thresholds would typically be imposed. When all loadings are freely estimated, the factor means will correspond closely with observed proportions of the first response category of the referent variable

once values are transformed using the appropriate cdf. Imposing equality constraints on loadings and thresholds, of course, leads to less correspondence between observed proportions and factor means because factor mean estimates are affected by the equality constraints.

Example 3.6: Comparison of Two Latent Variables with Ordinal Indicators

To briefly illustrate the relation between proportions for measured ordinal variables and the estimated factor mean, I examined a multiple indicator model using four items from the depression measure from the social exchanges data set. The items "bothered by things," "felt sad," "had the blues," and "felt depressed," with responses on a 4-point frequency rating, loaded on a general depression factor at Time 1 and Time 2. Means of both factors were estimated using WLSMV estimation for ordinal variables with theta parameterization, setting thresholds and loadings equal across time points. Delta parameterization or ordinal ML estimation also would be possible. Effects coding identification was used for factor variances and means, but any of the identification approaches could be used. The model had good fit to the data, with $\chi^2(33)=60.209$, $p < .001$, CFI=.983, Weighted Root Mean Square Residual (WRMR)=.925. The factor mean estimates for the depression factor at the two time points were –.382 and –.477. Using the normal cdf transformation of these values gave an estimation of the proportions of the number endorsing the first response category ("none of the time") as .649 and .683, which roughly approximate the average observed proportions of .610 and .616, respectively. Constraining the means to be equal resulted in a chi-square value that was not significantly larger of $\chi^2(34)=61.545$, $\Delta\chi^2(1)=1.33$, ns, using the weighted chi-square test, indicating that there was no significant change in the depression factor mean over the two time points.

Mixed Factorial Models: Between and Within-Subjects Factors

The ANOVA model which combines a between-subjects factor and a within-subjects factor is typically referred to as a *mixed factorial ANOVA*. It is a simple matter to build upon the previously discussed within-subjects models by adding a between-subjects grouping factor. Such a factor might include sociodemographic groups, such as gender and race, or treatment–control comparisons from an experimental or quasi-experimental design. In SEM, there are two general strategies for including such grouping variables – multigroup comparisons and the MIMIC model. Multigroup modeling involves a comparison of nested tests in which parameters are constrained across groups (see, for example, Bollen, 1989 or Kline, 2010, for an introduction). The MIMIC approach involves a single-group analysis with a binary predictor or set of orthogonal binary coded predictors to represent multiple groups, with the path coefficient between the grouping variable and the dependent variable representing the differences among groups. The MIMIC model can better control for confounding factors across groups by including covariates, whereas the multigroup approach only controls for covariate variation within group (Newsom, Prigerson, Schulz, & Reynolds, 2003). On the other hand, the MIMIC strategy assumes a structural invariance of the outcome variable across groups, because only one measurement model common to the groups is involved. MIMIC and multigroup models can produce different results when structural invariance is not met, the model is misspecified, distribution assumptions are not met, or there are unequal variances between groups (Hancock, 2001; Yuan & Bentler, 2006). Thus, the MIMIC model should be used with these precautions in mind and with preliminary establishment of measurement invariance.

The relation of multigroup or MIMIC models to ANOVA has nearly always been discussed in the context of cross-sectional designs and rarely discussed in the context of longitudinal designs or mixed ANOVA. Here I show how mixed within- and between-subjects designs can be investigated with SEM, focusing initially on the simplest case with a single-indicator at only two time points and in only two groups, but the models depicted in Figures 3.1, 3.2, or 3.3 can be readily extended to include more than two groups or time points.

Multigroup Strategy

Let us first consider the multigroup strategy with a single observed variable for each time point. If the model quantifies the difference score, as with the models shown in Figures 3.1a, 3.1c, or 3.1d, an interaction test equivalent to one obtained with a mixed factorial ANOVA is possible. There is an interaction to the extent that the mean difference is not equivalent across groups. An alternative conceptualization of the interaction considers whether there is a difference between groups at Time 1 is of lesser or greater magnitude than the difference between groups at Time 2. The latter interpretation, which cannot be explicitly investigated with the difference score model in Figure 3.1a, is especially of interest for comparing treatment groups in a pretest–posttest design, for example. Of course, the two alternatives for conceptualizing the interaction hypothesis are equivalent.

If the model estimates both Time 1 and Time 2 as in Figure 3.1b (equality constraint model with measured variables), however, there is no simple method of testing the interaction effect using either a multigroup or MIMIC strategy. Software programs that allow testing of complex constraints can be used to specify that the difference between factor means is equivalent in the two groups, although complex constraints are not currently available in some programs. For multiple indicator models that use single occasion identification in which the factor mean at Time 1 is set equal to 0 (Figure 3.3a), the interaction can be tested with a multigroup model, however, because the second factor mean represents the Time 1 – Time 2 mean difference.

MIMIC Strategy

The MIMIC modeling approach, in which a difference factor on a binary grouping variable (e.g., gender) is used to predict a factor representing a difference score, also provides a method of testing a mixed factorial interaction. Figure 3.4 illustrates some of the possible models. For any model which quantifies the repeated measures difference, such as a model with a pre-computed difference score (Figure 3.4a), a multiple-indicator model using the single occasion identification (Figure 3.4b), the contrast coding model (Figures 3.4c), or the latent difference score model (Figures 3.4d), a test of the interaction can be conducted by examining the effect of the grouping variable on the difference factor. Except for a model using a pre-computed difference score (Figure 3.4a) or the equality constraint model to identification (Figure 3.4b), this interaction test may be based on single-indicator or multiple-indicator models. Repeated measures main effects would need to be tested by respecifying the model without the binary predictor variable. Between-subjects main effects are less conveniently tested with a MIMIC model without respecifying the model so that an average mean across the time points could be obtained by using equality constraints on the factor means.

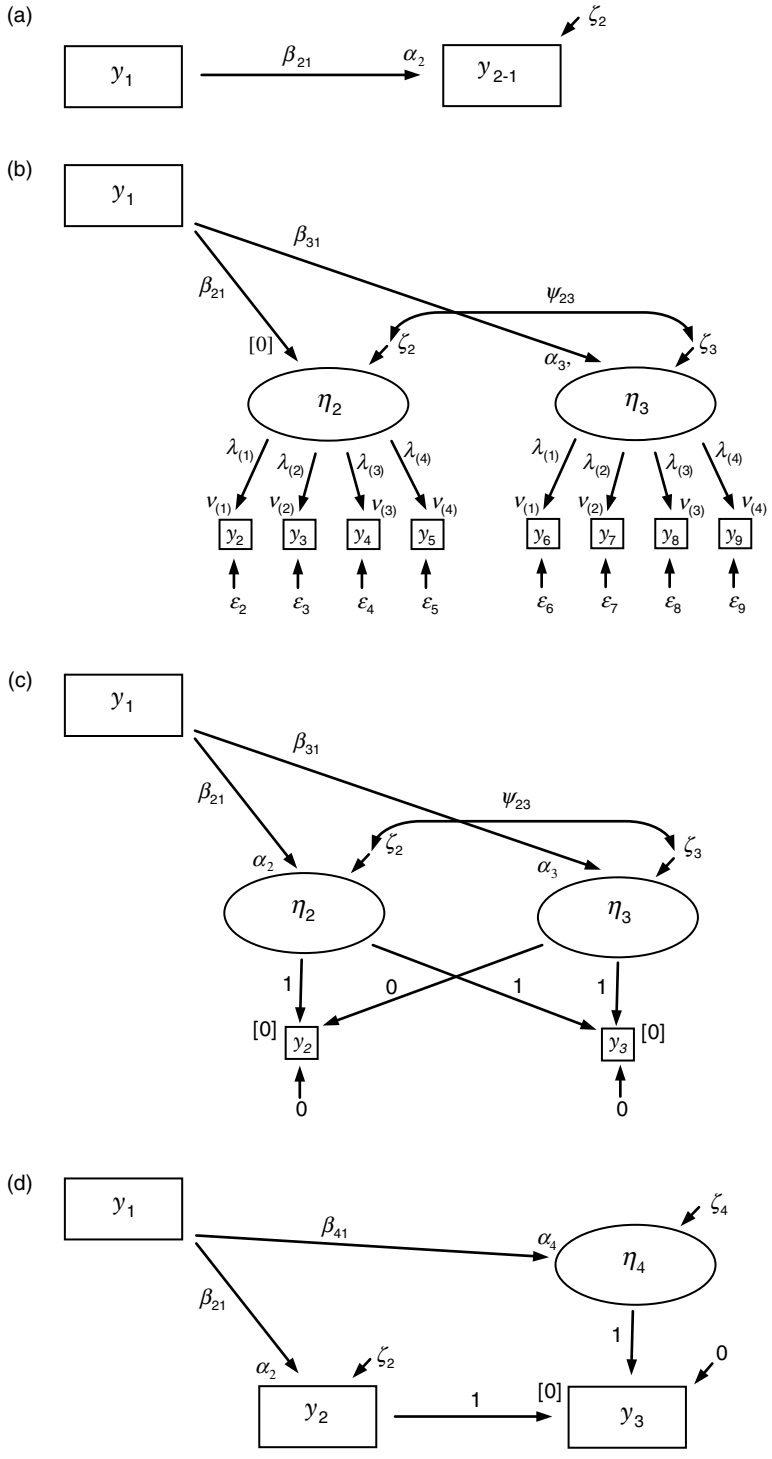

Figure 3.4 Four Examples of the MIMIC Approach to the Mixed Factorial ANOVA: (a) observed difference scores as outcome; (b) equality constraint model; (c) contrast coding model; (d) latent difference score model.

Relation to the ANOVA Model

To understand the connection to the mixed effects ANOVA, consider a MIMIC model, as illustrated in Figure 3.4c, involving a contrast factor with two observed variables that have loadings equal to –1 and +1. This model is given by the two equations below, where, y_1 is the binary grouping variable, η_2 is the intercept factor, and η_3 is the difference factor (given as η_2 in Equation [3.2]). We can then describe the model with the following equations:

$$\eta_2 = \alpha_2 + \beta_{21} y_1 + \zeta_2$$
$$\eta_3 = \alpha_3 + \beta_{31} y_1 + \zeta_3$$

Using this effect coding scheme, the intercept, α_2, will represent the average across groups, combining Time 1 and Time 2, or, in other words, the grand mean. The path coefficient, β_{21}, is the average difference of each group mean from the grand mean, and provides a test of the between-subjects main effect. The intercept, α_3, represents the average difference between Time 1 and Time 2, combining the two groups, which gives a test of the within-subjects main effect. The path coefficient for the difference factor regressed on the grouping variable, β_{31}, quantifies the interaction effect. These values provide all the information necessary for the mixed factorial ANOVA model (e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996) given below.

$$E\left(Y_{ijk}\right) = \mu + \tau_{\alpha j} + \tau_{\beta k} + \left(\tau_\alpha \tau_\beta\right)_{jk}$$

Using standard ANOVA notation in this equation, μ is the grand mean, $\tau_{\alpha j}$ is the deviation of the between-subjects mean from the grand mean, $\tau_{\beta k}$ is the deviation of the within-subjects mean from the grand mean, $(\tau_\alpha \tau_\beta)_{jk}$ is the interaction between the two factors, and subscripts *i*, *j*, and *k* represent individual, group, and repeated measurement indexes, respectively.

The coding scheme is unimportant for the significance test of the interaction, but the interpretation of the main effects would differ for an alternative coding scheme. With a dummy coding scheme, for example, α_2 and α_3 would represent the mean for group 0 and the difference between Time 1 and Time 2 for group 0, respectively, because the intercept represents the value of the outcome when the predictor is equal to 0. The mixed between- and within-subjects factorial ANOVA can be viewed as an analysis that defines changes as a difference score, because the repeated measures mean difference test is the same as a test of the hypothesis that the average difference score is different from zero. This definition of change differs from the definition implied by analysis of covariance (ANCOVA), which defines change as residual from an autoregression. The latter conceptualization of change is discussed at length in Chapter 4.

Main Effects

In the absence of a significant interaction, main effects for overall differences between groups or over time might be of interest. The specifications necessary to test the main effects with multigroup modeling differ depending on the particular repeated measures model. Main effects cannot be investigated in the single difference score model (Figure 3.1a). For the models in which factor means at Time 1 and Time 2 are estimated (Figures 3.1b and 3.3a), the between-subjects main effect, can be tested by holding the factor means to be equal over time while conducting a nested test of their equivalence across groups. This test, as in any between-subjects main effects test for two groups, is analogous to the

independent samples *t*-test with two groups if an average of the Time 1 and Time 2 measurements serves as the dependent variable. The repeated measures main effect, comparing the mean difference over time, can be tested by constraining the Time 1 factor means to be equal across the two groups while also constraining the Time 2 factor means to be equal across the two groups.

For the contrast coding model or the latent difference model, it is also possible to conduct main effects tests without changing the basic structure of the model.[12] The within-subjects main effect can be tested by comparing a model in which the contrast mean (η_3 in Figures 3.3b and 3.3c) is constrained to be equal to 0 in both groups. The test of the between-subjects main effect can be conducted by comparing the model with just the contrast factor means constrained to be equal across groups to a model in which both the contrast factor and the intercept factor means are constrained to be equal across groups.

Example 3.7: Mixed Factorial ANOVA Models

Marital status has often been linked to life satisfaction and emotional health. To investigate whether changes in positive affect over time differed for those who are married or unmarried, I conducted a mixed factorial analysis with one between-subjects factor (married vs. unmarried) and one within-subjects factor (Time 1 vs. Time 2 positive affect) using the social exchanges data set ($N=574$, with 284 unmarried and 290 married). The means for the 2×2 design are presented in Table 3.1. Results from the traditional mixed factorial ANOVA indicated that the main effect for marital status was significant, $F(1,572)=8.81$, $p < .01$, the main effect for Time was significant, $F(1,572)=15.43$, $p < .001$, and the Marital Status×Time interaction did not reach statistical significance, $F(1,572)=2.96$, $p=.086$. Thus, it appears that the change in positive affect does not strongly depend on marital status.[13]

This interaction model using measured variables cannot be tested using multigroup analysis and standard equality constraints, because the interaction test involves the test that the mean difference for the repeated measures factor (i.e., $\mu_2 - \mu_1 = 0$ in the population) is not equal in the two groups. It is possible to perform this test using complex equality constraints in some computer software programs. The method requires the computation of a constraint defined by the Time 2 – Time 1 difference in each group and then the specification that these constraints are equal across groups. Similarly, complex constraints can also be used to test the main effects, in which the levels of the between-subjects factor (i.e., marital status) are averaged and the within-subjects levels (i.e., Time 1 and Time 2 positive affect) are averaged, and then these averages are constrained to be equal across groups. The result of the test of these constraints indicated that the main effect for marital status was significant, $z=2.975$, $p < .01$, that the main effect for time was significant, $z=3.935$, $p < .001$, and that the interaction did not reach statistical significance, $z=1.724$, $p=.085$. Note that the squares of each of these values, $z^2=(2.975)^2=8.851$, $z^2=(3.935)^2=15.484$, and $z^2=(1.724)^2=2.972$, closely follow the *F*-values from the factorial ANOVA within rounding error.

A multigroup model can also be used to test the interaction using the contrast coding model. No special constraint features are required because the contrast factor represents the difference between Time 1 and Time 2 positive affect within each group. To test the interaction, the fit of the models with and without equality constraints across groups on the contrast factors were compared. The unconstrained model was just identified. The estimate of the mean difference in the unmarried group was –.061, $z=1.560$, ns, and in the married group was –.092, $z=1.557$, ns. These values correspond with the observed differences in the two groups, and their significance tests represent tests of simple effects of the within-subjects variable for the two levels of the between-subjects factor, marital status. The simple effects tests would ordinarily not be examined without a significant

Table 3.1 Means for the Equally Weighted Composite Measure of Positive Affect

Marital status	*Positive affect*		
	Time 1	Time 2	Time 2 – Time 1
Not Married	2.943	2.881	–0.062
Married	3.127	2.970	–0.157
Married – Not Married	0.184	0.089	–0.095

interaction. When the estimates of the mean of the contrast factor were constrained to be equal across unmarried and married groups, the chi-square was only marginally significant, $\chi^2(1)=2.963$, $p=.085$. This value also is approximately equal to the *F*-value for the interaction from the ANOVA.

The within-subjects main effect was tested by comparing a model in which the contrast factor mean was constrained to be equal to 0 in both groups, $\chi^2(2)=18.078$, to a model in which the contrast factor mean was estimated but constrained to be equal across groups, $\chi^2(1)=2.963$. The difference in fit was 15.115, which also was similar to the ANOVA result. A test of the between-subjects main effect was conducted by comparing the model with just the contrast factor means constrained to be equal across groups, $\chi^2(1)=2.963$ (same as the above model), to a model in which both the contrast factor and the intercept factor means are constrained to be equal across groups, $\chi^2(2)=12.434$. The difference, $\Delta\chi^2=9.471$, is similar to the *F*-value from the ANOVA.

We can conduct an analogous mixed factorial with multiple-indicator latent variables using several strategies. Just as with measured variables, latent variable means within groups or across groups can be compared, but it is difficult to test the interaction without the use of complex equality constraints that are not universally available. Single occasion identification provides one method of testing an interaction, however, because the second factor mean represents the mean difference for the repeated measures. For a multigroup comparison of the model shown in Figure 3.3a, I set the first factor mean equal to 0 as well as all loadings and measurement intercepts for the same indicators equal over time and across groups. Measurement residuals were allowed to be uniquely estimated across time points and groups. The covariance between factors was estimated, a necessary condition for the second factor to have the mean difference interpretation.

The mean of the second factor, representing the difference between the means, was –.129 in the not married group and –.169 in the married group. When the second mean was allowed to differ across groups, the overall chi-square, $\chi^2(96)=192.881$, was smaller than when the second mean was constrained to be equal across groups, $\chi^2(97)=197.884$. This difference was significant, $\Delta\chi^2(1)=5.003$, $p < .05$, indicating an interaction effect. This value differs slightly from the *F*-value in the ANOVA and model tests of the composite variable, because the same mean values are not being compared. The interaction test in this model involves factor means that take into account the loading and intercept equality constraints imposed across waves and across groups.

An alternative method of testing a mixed factorial interaction with multiple indicators is a multigroup comparison using the contrast coding model (a multigroup comparison of the model in Figure 3.3b). The referent identification method was used for intercepts and loadings, and all other pairs of intercepts and loadings were constrained to be equal over time and across groups. The chi square for the model allowing the contrast factor mean to vary across groups, $\chi^2(92)=174.068$, was only marginally lower than when the contrast factor means were constrained to be equal across groups, $\chi^2(93)=176.990$, $\Delta\chi^2(1)=2.922$, $p=.087$. Thus, there was no reliable evidence that the change in positive affect over time

differed by marital status. A MIMIC approach to testing the interaction also would be possible, where positive affect at Time 1 and Time 2 are regressed on binary marital status variable. Given the redundancy with the other strategies, I will not discuss the results of this analysis.

The MIMIC model test of the interaction hypothesis with multiple indicators is possible with any of the several modeling strategies and factor identification approaches, but I only report the contrast coding model (Figure 3.4c) with effects coding identification as an illustration (syntax for examples of the other MIMIC models and identification approaches are included on the website for the book). The intercept and the contrast factors were regressed on the marital status variable, and all loadings and intercepts for same-item pairs were set equal over time. This model had acceptable fit, $\chi^2(50) = 124.733$, CFI = .962, SRMR = .045. The test of the significance of the path coefficient of the contrast factor regressed on the marital status variable is the test of the mixed factorial interaction (Marital Status × Time). The unstandardized path estimate was –.096, $z = -1.729$, $p = .084$, suggesting a marginally significant tendency for there to be a greater decrease in the positive affect variable in the married group. The standardized coefficient was –.211, suggesting that this was a small to moderate effect. Although the factor means were computed differently from the analysis of the composite measure (i.e., an unequal weighting of the indicators), the square of the Wald ratio $(-1.729)^2 = 2.989$ was quite similar to the *F*-value or chi-square difference values obtained with the other methods of testing for the interaction of the composite variable.

Extensions

The models discussed in this chapter represent fairly simple hypothesis tests of repeated measures but provide an important stepping stone for more complex analyses. It is easy to imagine extending these analyses to additional time points or additional groups, or both, but it is also possible to incorporate these comparisons into more elaborate models. Latent growth curve models, latent difference score models, and time series models are all closely related extensions of the ANOVA model specifications presented here. The mixed models can be extended as well. Additional groups can be incorporated in several ways, through multiple group analysis, MIMIC style models with multiple dummy variables as predictors, or even through latent class analysis. Higher-order factorial designs are also possible using a MIMIC model with multiple categorical predictors and product variables or by combining the MIMIC and multigroup methods.

Estimates of mean differences over time do not have to be included as dependent variables in a structural equation model. There is no reason why contrast factors, for instance, could not serve as predictors of other variables or even as mediators. And, though I have discussed binary and ordinal variables, some SEM software programs provide estimation for counts or zero-inflated variables, and any of the models discussed can be extended to include these variable types. With extensions of the contrast coding model, it is also possible to modify these models to be equivalent to random effects repeated measures ANOVA for the analysis of small groups, such as family units (e.g., Newsom, 2002), or randomized-block designs. Covariates have not been mentioned but are simple to add in a number of ways, including the use of time-varying covariates, analysis of covariance, and experimental variables.

Comments

Let us return to the question posed at the beginning of the chapter. Why estimate repeated measures ANOVA with SEM? One important aim of this chapter has been didactic, to

explore some of the fundamental principles of analysis of change with structural equation modeling by illustrating their connection to ANOVA. With this, the groundwork has been set to develop a deeper understanding of more complex analyses, such as growth curve models and latent difference models. We have seen that the basic ANOVA model specifications also encompass conventional tests for binary and ordinal variables and can be expanded to include latent variables. Repeated measures ANOVA is often described as an analysis of mean change, but, with the material in this chapter, we have already begun to see how individual differences in change, topics discussed in greater depth in the following chapters, are connected.

In order to maximize clarity, there are a number of issues that arise in practice that have not been discussed, including violations of the multivariate normality assumption and missing data. Nonnormal continuous (dependent) variables can be used with appropriate estimation (e.g., robust adjustments or bootstrapping) but complicate likelihood ratio tests (see Chapter 1 for greater detail). These are inconveniences, though they are addressable. In addition, I have simplified examples in this chapter and others by omitting discussion of missing data, because consideration of missing data for each of the model variants and circumstances would make the material considerably more challenging. Missing data are rather conveniently handled in nearly all structural modeling packages through full information maximum likelihood methods under appropriate conditions (e.g. Arbuckle, 1996; Enders, 2013), and this translates into an important advantage for the convenience of the SEM approach over traditional ANOVA as well as several interesting possibilities for extensions to less time-structured designs. Another simplification has been the omission of correlated measurement residuals with multiple indicator models, but it is a topic I will address at greater length in subsequent chapters. The estimation of correlated residuals is common in longitudinal models and is often recommended to account for measurement artifacts. Their inclusion in the model may impact inter-factor correlations, which, in turn, may have some impact on the precision of the repeated measures tests.

Recommended Readings

A general understanding of how analysis of variance is related to structural equation models can be gleaned from discussions of between group comparisons. Graham (2008) shows a variety of statistical tests as special cases of structural equation models. A more detailed discussion of between-subjects ANOVA models using SEM is given by Thompson and Green (2006; Green & Thompson, 2012). See Hancock (2003) and Liu and colleagues (Liu, Rovine, & Molenaar, 2012) for more about the benefits of SEM compared with ANOVA. Rovine and Molenaar (2003) were the first to discuss repeated measures ANOVA models using SEM (see also Rovine & Liu, 2012). Voelkle (2007) also provides a general discussion of repeated-measures ANOVA models with SEM. For related articles on fixed and random effects ANOVA, see Newsom (2002) and Raudenbush (1993).

Notes

1 Although most SEM software programs allow the user to estimate structural paths among measured variables and between observed and latent variables, the original LISREL program syntax required that at least one latent variable be assigned to each observed variable in a structural model. Use of path model with measured variables, as in the model depicted in Figure 3.1a, requires setting the latent variable, η_1, equal to the observed variable, y_{2-1}, by specifying that a single loading, λ_{11}, be equal to 1, and that the corresponding measurement intercept, ν_1, and measurement residual, θ_{11}, both be equal to 0. Any of the models discussed in this chapter can be specified in this manner and will yield equivalent results.

2 Most SEM software programs and most authors compare the Wald ratio of parameter estimate divided by its standard error to the normal curve (z-distribution), assuming a large sample size. For small N, less than approximately 120 cases, it may be wise to consider a t-distribution as a more conservative test that takes into account the greater variability of the sampling distribution.
3 This connection made between the two models was not the first one. Meredith and Tisak (1990) noted the connection between repeated measures ANOVA and latent growth curves in one of their early papers on the topic.
4 See Raudenbush (1993) for a discussion of the general equivalence of multilevel models and random effects ANOVA.
5 Loglinear models do not directly test marginal homogeneity, but a test can be computed from the difference between the symmetry and quasi-symmetry models (Agresti, 2013). An alternative test for repeated ordinal measures is the Friedman test. The Friedman test is often thought of as a test of related medians, but it can also be conceived as an ordinal test and is equivalent to McNemar's for two related binary measures or Cochran's Q for several related binary measures. Although the Friedman test is of the same hypothesis as the loglinear marginal homogeneity test, the tests are not exactly equivalent.
6 Rovine and Molenaar (2003) show how the repeated measures ANOVA sum of squares and F-value can be obtained from a structural model that analyzes pre-computed difference scores. They also show how this method can be used for trend analysis and factorial models. Because their approach requires an input of difference scores as well as a constant vector of ones that make it less convenient for users, I do not discuss their method further.
7 The usual formula for Wilks's lambda is simplified here to show the computation when there are no group comparisons. The number of groups must be taken into account where multigroup comparisons are involved (see Tatsuoka, 1988, p. 91, for this generalization).
8 Effect codes could be used in this model if desired (e.g., –1, +1, 0 and –1, 0, +1), which would provide an equivalent omnibus test of equality of the three means. With this coding scheme, the mean for the intercept factor would equal the grand mean and mean of the contrast factors would represent the difference of their respective means from the grand mean. Note that also under this coding scheme, the loadings for the second contrast factor are the same as in the linear trend test illustrated below.
9 The estimation of the covariance between the factors is important, because, without this parameter in the model (i.e., assuming the covariance, ψ_{12}, is equal to 0), the mean of the second factor is simply an estimate of the factor mean at Time 2.
10 Remember that the effects of equality constraints or freely estimated parameters on the factor mean estimate is a moot point if referent intercept and referent loading identification are used, because the factor mean takes on the value of the observed mean for the indicator.
11 For simplicity, I have assumed that there are no covariances among measurement residuals of repeated indicators over time. In practice, covariances among measurement residuals for pairs of repeated indicators are likely to be estimated when there are sufficient degrees of freedom. These "correlated errors" will make the interpretation of the factor difference even more complex, because of the impact that the covariances have on the value of the covariance between the factors and on the loadings.
12 A similar process could also be used for the model using the single occasion identification approach, although it is also simple to alter the model specifications to allow both means to be estimated as described above.
13 As other research has shown, and as appears to be the case in this data set, it is important to distinguish between widowed and other non-married individuals in emotional health outcomes. Analysis of widowed as a separate group of adults in this data set does show differences in the changes in positive affect over time.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.
Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, *62*, 488–499.

Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods*, *2*, 208–218.

Arbuckle, J. L. (1996) Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques*, (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Enders, C. K. (2013). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller, *Structural equation modeling: A Second course* (2nd ed., pp. 493–519). Charlotte, NC: Information Age Publishing.

Fava, J. L., & Velicer, W. F. (1992). An empirical comparison of factor, image, component, and scale scores. *Multivariate Behavioral Research*, *27*, 301–322.

Graham, J. M. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics*, *33*, 485–506.

Green, S. B., & Thompson, M. S. (2012). A flexible structural equation modeling approach for analyzing means. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 393–416). New York: Guilford.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*, 373–388.

Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods*, *2*, 293–305.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*: *A Multidisciplinary Journal*, *13*, 59–72.

Liu, S., Rovine, M. J., & Molenaar, P. (2012). Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychological Methods*, *17*, 15–30.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122.

Mulaik, S. A. (2009) *Linear causal modeling with structural equations*. Boca Raton, FL: CRC Press.

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf.

Neter, J., Kutner, M. H., Nachtsheim, & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago, IL: Irwin.

Newsom, J. T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling*, *9*, 431–447.

Newsom, J. T. (2012). Basic longitudinal analysis approaches for continuous and categorical variables. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 143–179). New York: Routledge.

Newsom, J. T., Prigerson, H. G., Schulz, R., & Reynolds, C. F. (2003). Investigating moderator hypotheses in aging research: Statistical, methodological, and conceptual difficulties with comparing separate regressions. *International Journal of Aging and Human Development*, *57*, 119–150.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 459–496). New York: Marcel Dekker.

Rovine, M. J., & Liu, S. (2012). Structural equation modeling approaches to longitudinal data. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 243–270). New York: Routledge.

Rovine, M. J., & Molenaar, P. C. M. (2003). Estimating analysis of variance models as structural equation models. In B. Pugesek, A. Tomer, & A. von Eye (Eds.), *Structural equation modeling: Applications in ecological and evolutionary biology research* (pp. 235–280). New York: Cambridge.

Sankaran, M. (1963). Approximations to the non-central chi-square distribution *Biometrika*, *50*, 199–204.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. New York: Routledge.

Stoolmiller, M. (1994). Antisocial behavior, delinquent peer association, and unsupervised wandering for boys: Growth and change from childhood to early adolescence. *Multivariate Behavioral Research*, *29*, 263–288.

Tatsuoka, M. M. (1988). *Multivariate analysis*. New York: Macmillan.

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119–169). Greenwich, CT: Information Age.

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*, 4–70.

Voelkle, M. C. (2007). Latent growth curve modeling as an integrative approach to the analysis of change. *Psychology Science*, 49, 375.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 209–231). New York: Guilford.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: measuring the same construct across time. *Child Development Perspectives, 4*, 10–18.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Yuan, K. -H., & Bentler, P. M. (2006). Asymptotic robustness of standard errors in multilevel structural equation models. *Journal of Multivariate Analysis*, *97*, 1121–1141.

Keywords
repeated measures ANOVA, MANOVA, proportions, matched pairs, marginal homogeneity, latent difference score

4 Fundamental Concepts of Stability and Change

This chapter reviews essential concepts of stability and change, with an ultimate objective of promoting a deeper understanding of longitudinal models designed to investigate causes of change. In contrast to the previous chapter, which was more concerned with analyses that investigate whether change occurs, this chapter is concerned more with explaining which individuals change and identifying the potential causes of change. There is greater emphasis in this chapter and the next on the conditional or statistical control approach to change, which is better suited for investigating questions about causal precedence, than on a mean or individual differences approach to change discussed in some later chapters(e.g., Chapters 3, 7, and 9.

To truly grasp what it really means to predict change, we will need to reconsider some of the elementary concepts of stability and change. It seems like a simple matter to define stability as the absence of change and change as the absence of stability, but there are some core disagreements about what either of these terms really means. We begin by considering stability and change in continuous, binary, and ordinal variables at some length and then consider approaches to predicting change.

Stability and Change

General Concepts

The concept of stability can be an elusive one. One view of stability concerns the constancy of the absolute level of a measure. Using this sense of the term, we might say that the "income level of the sample was stable over time." Such a statement implies that dollar amount was the same value (or approximately the same) over time. Although applied to the group average here, we can apply a similar definition to an individual (Hertzog & Nesselroade, 1987). If the employee's income is stable over time, her income has the same absolute value each time it is assessed. Stability, however, can also imply that there is a strong correspondence between scores assessed at one point in time and scores assessed at another point in time (Kagan, 1980). We might call this *relative level stability*, because correlations quantify consistency in rank ordering more than consistency in the absolute level. Using this definition, we could say that income is highly stable over time if there is a high correlation between two scores or that an individual's income at one point in time is a good predictor of his or her income at another point in time. Average or individual income does not have to stay at the same absolute value over time for the scores to be considered stable under this definition.

Now consider change as the opposite of stability applying these two definitions. In the absolute level stability sense, change refers to an increase in the average or individual score over time. Income was not the same value; it increased or decreased. In the relative

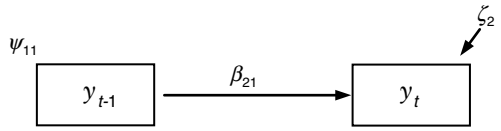

Figure 4.1 Autoregression with Observed Variables.

stability sense, change refers to a low correlation of scores over time. Income was not highly correlated. Absolute level stability and change are assessed by comparing means or difference scores ("change scores" or "gain scores"). As shown in Chapter 3, the average difference score is equal to the difference between the means, so the paired *t*-test or repeated measures ANOVA evaluate both quantities simultaneously. Less obviously, latent growth curve and latent difference score models also investigate absolute stability and change. Relative stability and change are assessed with correlation or regression, and analysis of covariance (ANCOVA) and cross-lagged panel analysis are based on this notion of stability and change. In the end, it will not always be so simple to classify structural models entirely into one of these two camps, and many models can incorporate some aspects of both definitions of change. A case in point is time series analysis, which often employs some elements of difference scores and autoregression.

The two conceptualizations of stability and change have been debated for decades (see Campbell & Kenny, 1999; MacKinnon, 2008; Finkel, 1995 for reviews), and it is not only difficult for researchers to decide which definition they should choose but it is also not always obvious which definition underlies a particular analysis. Rather than necessarily advocate for one definition over the other, I view my goal as distinguishing some of the key implications of the analyses associated with the two views of stability and change. Because many issues associated with absolute level stability and change have already been discussed, there will necessarily be more focus on issues associated with relative stability and change in this chapter.

Observed Continuous Variables

A predictive equation that models relative stability and change is the *autoregression*. Assume a simple structural model with a measurement, y_t, taken at a later time point, predicted by a measurement of the same variable taken at an earlier time point, y_{t-1} (Figure 4.1). Although I have used symbols y_t and y_{t-1} to refer to observed variables here, equivalent models, which will be addressed in due course, can be constructed using latent variables with a single indicator if the loadings are set to 1 and the measurement residual is set to 0. The observed variable may be either a single measured variable, such as an attitude question, or composite scores, such as an average or sum of several attitude questions.

$$y_t = \alpha_t + y_{t-1}\beta_{t,t-1} + \zeta_t \tag{4.1}$$

In this model, the regression path, $\beta_{t,t-1}$, sometimes referred to as the *autoregressive effect, stability path*, or *inertia* is the unstandardized regression coefficient that represents the increment in the value of the measure at time *t* for each unit increment in the measure at time *t*–1. The *i* subscript for an individual case is omitted from the variables and the error term for simplicity. The mean structure may or may not be of interest in the particular SEM application. If, for convenience, we assume the intercept is 0 (model has no mean

structure), then stating Equation (4.1) in terms of the disturbance (or residual), we have the equation

$$\zeta_t = y_t - y_{t-1}\beta_{t,t-1} \tag{4.2}$$

The equation defines the disturbance as the remainder of the score y_t after removing the portion of y_{t-1} that is associated with it. Viewed in a slightly different way, the disturbance is the difference between y_t and y_{t-1}, where y_{t-1} is weighted by the autoregression coefficient. This form of the equation shows that the disturbance will be equal to the difference score only if $\beta_{t,t-1}$ is equal to 1.

Assuming a positive association between the repeated measurements, which is likely to occur in practice in most instances, an individual's disturbance score and an individual's difference score both will be positive when the raw score increases over time.[1] Both will be negative if the raw score decreases over time. The magnitude of change will not be the same for the two computations, however. The magnitude of change of the disturbance score depends on the absolute level difference as well as the magnitude of the association (i.e., autoregression coefficient), whereas the magnitude of change of the difference score is just the amount of discrepancy between the raw values. The magnitude of change in the difference score will tend to be larger than the magnitude of the disturbance score in practice, because the regression coefficient will tend to be less than 1.

According to the relative level definition of stability, perfect stability occurs when the autocorrelation is equal to 1. This implies the same relative position of the scores at each time point rather than the same values as implied by difference scores. To see how the autocorrelation implies the same relative position, consider that the correlation coefficient can be stated in terms of the difference of standardized scores, z_t and z_{t-1} (e.g., Falk & Well, 1997).[2] Assuming standardized scores computed using the sample formula for standard deviation, the correlation between any two scores approaches 1 minus half of the average squared difference for large N.

$$r_{t,t-1} = 1 - \left(\frac{\sum (z_t - z_{t-1})^2}{2(N-1)} \right)$$

If there is no difference in the standardized scores, the second term is equal to 0 and the correlation becomes 1.[3] Because standardized scores give information about the relative position of the raw values, we see that the autocorrelation tends toward 1.0 to the extent that the relative position of scores remains the same (Spearman, 1904).

Without covariates, the standardized autoregression coefficient is equal to the autocorrelation coefficient. The formula for the standardized autoregression coefficient, computed in the usual way, suggests that if the standard deviations for y_{t-1} and y_t are equal, the unstandardized autoregression coefficient will be equal to the standardized autoregression coefficient.

$$\beta^*_{t,t-1} = \beta_{t,t-1} \left(\frac{sd_{y_t}}{sd_{y_{t-t}}} \right) = r_{t,t-1}$$

The standard deviations may not be exactly equal in most circumstances, but, with repeated measurements of the same variable, they will frequently be comparable. Under the condition of equal variances, the disturbance from the autoregression and the raw difference score will be equal only when the autocorrelation is perfect, a condition that

will occur infrequently in practice.[4] Rarely then will the two methods of quantifying stability yield identical results.

To summarize, we might say that the absolute level definition of stability implied by difference scores is stricter in the sense that perfect stability occurs only if all differences are exactly zero. The correlation must be perfect and the raw values must be identical for differences scores to indicate no change. The definition is less strict if we consider stability to refer to the average difference score being zero, because individual scores could vary. The relative level definition of stability implied by the autoregression, on the other hand, is less strict than either form of absolute level stability, because a measure may remain perfectly stable even if raw scores change to the extent that the relative position does not change.

Regression Toward the Mean. Scores tend to vary stochastically over time, with extreme scores tending toward the average and average scores tending toward the two extremes. This pattern suggests a negative correlation between the initial score, $t - 1$, and the difference score, t. High initial scores tend to decrease and low initial scores tend to increase. This *regression toward the mean* occurs whenever the autoregression coefficient is less than perfect or whenever the correlation is less than perfect, under the simplifying condition that means and standard deviations can be assumed to be equivalent (Campbell & Kenny, 1999).[5] Although part of this less-than-perfect relationship is due to measurement error in many instances, the less-than-perfect relationship can be due to many factors. The degree of regression toward the mean is inversely related to the strength of the relationship between the two repeated measures. Because the disturbance in the autoregression model is a function of the strength of relationship, the autoregression definition of change takes into account regression toward the mean. The difference score metric of change includes any deviation in the value over time, including any regression toward the mean. This does not suggest there is a "flaw" in the difference score per se, only that the autoregression "adjusts" for the correlation over time and the difference score does not.

Intercepts. For convenience, Equation (4.2) assumed that there was no mean structure and thus no intercept in the autoregression, so let us consider the case where the intercept is estimated. When the intercept is included in the model, an increase or decrease in the mean is taken into account in the value of the disturbance.

$$\zeta_t = y_t - \alpha_t - y_{t-1}\beta_{t,t-1} \tag{4.3}$$

Inclusion of the intercept ensures that the average disturbance is zero. Without its inclusion, the average disturbance is equal to what the intercept value would be in the model.

Restating the regression model from Equation (4.3) and inserting the formula for the intercept, $\alpha_t = \overline{y}_t - \beta_{t,t-1}\overline{y}_{t-1}$, implies a model with deviation scores.

$$\begin{aligned}\zeta_t &= y_t - \alpha_t - \beta_{t,t-1}y_{t-1} \\ &= y_t - \left(\overline{y}_t - \beta_{t,t-1}\overline{y}_{t-1}\right) - \beta_{t,t-1}y_{t-1} \\ &= \left(y_t - \overline{y}_t\right) - \beta_{t,t-1}\left(y_{t-1} - \overline{y}_{t-1}\right)\end{aligned}$$

If the regression coefficient is equal to 1, the model gives the difference between deviation scores. This suggests another way of contrasting the difference score from the disturbance value, because it becomes clear that disturbance involves the relative position of each score from the mean. Adding or subtracting a constant value from all y_{t-1} or all y_t scores will leave the disturbances unaffected, because the mean has been removed. In contrast, adding or subtracting a constant value from y_{t-1} or y_t, will affect the difference

scores. Therefore, altering the absolute level of the observed values may have nothing to do with the degree of stability in the autoregression model but everything to do with the degree of stability in the difference score model.

Example 4.1: Autoregression with Continuous Observed Variables

To illustrate a simple autoregression model with an observed variable at two time points, a composite index was computed by averaging three questions about the frequency of receiving unwanted advice in the social exchanges data set. The composite index was used for comparison with analysis examples later in the chapter. Syntax and data sets used in the examples are available at the website for the book.

A simple model following Figure 4.1 is just identified with no model fit information. The unstandardized autoregressive coefficient was significant, .371, $p < .001$, and indicated a little more than a third of a point increase in unwanted advice ratings at Time 2 (i.e., six months later) for every unit increase in unwanted advice ratings at Time 1. The mean at Time 1 was 1.979, and the mean at Time 2 was 1.983, with an average difference of –.004. The standardized autoregression coefficient was slightly higher ($\beta^* = .377$) than the unstandardized coefficient, reflecting a slight difference in the standard deviations of the measure at the two time points ($sd_{y_{t-1}} = .809$ vs. $sd_{y_t} = .822$). The square of the standardized coefficient, which is equal to the square of the correlation coefficient in this case, indicates that approximately 14% of the variance in unwanted advice at Time 2 was accounted for by unwanted advice at Time 1. These results provide one example where the average difference score is very near zero, suggesting very high degree of stability in the average level, yet the autoregression coefficient was far from perfect, suggesting a much more modest degree of stability.

Observed Binary and Ordinal Variables

Many of the concepts related to autoregression that have been discussed thus far apply similarly to binary or ordinal variables. In this chapter, I assume the reader is familiar with SEM estimation of binary and ordinal variables, so I only briefly review them. More detail on these estimators can be found in Chapter 1 and other sources (e.g., Finney & DiStefano, 2013; Muthén, 1993). Subtraction of a binary variable, coded 0 and 1, from another binary variable leads to only three possible values, –1, 0, and +1. Although –1 indicates a decline, 0 suggests a constant value, and +1 suggests an increase, researchers may be interested in distinguishing two types of cases that remain the same (e.g., "yes" and "yes" vs. "no" and "no"), both with a difference score equal to 0.

The generalized linear model formulation (McCullagh & Nelder, 1989) is convenient for conceptualizing the simple autoregression equation for binary or ordinal variables. In this formulation, the model is stated in terms of a linear equation with the dependent variable y^*, which represents an unobserved continuous propensity that underlies the binary or ordinal observed variable. An unobserved value on y^* exceeding some threshold results in an observed value on y of 1 instead of 0 (refer to Figure 1.3).

$$y_t^* = \alpha_t + \beta_{t,t-1} y_{t-1} \tag{4.4}$$

The slope from the model gives the change in the continuous y_t^* for each unit change y_{t-1}.

The continuous y^* variable must be linked to the observed y variable using a nonlinear link function. The link function depends on the estimation method used, where two common methods give logistic or probit coefficients. For ML, parameters are usually estimated as logistic coefficients that involve a natural log function link, and, for WLSMV

with delta parameterization, parameters are probit coefficients that involve a normal link function. For a simple autoregression model, the link between the linear equation and observed scores can be expressed as a transformation of the model parameters to the predicted probability that y_t is equal to 1, given the value of y_{t-1}. The transformation of logistic parameters uses the standard logistic cumulative distribution function (cdf), and the transformation of probit parameters uses the standard normal cdf, referred to below as Φ for either type.

$$p \mid y_{t-1} = \Phi\left(\alpha_t + \beta_{t,t-1} y_{t-1}\right), \tag{4.5}$$

The left-hand side of the equation is the conditional probability, $p|y_{t-1}$, that the observed variable y_t is equal to 1 given the predictor y_{t-1}. If the propensity on the unobserved y_t^*, the observed value on y_t will be equal to 1.

Binary Variables. Let us consider the logistic parameters from the autoregression model first. Although the right-hand side of Equation (4.5) still follows the linear model shown in Equation (4.1), Menard (2010) points out that the unstandardized autoregression coefficient in the logistic model implies a different scaling for y_{t-1} and y_t. The predicted value of y_t is a logarithmically transformed value (or logit transformation), $\ln\left[p/(1-p)\right]$, and the predictor variable in the equation, y_{t-1}, is not. The autoregressive effect therefore represents a nonlinear relationship, because the predicted value will increment nonlinearly with each unit increment in the predictor. This suggests a more complicated comparison between the difference score model and the unstandardized logistic estimate of the autoregressive effect.

For the binary case, the exponentiation of the intercept, $p \mid y_{t-1} = 1/\left(1+e^{\alpha_t}\right)$, estimates the proportion of cases $y_t = 1$ when $y_{t-1} = 0$ (e.g., proportion responding "yes" at Time 2 if they responded "no" at Time 1). The interpretation of the intercept α_t is the same as with continuous variables, because it represents the expected value of y_t when y_{t-1} equals 0. The exponentiation of the regression coefficient, $e^{\beta_{t,t-1}}$, gives the odds ratio, which can also be defined in terms of the observed proportions.

$$e^{\beta_{t,t-1}} = \frac{P\left(y_{t-1}=0 \mid y_t=0\right)/P\left(y_{t-1}=0 \mid y_t=1\right)}{P\left(y_{t-1}=1 \mid y_t=0\right)/P\left(y_{t-1}=1 \mid y_t=1\right)} = \frac{p_{00}/p_{01}}{p_{10}/p_{11}} = \frac{p_{00}p_{11}}{p_{10}p_{01}} \tag{4.6}$$

The odds ratio is the increased odds that y_t is equal to 1 if y_{t-1} is equal to 1 instead of 0. For a repeated survey, it is the odds of replying "yes" in the second administration compared with responding "no" given that the response was "yes" in the first administration compared with "no."

As with continuous measures, the disturbance (residual or error) is one approach to conceptualizing change, because it represents the remaining unaccounted-for variance once the association with prior measurement is taken into account. Using the generalized linear model framework, the disturbance can be interpreted as a remainder from y_t^* after accounting for the value of y_{t-1}, $\zeta_t^* = y_t^* - y_{t-1}$. This is a simplification though, because y_t^* represents an exponential link function in the logistic case. The dependent variable for the logistic model, y_t, has only two possible values, so the disturbance can take only two possible values for any given value of y_{t-1}. Across the range of values of y_{t-1}, the residuals may take on any number of possible values between –1 and +1, with their absolute values representing a conditional probability that y_t is equal to 1. The disturbance is typically assumed to have a logistic distribution with mean equal to 0 and variance equal to $\pi^2/3$. If the logistic distribution is assumed to be standardized with variance equal to 1, it has

a close resemblance to the standard normal distribution. The resemblance suggests that the logistic estimates will often closely resemble the probit estimates (Long, 1997). In fact, based on the relationship of the variances of the two underlying distributions, logistic regression coefficient will be approximately $\pi / \sqrt{3}$ larger than the probit regression coefficient.

Considering the standardized coefficient is one method of simplifying the interpretation of the autoregression model for non-continuous variables. With both measurements standardized, the standardized autoregressive coefficient, $\beta^*_{t,t-1}$, would represent a standard deviation change in y^*_t, for a unit standard deviation increase in y_{t-1}. The standardized coefficient can also be expressed in terms of cell proportions of the 2×2 table of repeated measures.

$$\beta^*_{t,t-1} = r_{t-1,t} = \frac{p_{00}p_{11} - p_{01}p_{10}}{\sqrt{p_{0.}p_{1.}p_{.0}p_{.1}}} \tag{4.7}$$

In this equation, p is used for observed probabilities with subscripts 00, 01, 10, and 11 representing cell proportions for repeated observations from the first and second measurement, respectively. Subscripts 0. and 1. represent marginal proportions for the first assessment, whereas .0 and.1 represent marginal proportions for the second assessment.[6] Note that Equation (4.7) is the same as the formula for the phi coefficient for the correlation between two binary variables. As with the equivalent Pearson correlation formula and the standardized regression coefficient, the numerator is the covariance of y_t and y_{t-1} and the denominator is equal to the product of their standard deviations. The numerator can also be related to the logistic regression coefficient from Equation (4.6), as it is equal to the log-odds, $\ln\left(p_{00}p_{11} / p_{01}p_{10}\right) = e^{p_{00}p_{11} - p_{01}p_{10}}$. Contrast this equation with the difference score model, which implies an unconditional difference between proportions, subtracting one row from one column (either $p_{0.} - p_{.0}$ or $p_{1.} - p_{.1}$). The logistic autoregressive model, on the other hand, can be said to be "conditional" in the sense that the (exponential) difference in proportions represents the difference in proportions for y_t depending on the value of y_{t-1}.[7]

Standardization methods vary for logistic regression models, so estimates may differ across software programs. The standardized estimate also can be interpreted as the standard deviation change in y^*_t associated with each standard deviation change in y_{t-1}. Although the standardized coefficient for logistic models will more closely approximate the value from Equation (4.7), the standardization method can vary by author and software program (Menard, 2010, pp. 90–91, provides a review). The fully standardized coefficient follows the same formula for continuous regression models, $\beta^*_{t,t-1} = \beta_{t,t-1}\left(sd_{y_{t-1}} / sd_{y^*_t}\right)$. The standard deviation of the predictor variable, y_{t-1}, is computed as usual. The standard deviation of y^* is slightly more complicated, $sd_{y^*_t} = \sqrt{\left(\beta^2_{t,t-1}\right)\mathrm{Var}\left(y_{t-1}\right) + \left(\pi^2 / 3\right)}$. The last term, $\pi^2 / 3$, is approximately 3.29, a value conventionally taken as the variance of the logistic distribution. The standardized coefficient is calculated in the same way for the probit estimate, except that the standard deviation for y^* is computed slightly differently, $sd_{y^*_t} = \sqrt{\left(\beta^2_{t,t-1}\right)\mathrm{Var}\left(y_{t-1}\right) + 1}$.

Another standardized metric that has been used to describe stability (Plewis, 1985) should also be mentioned. Although *Yule's Q*, which Yule named after Belgian statistician Quetelet (Yule, 1900), is bounded by –1 and +1, it is not identical to the correlation coefficient. Instead, it represents the proportion of cases that remain equal over time, with 0 representing the absence of stability and an absolute value equal to 1 representing perfect stability.[8] Yule's Q can be stated in terms of the odds ratio or the observed proportions.

$$\begin{aligned}\text{Yule's } Q &= \left(\frac{OR-1}{OR+1}\right) \\ &= \left(\frac{e^{\beta_{t,t-1}}-1}{e^{\beta_{t,t-1}}+1}\right) \\ &= \frac{p_{11}p_{00}-p_{10}p_{01}}{p_{11}p_{00}+p_{10}p_{01}}\end{aligned} \tag{4.8}$$

Ordinal Variables. The autoregression with ordinal variables quantifies a one-unit increment in y_t for each one-unit increment in y_{t-1}, taking into account the linearizing transformation. The appropriate link function connects the predicted ordinal values for the observed y_t to the unobserved continuous distribution of y_t^* through $C-1$ thresholds, where C is the number of ordinal categories. The slope can be interpreted as the change in y_t^* for each standard deviation change in y_{t-1}. As with binary variables, the autoregressive model is a conditional model of change because y_t values are a function of or dependent upon y_{t-1} values. The autoregressive or conditional model interpretation of change contrasts with the difference score or unconditional model interpretation of change discussed in Chapter 3 (e.g., Cochran–Mantel–Haenszel [CMH] or loglinear models).

The autoregressive model for ordinal variables can also be stated in terms of Equation (4.5) and the generalized linear model. With logistic estimates, the odds ratio can be computed to describe the odds that y_t will increment by 1 for a one-unit increase in y_{t-1}. Standardized coefficients are defined in the same manner as they are for binary variables, providing one way to conceptualize the magnitude of the stability coefficient. Stability may also be conceptualized in the same manner as Yule's Q. Goodman and Kruskal's (1954) statistic, $\hat{\gamma}$, extends Yule's Q to ordinal variables.

Example 4.2: Autoregression with Binary and Ordinal Observed Variables

An autoregression with binary variables was illustrated using data from the social exchanges data set. A yes/no question about whether the respondent had recently had a major health event was asked at two consecutive interviews six months apart. Table 4.1 gives the observed frequencies and proportions for each cell. Full ML estimation was used to obtain logistic estimates, and the model is just identified. The unstandardized autoregression coefficient was equal to 1.244 and was significant ($p < .001$). The odds ratio was 3.469, equal to $e^{1.244}$ in accordance with Equation (4.6), indicating that the odds of reporting a health event at Time 2 were about 3.5 times greater if the respondent experienced a health event at Time 1 as compared to not experiencing a health event at Time 1. The standardized coefficient was equal to .176, which matches the value obtained for the phi coefficient from a standard statistical software program. Alternatively, the standardized coefficient can be obtained by multiplying the unstandardized estimate by the ratio of the standard deviations of y and y^*. This computation produces the same value,

$$\beta_{t,t-1}^* = \beta_{t,t-1}\left(sd_{y_{t-1}} / sd_{y_t^*}\right) = 1.244(.261/1.843) = .176$$

where

$$sd_{y_t^*} = \sqrt{\left(\beta_{t,t-1}^2\right)\text{Var}\left(y_{t-1}\right)+\left(\pi^2/3\right)} = \sqrt{(1.244)(.068)+3.29} = 1.843$$

The standardized coefficient can also be obtained by applying Equation (4.7) to the proportions from Table 4.1. The square of the correlation is equal to .031 and indicates about

Table 4.1 Repeated Measures Results for Health Event

		Health event t = 2		
		No (0)	Yes (1)	Total
Health event $t = 1$	No (0)	413 ($p_{00} = .720$)	119($p_{01} = .207$)	532 ($p_{0.} = .0.927$)
	Yes (1)	21 ($p_{10} = .037$)	21($p_{11} = .037$)	42 ($p_{1.} = .073$)
	Total	434 ($p_{.0} = .756$)	140($p_{.1} = .244$)	574

3% of the variances in health events at Time 2 is accounted for by health events measured at Time 1. Applying Equation (4.8) to obtain Yule's *Q* gives a value of .553, one estimate of stability that suggests that approximately 55% of cases remain constant over time.

The same autoregressive model of health events was tested using the WLSMV estimator, which is more commonly used for binary and ordinal variables. The unstandardized estimate of the autoregression coefficient from this analysis was .759 and was significant ($p < .001$), representing the expected change in y^* for each increment in health events at Time 1 (i.e., from no event to at least one event). The standardized estimate was .194, which is slightly higher than the standardized estimate obtained with ML estimation.

The Effects of Measurement Error

Before discussing autoregression models with latent variables, let us first consider the effect of measurement error on autoregression coefficients and difference scores.

Autoregression Coefficients. Unstandardized regression coefficients are impacted only by measurement error in the independent variable, and standardized coefficients are impacted by measurement error in the independent or dependent variable. The attenuating effect of measurement error is a consequence of the increase in variance, which can be seen in the classical test theory formula, $\mathrm{Var}(X) = \mathrm{Var}(T) + \mathrm{Var}(e)$. The observed score variance is the sum of the variance of the true score, *T*, and the variance of the measurement error, *e*, and, as a consequence, a larger error variance leads to a larger observed score variance. The impact of measurement error on the unstandardized autoregression coefficient can be seen in the formula for the unstandardized regression coefficient, $\beta_{t,t-1} = \mathrm{Cov}(y_{t-1}, y_t) / \mathrm{Var}(y_{t-1})$, which only takes into account the variance of the independent variable. In fact, we can correct for measurement error by estimating the unstandardized regression coefficient taking into account the reliability of the independent variable, $\beta_{t,t-1}^{\text{corrected}} = \left[\mathrm{Cov}(y_{t-1}, y_t) / \left(sd_{y_{t-1}} \cdot sd_{y_t}\right)\right] / \rho_{t-1}$. The smaller the reliability of the independent variable, ρ_{t-1}, the more attenuated the obtained regression coefficient and the greater the discrepancy between the corrected value and the obtained value.

The correlation coefficient or the standardized regression coefficient is impacted by measurement error in both the independent and dependent variable, because the standard deviations of both variables are taken into account in their computation, $r_{t-1,t} = \beta^*_{t,t-1} = \mathrm{Cov}(y_{t-1}, y_t) / \left(sd_{y_{t-1}} \cdot sd_{y_t}\right)$. Thus, any inflation of the standard deviation in either variable due to measurement error will result in a smaller standardized coefficient, and the magnitude of effect will be assumed to be smaller than the case when no measurement error is present in the observed variables. The attenuation effect can be substantial, even for reliabilities that are considered "acceptable." An estimate of the disattenuated

correlation or simple standardized regression coefficient can illustrate how reliability plays a role. The disattenuation formula, due to Spearman (1904), $\beta^{*\text{corrected}}_{t,t-1} = \beta^{*}_{t,t-1} / \sqrt{\rho_{t-1}\rho_t}$, is equal to the standardized autoregression coefficient, $\beta^{*}_{t,t-1}$ divided by the square root of the product of the reliability of the measures, ρ_{t-1} and ρ_t.

To observe the impact of measurement error on the standardized autoregression coefficient, consider an example in which the obtained standardized regression coefficient is equal to .5 and reliabilities of the two measures are both equal to .7. With these values, the disattenuated estimate of the true relationship would be .714, which is of considerably larger magnitude (i.e., 50% variance accounted for instead of 25% variance accounted for). In many cases, the underestimate may be considerably less, because the reliability of the measures is higher. With a reliability of .9, the underestimate of the variance accounted for is only about 5.9%, although this magnitude of underestimation may be enough to arrive at incorrect conclusions. Although this shortcoming is potentially serious, it can be addressed by the estimation of latent variables.

Models with observed variables can incorporate correction for measurement error by constructing a latent variable with a single indicator at each time point and setting the loading equal to 1 and the error variance equal to $(1-\rho)\text{Var}(y)$. This procedure is not recommended for several reasons (see Deshon, 1998; McDonald, 1996) and is unnecessary in nearly all instances. Typically, if an estimate of reliability for a measure is available, multiple indicators can be used to estimate a latent variable.

Difference Scores. It is worthwhile to give some brief consideration to difference scores. Measurement error has no biasing effect on the average of the difference score, as random error has no biasing effect on the expected value of a measure (see Chapter 1 for more detail). Random measurement error may lead to overestimation or underestimation of individual scores, but, on average, it is expected to be 0. Thus, the difference score will be unaffected by measurement error on average,

$$\begin{aligned} E(y_t - y_{t-1}) &= E\left[(y_t + e_t) - (y_{t-1} + e_{t-1})\right] \\ &= \left[E(T_t) + E(e_t)\right] - \left[E(T_{t-1}) + E(e_{t-1})\right] \\ &= \left[E(T_t) + 0\right] - \left[E(T_{t-1}) + 0\right] \\ &= E(T_t) - E(T_{t-1}) \\ &= E(T_t - T_{t-1}) \end{aligned}$$

where $E(.)$ represents the expected or average value.

Measurement error will have an impact on significance tests and standardized coefficients when difference scores are predicted by another variable, because each takes the variance of the variables into account. The observed score variance will be equal to or greater than the true score variance whenever there is measurement error, $\text{Var}(T_t) \leq \text{Var}(y_t)$, so the variance of the difference score will be affected as a consequence. To see this, consider that the variance of the difference score will be a function of the variances of the two variables and their covariance.

$$\text{Var}(y_t - y_{t-1}) = \text{Var}(y_t) + \text{Var}(y_{t-1}) - 2\,\text{Cov}(y_t, y_{t-1})$$

To the extent that the repeated measurements are uncorrelated, the variance of the difference score will be equal to the sum of the variances of the two measurements. Assuming for a moment that the true score and the error are independent and, for convenience, that

errors at each time point are independent, we see the extent to which the variance of the difference score will be inflated by measurement error.

$$\begin{aligned}\mathrm{Var}\left(y_t - y_{t-1}\right) &= \mathrm{Var}\left[\left(T_t + e_t\right) - \left(T_{t-1} + e_{t-1}\right)\right] \\ &= \mathrm{Var}\left(T_t\right) + \mathrm{Var}\left(e_t\right) + \mathrm{Var}\left(T_{t-1}\right) + \mathrm{Var}\left(e_{t-1}\right) - 2\,\mathrm{Cov}\left(T_t, T_{t-1}\right)\end{aligned} \tag{4.9}$$

Measurement error in either of the two observed variables, then, will contribute to the variance of the difference score.

Although it might be expected that the reliability of the difference score will be lower if there is measurement error for the individual measures, it is perhaps surprising that the reliability of the difference score will tend to be lower than the reliability of the individual measures. The equation below for the reliability of the difference score, which assumes equal variances, is often used to illustrate that the difference score reliability, ρ_d, is lower than the average of the two separate reliabilities, ρ_{t-1} and ρ_t, whenever the autocorrelation, $r_{t,t-1}$, is larger than 0 (e.g., Kaplan & Saccuzzo, 2013).

$$\rho_d = \frac{\left[\left(\rho_{t-1} + \rho_t\right)/2\right] - r_{t-1,t}}{1 - r_{t-1,t}} \tag{4.10}$$

As Equation (4.10) indicates, if the two measurements are uncorrelated, the reliability is simply the average of the two reliabilities. For a positive autocorrelation, the reliability of the difference score decreases with increasing values. If the autocorrelation happens to be negative, which is a rarer case, then the difference score reliability is actually larger than the average of the two separate reliabilities. The reader can verify with a few hypothetical values that the reliability for the difference will be considerably lower than reliability of either measure for many likely values of the autocorrelation.

The importance of the lower reliability has been disputed and debated at length with many authors arguing theoretically or demonstrating empirically that reliability of differences scores may not always be low and that, even if reliability is low, this may not always have an important bearing on statistical power (e.g., Collins, 1996; Kisbu-Sakarya, MacKinnon, & Aiken, 2013; May & Hittner, 2003; Williams & Zimmerman, 1996). The apparent disconnect between the reliability of the difference score and statistical power is because reliability of difference scores is a complex function of several factors, including the autocorrelation, the ratio of the variances of the two separate scores, and the variance of the difference score.

The reliability of the difference score is more complicated than Equation (4.10) suggests, and there are several limiting factors on the unreliability of the difference score. This formula uses observed autocorrelation, $r_{1,t-1}$, and it could be argued that the autocorrelation of true scores makes more sense (Williams & Zimmerman, 1996). Because the autocorrelation of true scores will be larger in the presence of measurement error, the lower reliabilities of the difference score will be offset by replacing the observed score autocorrelation with the true score autocorrelation. It is also the case that as the ratio of the variances of the separate scores, $\mathrm{Var}(y_{t-1})$ and $\mathrm{Var}(y_t)$, departs from a 1:1 ratio, the reliability of the difference score increases. The reliability of the difference score also increases with an increase of the autocorrelation. Although the variance ratio and the autocorrelation may vary from variable to variable and across studies, use of the same measurement over time suggests that the ratio of variances will often be near 1. It is the case too that, for many data sets and many variables, the autocorrelation will be positive and of moderate magnitude. Aside from the effect of measurement error in y_{t-1} and y_t on the difference score, the relation between measurement error of individual scores and the reliability of

the difference score is further complicated if there is dependence of the errors of the individual scores (Williams & Zimmerman, 1996). Accounting for this dependence should increase the reliability of the difference score.

The other important factor in the reliability of the differences score is its variability, which can be shown in the following equation:

$$\begin{aligned}\rho_d &= \frac{\mathrm{Var}(T_t - T_{t-1})}{\mathrm{Var}(T_d) + \mathrm{Var}(e_d)} \\ &= \frac{\mathrm{Var}(T_t - T_{t-1})}{\mathrm{Var}(y_2 - y_1)} \\ &= \frac{\mathrm{Var}(T_t - T_{t-1})}{\mathrm{Var}(T_t) + \mathrm{Var}(e_t) + \mathrm{Var}(T_{t-1}) + \mathrm{Var}(e_{t-1}) - 2\,\mathrm{Cov}(T_t, T_{t-1})}\end{aligned}$$

This formula, which assumes $\mathrm{Var}(e_{t-1})$ and $\mathrm{Var}(e_t)$ are independent, illustrates that even for a fixed value of the true score variance and the error variance of each measure, reliability increases with the variance of the difference of the true scores (i.e., individual differences in absolute change) and approaches zero when there is little variability in difference score (Raykov, 2001). If all individuals in the study change similarly, for instance, the variance of the differences will be low, which will lead to a low reliability estimate for the difference score. The variability of the difference score does not necessarily have anything to do with whether change is accurately assessed, and, may in fact, may be indicative of low error variance of individual scores (Collins, 1996).

Example 4.3: Effects of Measurement Error on Autoregression Coefficients

An autoregression model of the composite measure of unwanted advice was used to illustrate measurement error attenuation. Cronbach's alpha was estimated for the items used in the composite score of Time 1 and Time 2 unwanted advice. Their reliability estimates were $\rho_1 = .82$ and $\rho_2 = .76$, respectively. Using the formulas above for the correction for measurement error, the corrected unstandardized coefficient was

$$\beta_{21}^{\text{corrected}} = \left[\mathrm{Cov}(y_1, y_2) / (sd_1 \cdot sd_2)\right] / \rho_1 = \left[.25 / (.809 \cdot .822)\right] / .820 = .376,$$

about 50% larger than the obtained value. The correction to the standardized coefficient was

$$\beta_{21}^{*\text{corrected}} = \beta_{21}^{*} / \sqrt{\rho_1 \rho_2} = .377 / \sqrt{(.820)(.757)} = .463,$$

about 23% larger than the obtained value. The autocorrelation (.377) and the individual scale reliabilities were used to estimate the reliability of the difference score with Equation (4.10), giving an estimate of .661. As expected, this value is considerably lower than either of the individual scale reliabilities.

Latent Variables with Multiple Indicators

I extend the discussion of autoregression and latent difference scores to latent variables with multiple indicators. The estimation of measurement error is an important advantage of latent variables, but this advantage does not extend equally to the two approaches to the analysis of stability. Because latent variables with binary or ordinal indicators can be

modeled in the same manner as latent variables with continuous indicators, given different estimation methods, I do not include any special sections on binary and ordinal variables below.

Autoregression. Multiple indicators of a latent variable can be used at each time point to estimate an autoregressive model. We know generally that use of latent variables has an advantage over observed variables by estimating measurement error. The autoregressive coefficient for observed scores will tend to be lower than the autoregressive coefficient for true scores. Latent variables can adjust for this attenuation in the autoregressive coefficient.

Multiple indicators account for measurement error by partitioning the variance due to the common factor and remaining variance. The loading represents the prediction of the observed indicator and the measurement residual represents unaccounted-for variances, which may be comprised of measurement error and specific variance. An estimate of reliability can then be derived from a ratio of the variance of the common factor, representing "true" score variance, to the observed score variance (Bollen, 1989). In the simple hypothetical case of a single indicator, the estimate of reliability of the measure is

$$\rho = \frac{\lambda_{jk}^2 \psi_k}{\mathrm{Var}\left(y_j\right)},$$

with ρ as the reliability estimate, λ_{jk}^2 as the unstandardized factor loading, and ψ_k as the variance of the factor. The true score variance is estimated by the numerator and the observed score variance is estimated by the denominator. This ratio mirrors the theoretical formula for measurement reliability. Not all measurement residual variance is measurement error, so this reliability estimate would tend to be an underestimate of reliability to the extent that there is any specific variance. The reader is referred to introductory texts for a more complete review of measurement error and the confirmatory factor analysis (e.g., Bollen, 1989; Kline, 2010; Brown, 2006; Maruyama, 1997). Raykov (1997) discusses estimation of Cronbach's alpha with SEM.

Difference Scores. Latent variable difference scores can be modeled in several ways. Several repeated measures specifications of ANOVA discussed in Chapter 3 accomplish this. Following from the discussion above regarding measurement and difference scores, we should expect no advantage of latent variable estimates in determining the averaged difference. This was demonstrated in Example 3.5. Equation (4.9) suggests that measurement error adds unnecessary variance to the variance of the true score difference. By extension, latent difference scores will have a smaller variance than observed variable difference scores to the extent that measurement residual variances are greater than zero. Hancock (2003) shows that compared with observed scores, between-group differences in latent variables will have smaller variances, smaller standard errors, and larger standardized effect sizes. Similar advantages hold for the repeated measures case.

Correlated Measurement Residuals. Correlated measurement residuals (or "correlated errors") are commonly estimated in longitudinal models to take into account unique variance that is common over time. Measurement residual variance contains variance due to measurement error as well as variance unique to an indicator but not common to the other indicators of the factor. Because measurement error is random, by definition, the measurement error component of the variance of the measurement residual cannot be correlated with other factors. The specific variance component of the measurement residual, on the

other hand, is systematic and may be correlated with other factors or may have autocorrelation over time. Take as an example a question about general happiness as an indicator of marital quality. If the respondent is unhappy about factors other than the marital relationship, he or she may have the same unhappy feelings at subsequent measurement time points. Under these conditions, an autocorrelation exists that is due to factors other than the true score of the intended construct, and, if it is not taken into account in the model, the estimate of stability will tend to be overestimated. This is an important advantage that longitudinal models with multiple indicators have over longitudinal path models with observed variables. Even when a path model with only observed variables can be identified with additional waves, the soundness of estimates of correlated measurement residuals for single indicators is questionable (Wiley & Wiley, 1974). Although inclusion of correlated measurement residuals with multiple indicator models could be decided on the basis of significance tests, they are generally included in longitudinal models a priori (Wheaton, Muthén, Alwin, & Summers, 1977). Except for the cost of degrees of freedom, there is typically no harm in including these estimates.[9]

The inclusion of covariances or correlations among measurement residuals for a multiple indicator longitudinal model is illustrated in Figure 4.2. For example, θ_{15} and θ_{26}, are examples of covariances between measurement residuals for indicators that are repeated over time. In matrix form, autocovariances among measurement residuals are represented in the off-diagonal elements of the theta matrix. The effect of such correlated residual variance can be illustrated with a simple example using path analysis decomposition. The covariance between two repeated observed measurements is equal to the sum of the compound paths between the two observed variables. Take indicators y_2 and y_6 from Figure 4.2. Their covariance can be decomposed into the compound path comprised of the product of the two loadings and the covariance between factors, $\lambda_{21}\psi_{12}\lambda_{62}$, and the covariance of the measurement residuals, $\mathrm{Cov}(\varepsilon_2,\varepsilon_6)=\theta_{26}$.

$$\mathrm{Cov}\left(y_2, y_6\right) = \lambda_{21}\psi_{12}\lambda_{62} + \mathrm{Cov}\left(\varepsilon_2, \varepsilon_6\right)$$

The two terms represent the two pathways that can be traced between y_2 and y_6 – the compound pathway through the loadings and the factor covariance and the covariance between the measurement residuals. If we rearrange the terms, it is clear that the estimated covariance between the factors depends, in part, on the measurement residual covariance.

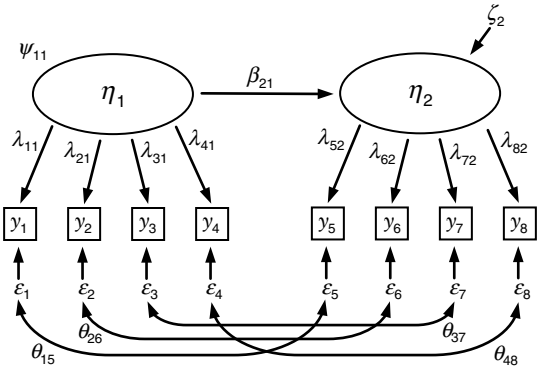

Figure 4.2 Autoregression with Latent Variables.

$$\psi_{12} = \frac{\mathrm{Cov}(y_2, y_6) - \mathrm{Cov}(\varepsilon_2, \varepsilon_6)}{\lambda_{21}\lambda_{62}}$$

If there is no measurement residual covariance, so that $\mathrm{Cov}(\varepsilon_2, \varepsilon_6) = 0$, then the factor covariance will be larger than if there is a measurement residual covariance, $\mathrm{Cov}(\varepsilon_2, \varepsilon_6) > 0$. Assuming the measurement residual covariance is equal to 0 when in fact it is not would overestimate the estimate of the factor covariance and, similarly, the autoregressive effect. This suggests that it is necessary to include measurement residual covariances among repeated indicators in longitudinal models in order to obtain the most accurate estimate of autoregression effects.

In the larger picture, the autocorrelation of specific variances may be caused by or confounded with other factors. In the hypothetical marital quality example, the specific variance for the happiness indicator could be confounded if caused by life stress or depression. Such factors can be explicitly modeled if measured in the data set, but, even if they are not, the proportion of shared variance in the factor covariance can be removed by estimating measurement residual covariances in the model. The inclusion of correlated measurement residuals in the model, therefore, serves to account for some of the third variables which may bias the autoregression estimate.

Specific Factors Model. An alternative approach to including correlations among measurement residuals is to model specific factors using the repeated measurements of each observed variable as indicators of a latent variable that represents the unique variance common to each indicator (Raffalovich & Bohrnstedt, 1987; see Finkel, 1995, p. 69). These specific factor latent variables represent stable variance over time due to unobserved causal factors that contribute the unique, systematic portion of the variance of the observed indicator. A simple example of the model approach is illustrated in Figure 4.3, which specifies a latent variable, η_3 through η_6, for each repeatedly measured indicator. The specific factors are uncorrelated with each other and with the conceptual factors, η_1 and η_2, modeled in the autoregression.

With two waves, the specific factor model is likely to encounter empirical underidentification problems without some additional constraints, such as equivalent loadings. Additional waves add to the likelihood that the model can be identified with fewer

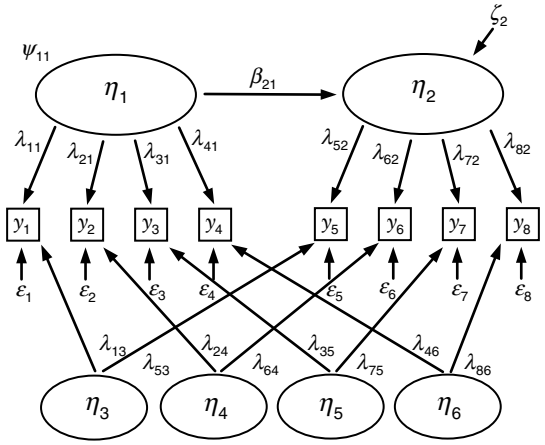

Figure 4.3 Specific Factors Model.

constraints. The specific factors model is an equivalent model to the correlated measurement residuals model and will have the same effect on the autoregressive estimates. The specific factors model provides some additional information compared with the correlated measurement residual model, because the proportion of variance due to each specific factor can be computed from the standardized estimates. This feature may be useful for describing and understanding the sources of longitudinal stability of the measure. The incorporation of specific factors longitudinal models is discussed further and illustrated in relation to the concept of method factors used with trait-state-error models in Chapter 6.

Example 4.4: Autoregression with Latent Variables and Correlated Measurement Residuals

An autoregressive model of unwanted advice with three indicators (as illustrated in Figure 4.2) was first tested without estimating correlated measurement residuals (i.e., assumed to be equal to 0). The model chi-square was significant, but alternative fit indices suggested a good fit, $\chi^2(10) = 34.778$, $p < .001$, CFI = .978, SRMR = .033. The unstandardized autoregression coefficient was .444, $p < .001$, and the standardized coefficient was .472. Note that these estimates are considerably higher than the model tested in Example 4.1 using an average of the individual indicators. The model was retested allowing for correlated measurement errors and also fit the data well, with a nonsignificant chi-square $\chi^2(7) = 12.108$, $p = .097$, CFI = .996, SRMR = .025. The unstandardized autoregression coefficient from this model was .428, $p < .001$, and the standardized autoregression coefficient was .453. Both of these values are lower than the model without correlated measurement residuals, suggesting that their omission led to an inflated estimate of the autoregressive effect. The inflation occurs because some of the stability in unwanted advice is due to stable specific variance in the individual indicators. The difference was not dramatic in this case, as the size of the residual correlations was not large, –.032, .034, and .093. It is easy to imagine that, when measurement residuals are more highly correlated, the bias in the stability estimate will be even larger if they are not estimated in the model.

Comments

Inclusion of multiple indicators and correlated measurement residuals has two valuable effects. Estimation of latent variables with multiple indicators helps mitigate the conservative bias of the autoregression effect that would occur if composite measures or other observed variables measured with error were the basis of the analysis instead. Estimation of correlated errors will reduce an otherwise positive bias of the autoregression effect that may occur if there is autocorrelation of specific variance that is not accounted for in the model. Although these two biases are countervailing, the degree of bias from either is never known without explicitly modeling them. In my experience, the autocorrelation of specific factors tends to have a fairly modest bearing on autocorrelation estimates, because measurement residual correlations are often small. The effect of measurement error, however, can be more substantial and analysis of latent variables can show fairly dramatically different results than when each variable is included as a composite scale score. Difference scores represent an alternative way to conceptualize stability and change, but because measurement error does not bias expected values, difference score averages are unaffected by measurement error. Estimation of latent difference scores, however, has advantages because the variance of the difference score can be reduced, with some attendant statistical benefits.

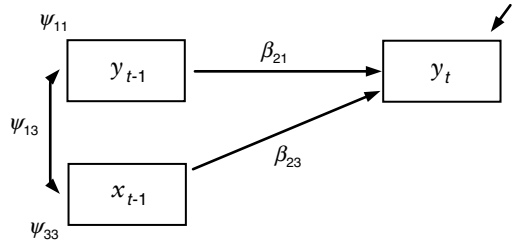

Figure 4.4 Cross-Lagged Effect with Observed Variables.

Predicting Change

Cross-Lagged Effects

Cross-lagged models are intended to examine causal precedence to determine whether one variable causes another. In the absence of experimental data, this will always be an imperfect endeavor. The basic *lagged regression* model estimates the effect of a hypothesized cause measured at an earlier time point on a hypothesized effect measured at a later time point, while controlling for the initial values of the effect variable. The model is sometimes referred to as the "conditional change model" or the "static score model." For example, we can imagine using perceived stress as a predictor of marital quality one month later, controlling for marital quality at baseline. Such a model is illustrated in Figure 4.4 and can be represented by the following equation, where y_t is a repeated measurement of y_{t-1}, and x is a predictor variable:

$$y_t = \beta_{21} y_{t-1} + \beta_{23} x_{t-1} + \zeta_t \tag{4.11}$$

I omit the intercept value (no mean structure) and assume observed variables temporarily. β_{21} is the autoregressive coefficient, and β_{23} is the *cross-lagged effect*.

There are two rationales for this type of model. First, the inclusion of outcome measured at the first time point, y_{t-1}, attempts to control for pre-existing differences in the outcome. In this sense, the model addresses the possible explanation for a relationship between the predictor at baseline, x_{t-1}, and the outcome at the later time point simply because there was already an existing relationship between the two variables at baseline (represented in the figure by the covariance ψ_{13}). Perceived stress may predict marital quality one month later because there was a pre-existing connection between the two variables. That pre-existing connection might be because something about the marital relationship was affecting perceptions of stress. To the extent that this concurrent association is removed, we control for the pre-existing relationship due to the opposite causal direction or due to other external factors.

The second rationale for including the dependent variable measured at the earlier time point is that the stable aspects of the dependent variable, represented by β_{21}, are removed, so that the cross-lagged effect, β_{23}, represents the effect of the predictor x_{t-1} on changes in the dependent variable y_t. After controlling for the earlier measurement of marital quality, we see the effect of perceived stress on changes in marital quality. It is worth noting that for a model that includes a cross-lagged path, the autoregression coefficient becomes a partial regression coefficient, representing stability in the dependent variable once the effects of the predictor have been removed from the association. In other words, the autoregression represents the stability of marital quality over time, removing any portion

of the association that may have been due to levels of perceived stress. Certainly, the definition of stability and change in this context should be taken to be specific to how they are defined under the autoregression model.

Analysis of Covariance. If the predictor x_{t-1} is a binary variable, the model shown in Figure 4.4 represents a special case – the ANCOVA model, in which two groups are compared on a dependent variable while controlling for initial differences on the dependent variable (Aiken, Stein, & Bentler, 1994). In the ANCOVA model, the cross-lagged path is the group-difference of the dependent variable at Time 2 controlling for initial differences between the two groups at Time 1. A common example would be a comparison of the difference between intervention and control groups controlling for pre-existing differences on the outcome variable. The unstandardized coefficient for the cross-lagged path represents the mean differences on the Time 2 measurement of the dependent variable, y_t, and the intercept, α_t, represents the mean of y_t when the grouping variable, x, is equal to 0 (e.g., the mean of the control group). Standard software procedures for ANCOVA, however, center the values of the covariate by subtracting its sample mean, and this produces the "adjusted mean" estimate. In the structural modeling analogue, adjusted means can be estimated either by centering y_{t-1} or by setting its mean, α_{t-1}, equal to 0 in the model. Setting α_{t-1} equal to 0 would have a detrimental effect on fit but would produce an appropriate estimate of the adjusted mean. With rescaling of y_{t-1}, the intercept α_t will be equal to the adjusted mean for the group coded 0 on the predictor, x_{t-1}. The interpretation of the intercept can be altered by centering x_{t-1}, producing an intercept that represents the value of y_t when y_{t-1} and x_{t-1} equal their means (i.e., the grand mean of y_t).

The ANCOVA model can be extended to G groups by including a set of $G-1$ dummy variables for a multiple category independent variable. The ANCOVA model is a conditional model of change, because initial values of the dependent variable are taken into account, and contrasts with the mixed factorial repeated measures ANOVA discussed in Chapter 3, which is an unconditional model of change. As illustrated in that chapter, the regression coefficient for a binary variable predicting a difference score represents the estimate of the interaction effect.

Example 4.5: Cross-Lagged Effects and ANCOVA

A model with a cross-lagged effect as illustrated in Figure 4.4 was tested using a composite of the positive affect measure from the social exchanges data set. The composite was calculated as an equally weighted average of the five items from the scale. The model investigates whether negative social exchanges in the form of unwanted advice are predictive of positive affect six months later, controlling for initial levels of positive affect. The model is just identified and equivalent to a multiple regression model with two predictors, with positive affect at Time 2 regressed on positive affect at Time 1 and the unwanted advice composite measure. Positive affect at Time 1 significantly predicted positive affect at Time 2, $\beta = .490$, $\beta^* = .465$, $p < .001$, but unwanted advice at Time 1 did not predict later positive affect, $\beta = -.029$, $\beta^* = -.036$, ns. The nonsignificant effect is consistent with prior research that negative social exchanges do not have a longitudinal association with positive affect (Newsom, Nishishiba, Morgan, & Rook, 2003).[10]

A different example from the same data set was used to illustrate an ANCOVA. Specification of the ANCOVA model followed those shown in Figure 4.4 but with a binary predictor. The model examined differences between those who did and did not report a major health event at Time 1 on the composite measure of unwanted advice at Time 2, controlling for any initial differences in unwanted advice at Time 1. To demonstrate

equivalence with traditional ANCOVA, the mean structure was estimated and unwanted advice at Time 1 (the covariate) was centered by subtracting its mean. Results indicated a significant positive cross-lagged effect, $\beta = .262$, $\beta^* = .084$, $p < .05$, indicating that those reporting a health event at Time 1 were more likely to receive unwanted advice six months later after equating the two groups on their initial levels of unwanted advice. The results match those obtained from an ANCOVA procedure using a standard software program, where the significance tests in the two programs were equal, $t^2 = (2.166)^2 = F = 4.67$. The intercept value from the structural model, 1.963, was equal to the adjusted mean estimate for the 0 group (i.e., those who reported no health event at Time 1) requested from the ANCOVA output. Moreover, the unstandardized coefficient for the cross-lagged effect (.262) was equal to the difference between the two adjusted means (2.225 – 1.963) from the ANCOVA.

Comments

This kind of application of ANCOVA is frequently used for analysis of non-experimental intervention studies. Researchers should keep in mind that changes in an outcome will be harder to predict when the autoregression effect is large. A highly stable variable will require a powerful cause to produce changes in it. The covariance between the presumed cause and effect measured at the initial time point takes into account any causal effect that is occurring simultaneously between them, regardless of the direction, as well as any third variables that may be responsible for their relationship. This does not necessarily indicate that all variables confounded with the lagged effects of the predictor will be controlled for, however. Relatedly, the cross-lagged effect does not necessarily take into account any simultaneous effects of the predictor that occur at the later time point. For example, although perceived stress one month earlier may not predict later marital quality, it may still be that perceived stress one month after baseline predicts marital quality one month after the baseline. Finally, the model in Equation (4.11) does not provide any information about the possibility that there may also be a cross-lagged effect over the given period operating in the opposite direction.

Predicting Difference Scores

Mixed Factorial Anova. To contrast the cross-lagged model with a model that predicts difference scores, examine Figure 4.5 for a moment. A variable, x_1, such as perceived stress measured at an initial time point, predicts the difference score for another variable, $y_{2-1} = y_2 - y_1$, such as marital quality. This model is the same as a simple regression, $y_{2-1} = \beta_{21} x + \zeta_2$, and is therefore sometimes referred to as the unconditional model of change. The dependent variable at time t is not "conditioned on" or does not control for the dependent variable measured at time $t - 1$. When the predictor variable is binary, the unconditional model is equivalent to the mixed factorial ANOVA discussed in Chapter 3.

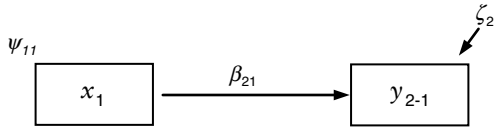

Figure 4.5 Difference Score Prediction or "Unconditional" Model of Change.

The analysis then contrasts with the ANCOVA equivalence of the cross-lagged model just discussed.

In the mixed ANOVA model implied by the difference score simple regression, the test of β_{21} is a test of the interaction of the between-subjects factor, represented by the binary variable, and the repeated measures factor, represented by the difference score. With some algebra, the unconditional (difference score) model can be restated in terms of Equation (4.11) (Finkel, 1995; Plewis, 1985).

$$y_{2-1} = \beta_{23}x + (\beta_{21} - 1)y_1 + \zeta_2 \tag{4.12}$$

The conditional model of Equation (4.11) also can be restated in terms of the unconditional model (following Equation [4.2]).

$$(y_2 - \beta_{21}y_1) = \beta_{23}x + \zeta_2 \tag{4.13}$$

The left-hand side of the equation is equal to y_{2-1} if β_{21} is equal to 1. A comparison of Equation (4.12) and Equation (4.13) highlights the contrast between the conditional and the unconditional models of change – they are equivalent only if the autoregression coefficient β_{21} is equal to 1.

Lord's Paradox. The fact that the conditional and the unconditional models of change differ from one another suggests that the results from the two analysis approaches will usually produce estimates of the effect of the predictor on change in the dependent variable that are not equivalent. Although the results of the two approaches may lead to the same conclusion in some instances, they may lead to different conclusions in other instances. The discrepancy in the results of the two approaches predicting change is referred to as *Lord's paradox* (Lord, 1967), named after the psychologist Frederic Lord who noted the possibility that the two methods of predicting change sometimes lead to different conclusions. The closer the autoregression effect is to 1, the more likely the results from the two approaches will lead to congruent conclusions.

Lord's paradox can be stated in terms of regression toward the mean (Campbell & Kenny, 1999), where extreme scores move toward the mean over time and scores near the mean move toward the extremes over time. Whenever the autoregressive effect is less than 1, a predictor or covariate associated with baseline values of the dependent variable can be mistaken as a cause of the change over time if regression toward the mean is not taken into account. Figure 4.6 illustrates this point with a simple case in which the lines represent changes in an observed score, y, measured at two time points. Dotted and solid lines represent two groups measured by a binary predictor variable, x. An association between x and change scores could occur because of regression toward the mean if there are differences between the two groups at baseline (i.e., values of x are associated with values of y at baseline). Because high initial scores are decreasing over time and low initial scores are increasing whenever repeated measures of y are not perfectly correlated over time, an association of the difference scores with values of x could occur for reasons that are meaningful or not.

The classic example of regression toward the mean as a methodological artifact occurs when individuals initially scoring low on a test are compared to individuals initially scoring high on a test in a non-experimental intervention study. Scores in the intervention group appear to increase relative to the comparison group because of regression toward the mean. The autoregression approach takes into account the association of x and y at baseline through statistical control. Although the association between x and difference scores on y supplies reasonable estimates of which cases change over time in what direction, the changes in y may be due to initial differences in y that are associated with x.

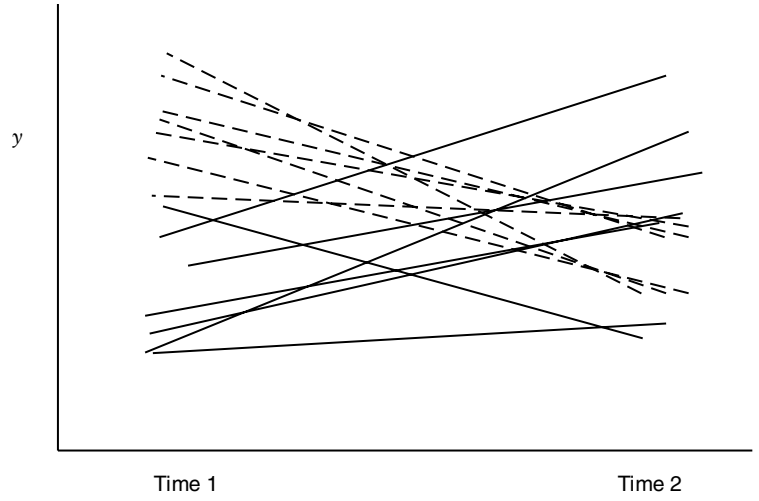

Figure 4.6 Regression Toward the Mean for Two Chosen Extreme Groups.

Example 4.6: Comparing Cross-Lagged and Difference Score Models

To compare to the cross-lagged model in Example 4.5, a second model was tested that estimated the effect of unwanted advice at Time 1 on the positive affect difference score (Time 2 positive affect minus Time 1 positive affect). The average difference score for positive affect was –.009, suggesting little average change in this variable over time. Though there was little average change, it still might be worthwhile to account for individual differences in changes scores with a predictor. The effect of unwanted advice on the difference score was not significant, $\beta = .041$, $\beta^* = .050$, ns, however.

Although both the lagged regression model from Example 4.5 and the difference score model found no significant effect of unwanted advice, it is worth noting that the coefficients were in opposite directions. The cross-lagged model suggested that higher levels of unwanted advice were associated with declines in positive affect, whereas the difference score model suggested that higher levels of unwanted advice were associated with gains in positive affect. These results seem to be at odds with one another and serve as a case of Lord's paradox (albeit a weak one). Higher levels of unwanted advice were associated with lower levels of positive affect at Time 1. The cross-lagged regression controls for initial differences in positive affect when examining the effect of unwanted advice on later positive affect. The positive coefficient for unwanted advice in the difference score model is likely to be due to different patterns of change for those with different initial levels of positive affect. If those with high initial levels of positive affect tended to decrease in affect over time (and those with low initial levels of positive affect tend to increase over time), there would be a negative correlation between initial positive affect and change in positive affect (i.e., consistent with regression toward the mean and illustrated by Figure 4.6). The negative correlation between initial levels of unwanted advice and initial levels of positive affect would then imply a positive association between initial unwanted advice and positive affect difference scores. This example provides an illustration of how the lagged regression and the difference score approaches test different hypotheses.

Comments

The distinctions between the difference score and the lagged regression models serve only to clarify that the two approaches do not have the same definition of change or stability, and I leave it to others to argue that one approach is preferable to the other (e.g. Allison, 1990; Cronbach & Furby, 1970; Dwyer, 1983; Edwards, 1995; Rogosa, 1988; Tisak & Smith, 1994). There are several points that should be made, however, in order to provide the reader with a full understanding of the trade-offs. One point is that it never makes sense to include the prior measurement of the dependent variable as a predictor when the difference score is used as the outcome. Equation (4.12) shows that this would be an analysis that is duplicative of the conditional model. In a related vein, this equation also illustrates the regression to the mean phenomenon, because there will tend to be a negative effect of prior measurement of the dependent variable whenever the autoregressive estimate is positive and less than 1. Regression toward the mean is a potential artifact for the difference score analysis if the goal is to discover the causal direction of the relationship.

I am asked all too often if the difference score approach might not be better, I suspect, because the lagged regression does not yield the desired results. The experience is not uncommon, particularly when the variable is highly stable (i.e., highly correlated) over time. But this would be the wrong reason for wanting to use the difference score approach. The right reason for predicting difference scores is the goal of discovering more about which cases change over time.

The two approaches ask different questions and define change and stability differently. The conditional (lagged regression model) analysis asks the question whether the predictor is associated with the dependent variable at a later time point assuming that cases are equated initially. Adjusting or "equating" initial values of the dependent variable according to values of the predictor (i.e., statistically removing the effects of the predictor) is potentially fallible, because there may be an under-adjustment, or, in some cases, an over-adjustment, to the baseline values of the dependent variable (Campbell & Kenny, 1999). One of the reasons for the under-adjustment is the presence of measurement error that attenuates the correlation between the predictor and baseline values of the dependent variable, a problem that is addressable when latent variables with multiple indicators are modeled.

Latent Variables with Multiple Indicators

Cross-lagged regression models with observed variables may be generalized to structural models of latent variables with multiple indicators. Use of latent variables with multiple indicators can take measurement error and specific variance into account in the estimate of autoregressive and cross-lagged effects.

Unidirectional Models. Unidirectional models, or models that only examine the effect of x on y and not the effect of y on x, may be used when the researcher has an a priori hypothesis about the direction of causality or if data on the predictor is not available at both time points. There are many published examples in which longitudinal models are tested without explicit interest in the hypotheses about the opposite causal direction. Such unidirectional models can be tested with regression analysis, but the use of multiple indicators will provide more accurate estimates of autoregressive and cross-lagged effects. I will discuss bidirectional longitudinal models in the next chapter, Chapter 5, devoted entirely to cross-lagged panel models.

A unidirectional model with latent variables is illustrated in Figure 4.7. Estimation of covariances among measurement residuals for repeated indictors over time will improve the autoregressive path estimate and, indirectly, the cross-lagged path estimate

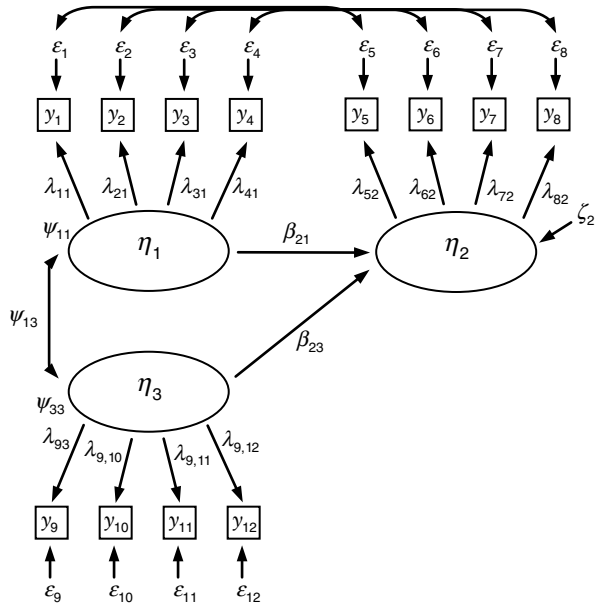

Figure 4.7 Cross-Lagged Effect with Latent Variables.

by removing stable specific variance from the autoregression between the latent variables. For the sake of contrast, consider a variation of this model in which the predictor variable, x_{t-1}, does not have multiple indicators (see Figure 4.4). To the extent that the predictor has measurement error, the cross-lagged path, β^*_{23}, would be underestimated and, because the correlation between the two exogenous variables, ψ^*_{13}, would be underestimated, the autoregression estimate, β^*_{21}, would tend to be overestimated as a consequence. Just from this relatively simple model, it becomes clear that measurement error has a complex effect on the model estimates and potential conclusions when there are multiple predictors in the model and partial regression estimates are involved. For larger models with additional exogenous variable, each additional observed variable included as a predictor, of course, adds further complexity, and inaccuracy, to the estimates of the autoregression and the cross-lagged effects.

It should be noted that the fit of the model in Figure 4.4 or its latent variable counterpart in Figure 4.7 has nothing to do with the correctness of the causal model, because the structural portion of the model is just identified, or, in other words, saturated with all possible relations among the variables estimated. Any lack of fit of the latent variable model with cross-lagged effect, therefore, is due to lack of fit of the measurement portion of the model. The sources of poor fit may be that the factor structure is incorrect, measurement invariance does not hold, or additional correlated measurement residuals are needed (e.g., between two indicators of the same factor).

Measurement Invariance. Good practice is to investigate all aspects of measurement invariance prior to testing other substantive hypotheses. This issue is discussed at greater length in Chapter 2. Imposition of longitudinal equality constraints is common for any models with multiple indicators. Loadings, measurement residual variances, and measurement intercepts may be constrained over time, depending on what makes sense empirically and theoretically. Most authors recommend, at a minimum, testing for equality of loadings

for repeated indicators over time (e.g., $\lambda_{21} = \lambda_{52}$) and then constraining these parameters to be equal over time if they are found to be equivalent (i.e., weak invariance). Because the focus of lagged-regression models is typically on prediction rather than means, the model does not need to estimate a mean structure, and therefore testing of equality constraints of measurement intercepts for repeated indicators would not be required. Notice that it would not make sense to set variances for exogenous factor variances equal to the disturbance variance for the model in Figure 4.7, for example, because the variance for the exogenous factor, $\psi_{11} = \text{Var}(\eta_1)$, is unconditional and the variance of the disturbances, $\psi_{22} = \text{Var}(\zeta_2)$, is conditional representing unaccounted-for variance. Similarly, if factor means and intercepts for the structural model are estimated, longitudinal constraints would not make sense. If means are estimated in this model, α_1 is an estimate of the unconditional mean of the exogenous factor, η_1, whereas α_2 is an estimate of intercept for the structural regression (or conditional mean of η_2).

Identification Choice. The factor identification specification does not affect the estimate of the unstandardized or standardized autoregression coefficient when longitudinal equality constraints are used on the factor loadings. Although the factor variance is dependent on how the factor is identified, a change in the identification approach will proportionately change the variance as long as the same approach is used at each time point and equality constraints on the loadings are used. Thus, the autoregression estimate is unaltered by the change in variance, because the variances of the two variables are changed proportionately.

The unstandardized cross-lagged path coefficient is sensitive to changes in the factor variance of the dependent variable, however. That is, for the model in Figure 4.7, the value of β_{23} will be altered if the method for identifying the factor variances of η_1 and η_2 is altered. Choosing a different referent indicator to identify the factors (e.g., setting $\lambda_{(2)}$ equal to 1 instead of $\lambda_{(1)}$) will produce a different estimate of the cross-lagged path coefficient, because the exogenous variable η_1 and the covariance of the exogenous variables η_1 and η_3 are a function of the variance of the observed variance referent indicator. The results under different identification strategies are statistically equivalent, however, because the significance tests and standardized estimates are unaffected to the arbitrary scaling of the factor variance.[11] Standardized coefficients and standard errors involve the standard deviations of both x and y, so these quantities are unaffected by scaling changes that affect the variances.[12]

Example 4.7: Unidirectional Cross-Lagged Effect with Multiple Indicators

The effect of unwanted advice on positive affect was examined in a model following Figure 4.7. Latent variables for positive affect at Time 1 and Time 2 with five indicators, longitudinal equality constraints on loadings, and correlated measurement residuals were specified. A latent variable for unwanted advice with three indicators was specified at Time 1. Effects coding identification (Little, Slegers, & Card, 2006) was used for each of the latent variables for the initial model. Although the chi-square was significant, $\chi^2(61) = 154.489$, $p < .001$, the model appeared to have a reasonably good fit to the data based on alternative fit indices, CFI = .964, SRMR = .044. The autoregressive effect was significant, $\beta = .600$, $\beta^* = .565$, $p < .001$, suggesting positive affect was fairly stable over the six-month period. The cross-lagged coefficient was not significant, however, $\beta = -.015$, $\beta^* = -.018$, ns. Although the conclusions from these results are consistent with those drawn from Example 4.5, the results differ for two reasons. Most important, the present model takes into account measurement error as well as correlated measurement residuals due to correlated specific variance. The most apparent difference in this particular example is in

the estimates of the autoregressive effects, with $\beta^* = .465$ for the observed variables model and $\beta^* = .565$ in the latent variable model. Much less important, because the effects coding identification approach was used for the latent variables, the variance of each of the variables was a weighted function of the loadings.

A second latent variable model was tested using a referent loading identification of the factors. The first indicator of each factor was chosen for the referent (arbitrarily in this case), and longitudinal equality constraints were imposed on the remaining loadings for the positive affect factor. Correlated measurement residuals were included as before. The fit of the model and degrees of freedom are identical to the effects coding identification model as expected. The unstandardized coefficient for the autoregressive effect was also identical to the effects coding model, $\beta = .600$, because the rescaling of the variance of the positive affect factor was proportionate in the two factors when equality constraints on loadings were imposed. The unstandardized coefficient for the cross-lagged effect differed slightly from the effects coding model –.013, ns, as a result of different variance scaling of the positive affect variable. The standardized effects and p-values were identical for all coefficients across the two identification approaches.

The ANCOVA Model

The analysis of covariance model can be estimated with latent variables to examine the effect of a binary predictor a dependent latent variable estimated at Time 2 while controlling for the dependent latent variable estimated at Time 1. The structure of the model is the same as that depicted in Figure 4.7, except that η_3 is replaced by a binary predictor x. Conceptually, the unstandardized cross-lagged path coefficient represents group differences in the adjusted means of η_2 controlling for η_1. The intercept, α_2, represents the adjusted mean for the group coded 0 on the binary predictor.

There are several differences from the ANCOVA with observed variables that should be kept in mind. The factor variances, the exogenous covariances, and the unstandardized cross-lagged coefficient are a function of the identification method. Thus, the interpretation of the intercept as an adjusted mean and the unstandardized cross-lagged coefficient as the difference in adjusted means must take into account the scaling of the factor variance under whichever identification approach is used. For the referent identification approach, the adjusted mean values are based on the observed mean of the indicator variable used as the referent, but the "adjustment" is also a function of the disattenuated cross-autoregression and lagged effects that occur with latent variables. The adjusted mean differs from the adjusted mean estimate in standard ANCOVA procedures unless the indicators for the covariate (η_1) have been centered prior to analysis or the factor mean has been set to equal 0.

The inclusion of longitudinal equality constraints on loadings or measurement residuals also modifies the interpretation of the means and variances in comparison to model using only observed indicators. When measurement equality constraints are imposed, the results represent an ANCOVA when the measurement properties are assumed to be unchanging. The latent variable ANCOVA may differ from an ANCOVA on observed variables, depending on the identification approach to the factor variance and factor mean.

An identification or model specification that takes into account all of the observed indicator means and variances may be the most reasonable. A simplistic method for accomplishing this is to set all measurement intercepts equal to 0 and set all loadings for a factor to be equal to 1. Such constraints will define the factor variance and mean in terms of an equally weighted composite. These specifications will detract from model fit, however,

because of the implicit assumption that all indicators for a factor have equal loadings and measurement intercepts. A more optimal strategy that bases the factor mean and variances on information from all of the indicators is the effects coding identification approach, which creates factor means and variances that are a weighted function of all the indicators. Because the effects coding identification approach does not make assumptions about the relative values of the loadings or measurement intercepts, it does not detract from the fit of the model (i.e., the fit is equivalent to the referent identification or factor variance identification approaches).

Naturally, the results obtained with effects coding identification will not equal a traditional ANCOVA conducted on a simple composite of the observed indicator variables, because the effects coding uses an unequal weighting of indicators and because measurement error is taken into account. Nor will the results obtained from an ANCOVA model using effects coding identification equal results obtained from an ANCOVA model using referent identification or the single-occasion identification approaches, because the factor means and variances are defined differently. In practice, however, results may not always differ appreciably for different identification approaches as long as the observed indicators are measured on a similar scale and have similar means, and loadings on the factor are all high. In any event, careful thought should be given to the identification approach in order to arrive at an optimally desired interpretation of the results in terms of group mean differences and adjusted means.

Example 4.8: ANCOVA with Multiple Indicators

A conceptual replication of the ANCOVA model tested in Example 4.5 was tested with unwanted advice as a latent variable. This model, which follows the general structure of the model shown in Figure 4.7, used effects coding identification, equal factor loadings over time, and autocorrelated measurement residuals for the unwanted advice factor. The cross-lagged effect of health events at Time 1 investigated the hypothesis that those with and without a health event six months earlier would differ on unwanted advice at Time 2, after controlling for any pre-existing differences in unwanted advice at Time 1. To improve the interpretation of the intercept, the indicators for unwanted advice factor were centered prior to the analysis. Centering the indicators sets the factor mean at Time 1, so that the intercept, α_2, can be interpreted as the adjusted mean difference between the two health event groups on the unwanted advice factor.

The model had a significant chi-square, $\chi^2(14) = 72.233$, $p < .001$, but acceptable relative fit indices, CFI = .950, SRMR = .063. The autoregressive coefficient was significant, $\beta = .427$, $\beta^* = .447$, $p < .001$. The cross-lagged path $\beta = .293$, $\beta^* = .109$, $p < .05$, indicated a significantly higher level of unwanted advice at Time 2 for those experiencing a health event at Time 1, after controlling for any group differences in unwanted advice in the two groups at Time 1. The intercept was equal to 1.970, which can be interpreted as the adjusted factor mean of unwanted advice at Time 2 for the group with no health events at Time 1.The unstandardized cross-lagged path coefficient represents a .293 difference between the groups on the adjusted factor mean at Time 2.

Comments

I began this chapter with discussion contrasting the absolute level and relative level stability and change definitions to provide a foundation for interpretation of models used to predict change. The absolute level definition underlies difference score models, whereas the relative level definition underlies cross-lagged models. Either view of stability or change is

legitimate; but they do not always lead to the same conclusions, because they differ in how they define stability and change.

With difference scores, the implicit meaning of "stability" is that the measure maintains exactly equal values over time. The difference score is simple to understand, and is, therefore, often used for the purpose of describing average change or the variability in change. Prediction of differences scores provide information about whether variation in the scores over time is related to the predictor, answering questions such as "whose score is most likely to increase or decrease over time?" In contrast to the lagged regression approach, prediction of difference scores is not intended to address whether the predictor variable is an antecedent or consequence of the dependent variable. The use of the difference score in predictive analyses implies a near perfect or perfect correlation between the two measurements in this regard. In most instances, then, the difference score will imply a closer correspondence between the repeated measurements than is likely to be the case empirically and does not take into account regression toward the mean.

The cross-lagged regression approach, on the other hand, is rarely used to quantify change for descriptive purposes and is typically used for investigating questions about causal precedence by taking into account any pre-existing differences in the dependent variable. The disturbance from the autoregression will be impacted by exact differences, but only to the degree that there is a correlation between η_{t-1} and η_t, and thus takes into account changes over time due to regression toward the mean. Keep in mind that a significant autoregressive effect or cross-lagged effect does not indicate a causal relationship. Any autoregressive effect may be causal or it may be due to other factors. Family income level from one generation may cause family income level in the next generation through inheritance or education. It also may be the case, however, that family income level in the two generations is due to outside economic conditions, such as the employment market. Similarly, a cross-lagged effect of country of origin on family income in the second generation may or may not be a causal one. Both of these examples underscore the importance of including relevant covariates in the model, a message that should not be undermined by my focus in this chapter on simple examples for the purposes of illustration.

An additional caveat for this chapter is that all of the models discussed here assume linear relationships. As such, each coefficient is an estimate of the average or constant effect of the variable. This may not always be the case. If a predictor has different cross-lagged effects for different levels of that variable, a nonlinear relationship is implied. In the case of the autoregression, if some individuals that have a higher score at an earlier time point are more likely to have stable scores over time than individuals with a lower score at the earlier time point, then a nonlinear autoregression effect exists (Dwyer, 1983). As with multiple regression analysis, such relationships can be investigated by using transformations of the predictor variable.

The exploration of the fundamental concepts of change and stability provide a necessary background for more complex models in the chapters to come. It will become clear that many of the principles discussed here are essential, integral parts of latent growth curve models, latent difference score models, latent trait-state-error models, and time series models.

Recommended Readings

There are several sources that provide further background on the concepts and types of longitudinal analyses discussed in this chapter. For the reader interested in review of some basic principles of regression analysis with two waves, there are several good sources (e.g., Newsom, 2012; Taris, 2000). Plewis (1985) has an excellent overview of many

fundamental issues related to analysis of change, including change scores, lagged regression analysis, discussion of continuous and categorical variables, and basic SEM concepts for longitudinal data. MacKinnon (2008, pp. 193–199) provides an excellent brief summary and references to the change-score-versus-lagged-regression debate. Finkel (1995) discusses many fundamental issues covered in this chapter regarding lagged regression ("static-score" models) and difference score analyses as background for more in-depth treatment of full cross-lagged panel models. For a more mathematical discussion of change, see Coleman (1968). Finally, the reader may gain no deeper insight into regression toward the mean and analysis of change over two waves if Campbell and Kenny's (1999) comprehensive treatise is given a careful read.

Notes

1 The autocorrelation may certainly be negative, but this is a considerably less common case. A negative autocorrelation may occur when there is a highly cyclical pattern. A negative autocorrelation will tend to result in individual residuals that are negative when there is a raw score increase and individual residuals that are positive when there is a raw score decrease. In contrast, the difference score would be unaffected by the negative autocorrelation and the direction of the value would always correspond with the direction of the change in raw values. This more exceptional case has received considerable attention from some authors that disagree about its significance for deciding on the reasonableness of the lagged regression approach (e.g., Campbell & Kenny, 1999; Rogosa, 1988).

2 Standardized scores are used here just for the purpose of illustration. Standardization of scores prior to calculating difference scores is generally discouraged in practice, because the meaning of the difference score is altered when the original variance information is lost (Rogosa, Brandt, & Zimowski, 1982). When variables are standardized, the difference score resembles a relative level difference rather than an absolute level difference.

3 The correlation also would be perfect, of course, if half the average squared difference happens to be 0 as well.

4 It is also interesting to consider the opposite circumstance where the correlation is equal to 0. The regression coefficient will be equal to 0, and the residual will simply be equal to the score at the second time point, t.

5 The process of regression to the mean could also be characterized as *entropic decay* (Dwyer, 1983), connoting the opposite of stability or inertia.

6 It is instructive to compare this formula with the formula for the standardized residual from the McNemar test, $r_m = (p_{01} - p_{10}) / \sqrt{p_{01} + p_{10}}$, where $r_m^2 = \chi_m^2$, for contrast with the test of unconditional change for a binary variable discussed in Chapter 3.

7 The term "conditional" in this context is used differently than it is in the Chapter 3 discussion of the conditional logistic model approach to marginal homogeneity. Here, I mean "conditional" in the sense that the value of y_t is dependent upon the value of y_{t-1}. The test of marginal homogeneity is a test of a simple difference, and therefore is not conditional in this sense. The "conditional logistic model" for marginal homogeneity differs from the autoregressive logistic model discussed in the present chapter, because marginal homogeneity model uses only the cases in the discordant cells (e.g., "yes" at Time 1 and "no" at Time 2 and "no" at Time 1 and "yes" at Time 2) in the analysis.

8 If one of the cells has a zero frequency or proportion, Yule's Q may be misleading by suggesting perfect stability or instability.

9 One could test correlated measurement residuals for significance, and set them to be equal to 0 (i.e., omit them) when they are nonsignificant, with the rationale that any estimated correlations greater than 0 due to random chance would reduce precision of the other model estimates. Although this rationale has some merit, there are two precautions. One is that the sample size may not be sufficient to detect true underlying relationships. Another precaution is that these modifications constitute post hoc changes.

10 There does tend to be a consistent finding that negative social exchanges are predictive of negative affect, however, as reported in this study and others.

11 These conditions hold for models with longitudinal equality constraints on loadings and the true state of affairs is that invariance holds. Where factor loadings are not equivalent over time, the researcher will typically seek to modify the measurement model so that consistent measurement properties will be used in the longitudinal model. Imposing equality constraints when loadings are unlikely to be invariant is ill advised, however, because results will be biased by incorrect model constraints. If loadings are exactly equal over time, there would be no effect on path estimates whether equality constraints were imposed or not. If loadings are not exactly equal, however, the path estimates obtained for models with and without loadings constrained would differ.

12 The standardized path coefficients will be unaffected by referent choice, and, when the standard error is based on the standardized estimate, its significance also will be unaffected by referent choice. Standardized coefficients require an alternative standard error computation, because they are scaled by a ratio of the standard deviations, which should be treated as a random variable in the computation of the standard error (Bollen, 1989, p. 125). Many SEM (or regression) software programs do not calculate separate standard errors for the standardized coefficient, leaving the researcher with the inconvenience of having to compute standard errors for each coefficient. Given critiques of the standardized coefficients (e.g., Baguley, 2009; Duncan, 1975; Tukey, 1969), the wisest course is to inspect and report both the unstandardized and standardized effects. In most cases, statistical conclusions are likely to agree. When they disagree, an attempt should be made to understand reasons for the discrepancy (e.g., dramatically different variances) in the context of the data and research problem.

References

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, *62*, 488–499.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary, NC: SAS Institute Inc.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Erlbaum.

Coleman, J. S. (1968). The mathematical study of change. In H. M. Blalock, Jr. & A. B. Blalock (Eds.), *Methodology in social research* (pp. 428–478). New York: McGraw-Hill.

Collins, L. M. (1996). Is reliability obsolete? A commentary on "Are simple gain scores obsolete?". *Applied Psychological Measurement, 20*, 289–292.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" – or should we? *Psychological Bulletin, 74*, 68–80.

Deshon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, *3*, 412–423.

Duncan, O. D. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, *72*, 177–182.

Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences*. Oxford: Oxford University Press.

Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes*, *64*, 307–324.

Falk, R., & Well, A. D. (1997). Many faces of the correlation coefficient. *Journal of Statistics Education, 5*, 1–18.

Finkel, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 439–492). Charlotte, NC: Information Age Publishing.

Goodman, L.A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.

Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods, 2*, 293–305.

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development*, 93–109.

Kagan, J. (1980). Perspectives on continuity. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 26–74). Cambridge, MA: Harvard University Press.

Kaplan, R. M., & Saccuzzo, D. P. (2013). *Psychological testing: Principles, applications, and issues* (8th ed.). Belmont, CA: Wadsworth.

Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2013). A Monte Carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement, 73*, 47–62.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304–305.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.

Maruyama, G. (1997). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

May, K., & Hittner, J. B. (2003). On the relation between power and reliability of difference scores. *Perceptual and Motor Skills, 97*, 905–908.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research*, *31*, 239–270.

Menard, S. (2010). *Logistic regression* (2nd ed.). Thousand Oaks, CA: Sage.

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Newsom, J. T. (2012). Basic longitudinal analysis approaches for continuous and categorical variables. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 143–179). New York: Routledge.

Newsom, J. T., Nishishiba, M., Morgan, D. L., & Rook, K. S. (2003). The relative importance of three domains of positive and negative social exchanges: A longitudinal model with comparable measures. *Psychology and Aging*, *18*, 746–754.

Plewis, I. (1985). *Analysing change: Measurement and exploration using longitudinal data*. Chichester, UK: Wiley.

Raffalovich, L. E., & Bohrnstedt, G. W. (1987). Common, specific, and error variance components of factor models estimation with longitudinal data. *Sociological Methods & Research, 15*, 385–405.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173–184.

Raykov, T. (2001). Studying change in scale reliability for repeated multiple measurements via covariance structure modeling. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 217–229). Lincolnwood, IL: Scientific Software International.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–210). New York: Springer.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*, 726–748.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*, 72–101.

Taris, T. W. (2000). *A primer on longitudinal data analysis*. London: Sage.

Tisak, J., & Smith, C. S. (1994). Defending and extending difference score methods. *Journal of Management, 20*, 675–682.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing reliability and stability in panel models. *Sociological Methodology 1977, 8*, 84–136.

Wiley, J. A., & Wiley, M. G. (1974). A note on correlated errors in repeated measurements. *Sociological Methods & Research*, *3*, 172–188.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*, 59–69.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions, Series A, 194*, 257–319.

Keywords
lagged effects, autoregression, differences scores, gain scores, change, stability

5 Cross-Lagged Panel Models

The primary goal of cross-lagged panel analysis is to investigate the causal direction of the relation between two variables over time. The term *panel study* typically refers passive observational studies with discrete time intervals, usually with no experimental (or even quasi-experimental) design elements. Early work by sociologist Paul Lazarsfeld promoted the investigation of repeated surveys to explore causal hypotheses (Lazarsfeld & Fiske, 1938; Lazarsfeld, 1940), referring to such studies as panel designs. Campbell and Stanley (1963) originally proposed that the analysis of correlations from cross-lagged panel studies in which two variables are measured at two points could be used to investigate the question of whether one variable causally precedes another.[1] Duncan (1969) later discussed simultaneous regression models within a path analysis framework to control for initial values of each dependent variable.

This chapter is an introduction to many of the issues related to conducting cross-lagged panel analysis with SEM, as well as several closely related topics, including simplex models, mediation, and models for continuous time. The cross-lagged panel structural model builds upon the autoregression and unidirectional cross-lagged modeling approach to stability and change discussed in the last chapter. For this reason, there will be less discussion of difference scores, except in a few instances. I will return to difference scores again in the chapter on latent difference score models (Chapter 9). There also is less focus on latent means, although all of the models discussed can add mean structures when they are of interest. Preliminary analyses to investigate longitudinal measurement invariance should be conducted before testing cross-lagged models.

Cross-Lagged Panel Models

Observed Variables

In many cases, both variables are measured on repeated occasions, and it is possible to investigate hypotheses about both causal directions. Models with autoregressive and a single cross-lagged effect can be tested with regression analysis, but, with SEM, both directional hypotheses can be tested in the same model. The intention is generally to investigate which variable causally precedes the other, referred to as *Granger causality* in time series analysis (Granger, 1969). A model for observed variables with two directional effects, the *cross-lagged panel model*, is illustrated in Figure 5.1. Observed variable y_2 is a repeated measurement of observed variable y_1, observed variable y_4 is a repeated measurement of observed variable y_3, β_{21} and β_{43} are autoregressive paths, β_{23} and β_{41} are cross-lagged paths, ψ_{13} is the covariance between the two exogenous variables, and ψ_{24} is the covariance between the endogenous disturbances. Thus, there are two structural equations.

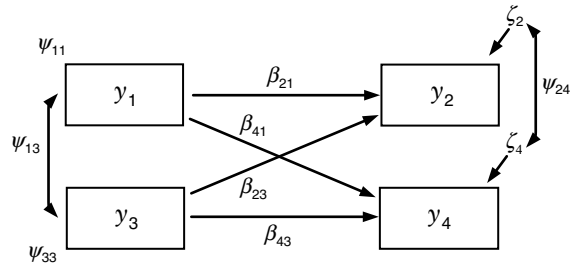

Figure 5.1 Cross-Lagged Panel Model with Two Observed Variables.

$$\begin{aligned} y_2 &= \beta_{21}y_1 + \beta_{23}y_3 + \zeta_2 \\ y_4 &= \beta_{43}y_3 + \beta_{41}y_1 + \zeta_4 \end{aligned} \tag{5.1}$$

These two equations are identical to the equations of two separate regression models.

The model is just identified, with degrees of freedom equal to 0.

$$df = \frac{J(J+1)}{2} - q = \frac{4(5)}{2} - 10 = 0$$

The value J is equal to the number of observed variables and q is equal to the number of free parameters in the model. The ten freely estimated parameters in this model represent six structural parameters ($\beta_{21}, \beta_{23}, \beta_{42}, \beta_{43}, \psi_{13}, \psi_{24}$) and four variable variances.[2] Inclusion of the exogenous covariance in the model, of course, ensures that path coefficients are partial effects controlling for the other variable. Although the covariance between the disturbances does not need to be estimated, it is generally included to account for any remaining synchronous association between the two endogenous variables once the exogenous predictors are taken into account. The exogenous or endogenous covariances may be due to synchronous causal effects at either time point or any extraneous variables responsible for the association between the variables. When this covariance is larger than zero, there will be some model misfit if it is not freely estimated in the model.

The general goal of cross-lagged panel models is to investigate a hypothesis about causal directionality. Does x cause y or does y cause x? For example, does psychological stress lead to marital discord or does marital discord lead to psychological stress. If the cross-lagged effect is significant in one direction but not the other, findings are consistent with the hypothesis that the causal effect works in one direction but not the other. Along the same lines, if neither cross-lagged path is significant, then results are not consistent with causation for either direction, at least given the length of the time lag and the sample size for the study. Results, however, may indicate that both of the cross-lagged effects are significant, which may suggest that the causal effects work in both ways. Such a result would still leave a question for many researchers as to whether one effect is stronger than the other. It is possible to impose equality constraints on the cross-lagged effects and, using a one-degree-of-freedom likelihood ratio test, investigate whether one cross-lagged path is larger than the other. This test is likely problematic, though, because comparison of the unstandardized coefficients would assume that the variances of the two exogenous variables, y_1 and y_3, are equal to one another. Any arbitrary rescaling of one of the variables, for instance, could lead to a considerably different cross-lagged path estimate in one direction, which could lead to different conclusions about the relative strength of the cross-lagged effects drawn from the nested comparison.

Standardized estimates offer one possible solution to the dilemma. Informal comparisons of the standardized coefficients may be made, but caution is needed in interpreting them. There have been many critics of the standardized coefficients (e.g., Baguley, 2009; Duncan, 1975; Tukey, 1969). The case is most often made against comparing standardized coefficients across groups, because differing variances can be responsible for apparent differences in standardized coefficients even though the unstandardized coefficients do not differ (Kim & Feree, 1981). Similar concerns could be raised about comparing different variables or the same variable over time, though. In particular, changes over time in the variance of one variable but not the other may produce misleading conclusions about the difference in magnitude of the cross-lagged coefficients. Pre-standardizing the variables prior to analysis would equate their variances, but this approach would only hide any issues with unequal variances and loses valuable information about the original metrics of the variables (Cudeck, 1989). Standardized estimates do provide some useful information about relative magnitudes of the effects, as they are connected to common definitions of variance accounted for by a predictor. The wisest course, however, is to inspect and report both the unstandardized and standardized effects, being mindful of the characteristics of the data.

One potential problem with the cross-lagged panel model using observed variables is that autoregressive and cross-lagged effects will be biased by any measurement error present. Measurement error could be accounted for by constructing single-indicator latent variables and setting the measurement residuals using their reliability estimates. It is fairly simple to verify this phenomenon from the correction for attenuation formula for standardized coefficients. For example, it is easy to compare $\beta^*_{23,\text{corrected}} = \beta^*_{23} / \sqrt{\rho_2 \rho_3}$ and $\beta^*_{41,\text{corrected}} = \beta^*_{41} / \sqrt{\rho_4 \rho_1}$ using different reliability estimates, ρ_j, to see that their values could alter the conclusions about the magnitude of the cross-lagged effects (for further discussion of the effects of measurement error on autoregression and cross-lagged effects, see Chapter 4).

Measurement error correction with single indicators could be considered better than assuming there is no measurement error, as is the default with observed variables, but it is also fraught with some potential problems and can easily create misguided conclusions (Deshon, 1998; McDonald, 1996). Nonetheless, there may be rare instances in which multiple indicators are not available but good estimates of reliability are available, and, in such cases, it might be reasonable to use this measurement error correction approach. It may be best, however, to use a range of values of reliability and verify that the overall conclusions do not change with the values.

Example 5.1: Cross-Lagged Panel Model with Observed Variables

A bidirectional cross-lagged panel model with observed variables following the model depicted in Figure 5.1 was tested using the social exchanges data set. Syntax and data sets used in the examples are available at the website for the book. The observed variables used in the analysis were composite indexes based on an equally weighted average of their respective scale items. This model was an extension of the model tested in Example 4.5 that only examined the cross-lagged effect of unwanted advice on positive affect. That model found no significant effect of unwanted advice on positive affect six months later, controlling for the prior level of positive affect. This result is consistent with earlier research showing that negative social exchanges predict negative affect but not positive affect longitudinally (e.g., Newsom, Nishishiba, Morgan, & Rook, 2003). It can be argued, however, that negative social exchanges are at least partly a function of a less positive outlook or that those who tend to be less satisfied engender conflict and evoke criticism, so that the

direction of causality may work in the opposite direction. Results from the analysis indicated the autoregressive coefficients for positive affect, $\beta = .490$, $\beta^* = .465$, $p < .001$, and unwanted advice, $\beta = .344$, $\beta^* = .349$, $p < .001$, were significant. The cross-lagged coefficient for unwanted advice at Time 1 predicting positive affect at Time 2 was nonsignificant as in the earlier unidirectional model, $\beta = -.029$, $\beta^* = -.036$, ns. The cross-lagged effect for positive affect at Time 1 predicting unwanted advice at Time 2, however, was significant, $\beta = -.198$, $\beta^* = -.155$, $p < .05$. These results serve as an example of how investigation of the opposite causal directional hypothesis may be unexpectedly informative.

Comments

By estimating partial effects controlling for earlier time points, the cross-lagged panel model certainly offers advantages over examining cross-sectional relationships or cross-lagged analysis using correlations. As the fit of the model is not relevant in a just identified model with only two waves of data, the cross-lagged panel model with observed measures offers few if any advantages over separate regression analyses. Perhaps the biggest concern is that autoregressive and cross-lagged effects are likely to be biased by measurement error. Moreover, there is a potential problem with conclusions from cross-lagged effects whenever the reliability of the two variables is not equivalent. It is possible to construct examples that alter the significance of the cross-lagged effects and thus the conclusions drawn from them simply by altering the reliabilities of the variables. True values would be even further obscured by reliabilities that change over time for one variable or another. To the extent that specific variance is a component of the observed variance for each variable, stability will be overestimated. Analysis of observed variables has no reasonable method of accounting for such specific variance.

Latent Variables

Cross-lagged models using latent variables with multiple indicators address a number of the challenges associated with the cross-lagged models discussed thus far. Most importantly, measurement error can be estimated with multiple indictors, providing more accurate estimates of autoregressive and cross-lagged effects. Through initial invariance tests and estimation of latent variables in the model, we can also address the potential ramifications of unequal reliabilities or changing reliabilities. An additional advantage of the latent variable approach is the ability to estimate autocorrelations among measurement residuals, which remove stable specific variance that would otherwise inflate our estimate of stability in the construct over time.

An illustration is shown in Figure 5.2. The mean structure is not depicted for this model and may or may not be of interest, depending on the research questions. If the mean structure is estimated, the same identification approach for identifying the factor means should be used as the approach used for identifying the factor variances. The autoregressive and cross-lagged effects can be written as two structural equations:

$$\eta_2 = \beta_{21}\eta_1 + \beta_{23}\eta_3 + \zeta_2$$
$$\eta_4 = \beta_{43}\eta_3 + \beta_{41}\eta_1 + \zeta_4$$

β_{21} and β_{43} represent the autoregressive paths, whereas β_{23} and β_{41} represent the cross-lagged paths. The parameters ψ_{12} and ψ_{43} represent covariances among the exogenous factors and the disturbances, respectively. η_2 and η_4 are latent variables that may have any number of indicators, where I will assume longitudinal equality constraints on all loadings, indicated

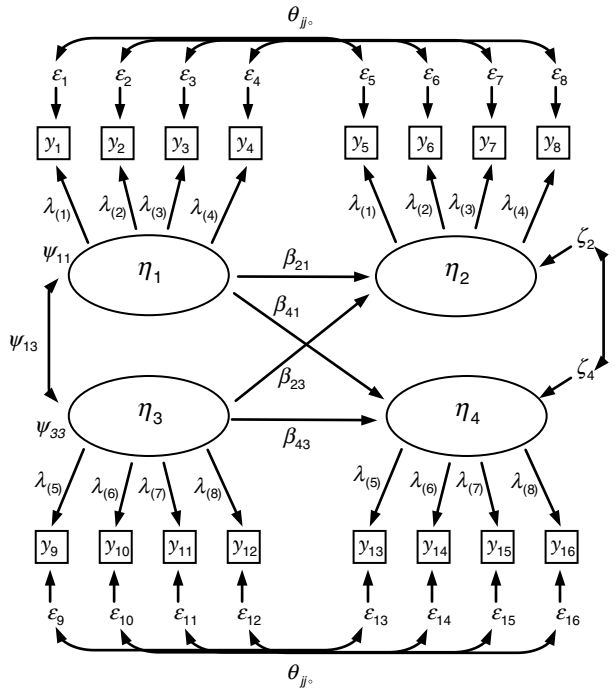

Figure 5.2 Cross-Lagged Panel Model with Two Latent Variables.

by the parenthesis subscript for each loading, (.). I also illustrate the effects coding identification of the factors in this model, which is why no referent indicators are shown. This identification approach is not necessary, however. Correlated errors are included for each repeated pair of indicators, which are designated with $\theta_{jj\circ}$ with $j \neq j\circ$.

Identification. As long as each of the latent variables in the two-wave cross-lagged panel model has more than one indicator, the model will be overidentified and have positive degrees of freedom. In general, for any cross-lagged panel model with all loadings constrained to be equal over time and all pairs of repeated indicators allowed to correlate, the degrees of freedom can be given by the heuristic formula for J observed variables as indicators of the factors.

$$df = \frac{J(J+1)}{2} - [2J + 8]$$

The formula is derived from the standard degrees of freedom rules (Rigdon, 1994) but is abbreviated for the cross-lagged panel model with latent variables as long as the model specification follows the general template shown in Figure 5.2. The second term in brackets is simply the number of freely estimated parameters, q, with $2J$ representing each pair of freely estimated loadings held constant over time plus the number of correlated errors. The number 8 is equal to six structural paths estimated (i.e., factor covariances, autoregressive paths, and cross-lagged paths) plus two exogenous factor variances. The identification approach does not impact the calculation of degrees of freedom. Employing this formula, we can see that even with two indicators of each variable, a cross-lagged panel model would be theoretically overidentified ($df = 12$) if equality constraints on the loadings

were included. Even though theoretically identified, latent variables with two indicators are generally not recommended and are often empirically underidentified, leading to estimation difficulties (e.g., nonpositive definite matrices).

Although it is clear from the calculation of degrees of freedom that any cross-lagged panel model involving latent variables with at least two indicators and specified in this manner will be overidentified, the positive degrees of freedom is due to the measurement portion of the model. The structural portion of the model is saturated, as all possible structural paths are estimated, so model fit is not an indication of the correctness of the predictive paths. After all, if relationships among all latent variables are specified, the structural model is not falsifiable. All other specifications using the same measurement model but different paths, such as directional paths between η_1 and η_3 or between η_2 and η_4 instead of non-directional covariances, are equivalent models and would not be distinguishable from the model in Figure 5.2.

The constraints on factor loadings presuppose investigation of longitudinal measurement invariance, but initial tests of factor invariance specifying factor autocorrelations instead of autoregressive paths can provide useful information about the equality of factor variances over time. Ideally, both factor variances and factor loadings are invariant over time. As we saw in Chapter 1, the two elements are interdependent. Changes in variances, even if loadings are found to be invariant over time, may still signal changes in underlying measurement reliability over time. To the extent that loadings are constrained to be equal in the analysis, any differences in factor variances may complicate the interpretation of the autoregressive effects and the cross-lagged effects. The unstandardized regression coefficient is partially a function of the predictor variance, and the standardized regression coefficient is a function of both the predictor and the outcome variances. I have discussed cross-lagged panel models with all observed variables or all latent variables, but it also is possible to investigate the directional relationships between one observed variable and one latent variable. As long as the observed variable might contain measurement error, however, inferences about the relative effects of the cross-lagged paths can be biased by the unequal measurement error in the two variables.

Correlated measurement residuals may occur between different variables, which could have biasing effects on cross-lagged estimates if ignored in the model. An indicator of η_1 and η_3 may be associated with one another over and above the association between their common factors. For example, measures of social support and depression sometimes both contain a question related to loneliness. In a model examining the cross-lagged effects of social support and depression over time, the covariance between the two factors would be inflated if specific variance related to the loneliness indicator of each factor was correlated. Cross-lagged effects would be overestimated to the extent that specific variance for the two indicators was not accounted for by inclusion of measurement residual covariances between indicators of the social support and depression factors in the model. Constraining these covariances between measurement residuals to be equal over time would make sense to ensure consistency in estimates over time.

Example 5.2: Cross-Lagged Panel Model with Multiple Indicators

Using latent variables instead of observed variables, a cross-lagged panel model parallel to the model tested in Example 5.1 was specified to investigate the bidirectional effects of positive affect and unwanted advice. For this model (following Figure 5.2), factors were identified using effects coding (Little, Slegers, & Card, 2006), loadings for repeated indicators were set equal over time, and correlated measurement residuals among repeated indicators were estimated. The model chi-square was significant, $\chi^2(96) = 226.516$, $p < .001$,

but alternative fit indices suggested acceptable fit, CFI = .959, SRMR = .042. The autoregressive paths were significant, $\beta = .600$, $\beta^* = .565$, $p < .001$, for positive affect and $\beta = .390$, $\beta^* = .414$, $p < .001$, for unwanted advice, although their magnitudes were larger. As with the prior model, the cross-lagged effect of positive affect on unwanted advice was significant, $\beta = -.227$, $\beta^* = -.182$, $p < .001$, but the effect of unwanted advice on positive affect was not, $\beta = -.015$, $\beta^* = -.019$, ns. The model also provides information about the synchronous correlation between the two variables at Time 1 ($\psi_{13} = -.220$, $p < .001$) and the correlation among the disturbances ($\psi_{24} = -.144$, $p < .10$). This latter marginally significant result suggested that there may be some remaining association between the two factors after accounting for their prior time points. The total variance accounted for in positive affect and unwanted advice at Time 2 was approximately 32% and 24%, respectively. Although this model differed from the observed variable model in Example 5.1 in that the factors are a function of unequal weighting of indicators (see Chapter 1 for details), the primary differences in the magnitude of the coefficients stems from the fact that measurement error and correlated specific variance are taken into account in the latent variable version of the model.

As an illustration of how scaling of the factor can influence the model results, the same latent variable cross-lagged model was tested using referent loading identification instead. The model fit, degrees of freedom, statistical significance tests, and standardized solution for this model were identical to those obtained when the effects coding identification approach was used. The unstandardized estimates for the cross-lagged paths, however, differed somewhat from the effects coding model, $\beta = -.246$ for positive affect predicting unwanted advice and $\beta = -.014$ for unwanted advice predicting positive affect, because the factor variances differed as a result of the identification approach. Use of the single occasion identification method (setting exogenous factor variances equal to 1) would also be possible with this model and would lead to identical fit, standardized solution, and significance tests. The unstandardized cross-lagged effects would differ as a consequence of how the factor variances are defined.

Inclusion of Covariates

Discussion thus far has focused only on the longitudinal association of two conceptual variables. Rarely in other contexts do we consider such simple models in the absence of other variables, yet many applications of cross-lagged panel models omit covariates. As with the association between any two variables, third variables may be partially or completely responsible for the relationship. Consider a longitudinal study of employee health and productivity. The amount of sleep may be related to general health and to productivity. Omission of sleep may affect the covariance of health and productivity at any time point or could impact the cross-lagged relationships among the two variables. Continuing problems with sufficient rest over time may impact the stability of health as well.

Covariates can be included in cross-lagged panel models at any time point. Let us begin by considering the inclusion of a covariate at the initial measurement occasion, as illustrated in Figure 5.3. As with any multiple regression model, partial regression coefficients are generally biased to the extent that a covariate is associated with other predictors in the model and the outcome but omitted from the model.[3] For the autoregression effect (β_{21}), this implies that if the covariate η_5 is correlated with η_1 and η_2, then the autoregressive path will be biased if the covariate is not included in the model. For example, if η_1 and η_2 represent health and η_5 represents sleep, health will appear more stable to the extent that sleep is related to health at both time points. Persistent poor health may be due to

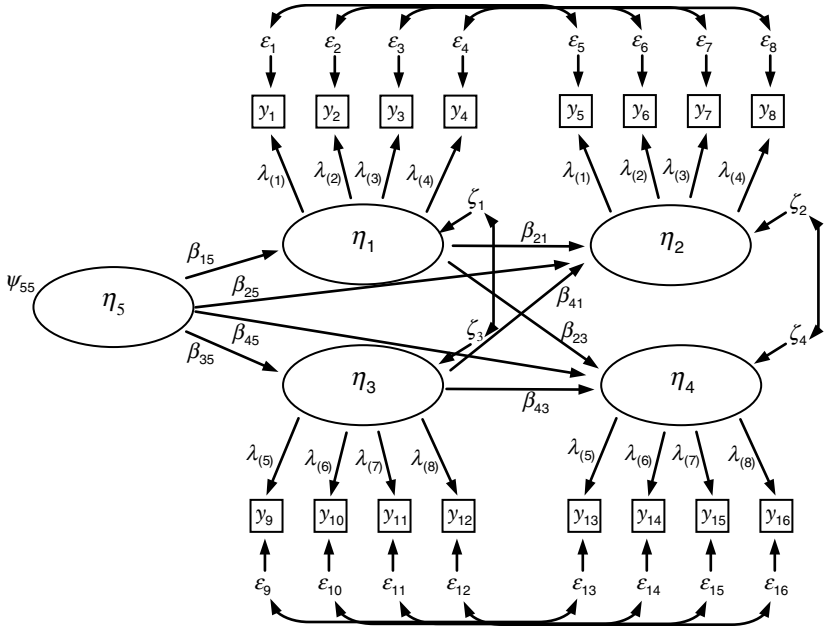

Figure 5.3 Cross-Lagged Panel Model with Time-Invariant Covariate.

persistent poor sleep. This phenomenon can also be viewed as the classic omitted variable (or misspecified model) problem, where the disturbance, ζ_2, will be correlated with the predictor η_1 if the covariate is not in the model.

When considering the cross-lagged effect of η_1 predicting η_4, the covariate, η_5, would need to be associated with η_1 and η_4 in order to have a substantial impact on the cross-lagged path, β_{41}. If η_5 is sleep, its omission would bias the cross-lagged effect of health on productivity only to the extent that sleep affects productivity at a later time point controlling for initial levels of productivity and initial levels of health. Such a relationship is not inconceivable if there were a causal relationship between sleep and productivity over and above any effects of health on productivity. Unless there is a strong theoretical rationale, it will usually make sense to estimate a lagged path between the covariate and both endogenous variables, rather than just one (i.e., η_2 and η_4, as depicted in Figure 5.3), because there may be a causal effect of the covariate on either outcome. Obviously, measurement error in the covariate will diminish its role as an adequate control variable, and autoregression and cross-lagged estimates will be biased to the degree that measurement error is not taken into account. Multiple indicator latent variables included as covariates will provide the most accurate estimates of the parameters in the model.

When the covariate is repeatedly measured over time, a three-variable cross-lagged panel model can be estimated (discussed in greater detail below under "Additional variables"). This is an extension of the previous model, but including the covariate as a third outcome. Such models become complex quickly, especially if other covariates or model features are included. With the addition of more variables, the analyses become more exploratory and less focused on a theoretically driven analysis. An important possible goal of a three-variable cross-lagged panel model is to examine mediational pathways, a special topic addressed below.

Example 5.3: Cross-Lagged Panel Model with Covariate

To illustrate the inclusion of a covariate, a different cross-lagged model was tested (as depicted in Figure 5.2). This model examined the association of companionship support with depression. A wide array of studies has found that higher social support is associated with lower depression, although there have been inconsistent findings when this direct relationship has been examined longitudinally. It is possible that one reason for this association is that those who are more depressed initially receive less support because they are avoided by others or they are less likely to seek out companionship from others. The model used WLSMV estimation that treated the indicators for the depression scale, measured on a 4-point scale, as ordinal. The fit of this model overall was good, although the chi-square value was significant, $\chi^2(69) = 120.979$, $p < .001$, CFI = .973, WRMR = .939.[4] An initial cross-lagged model without a covariate suggested that companionship did not predict depression, $\beta = -.081$, $\beta^* = -.067$, ns, but that depression significantly predicted companionship, $\beta = -.109$, $\beta^* = -.130$, $p < .05$. This result appears to be consistent with the hypothesis that depressed individuals may receive less support because others avoid them or they do not seek out companionship.

A subsequent cross-lagged panel model was then tested using positive affect measured at Time 1 as a covariate predicting both Time 2 outcomes (see Figure 5.3). This time-invariant covariate might address the hypothesis that a diminished joie de vivre might be confounded with the longitudinal effect of depression on companionship. This model also had adequate fit, $\chi^2(140) = 245.386$, $p < .001$, CFI = .959, WRMR = .959. Results indicated that the cross-lagged effect of depression on companionship was no longer significant after accounting for initial levels of positive affect, $\beta = .007$, $\beta^* = .009$, ns. Positive affect at Time 1, on the other hand, significantly predicted depression, $\beta = -.406$, $\beta^* = -.190$, $p < .01$, and companionship at Time 2, $\beta = .368$, $\beta^* = .261$, $p < .001$. These results underscore the importance of including relevant covariates in the cross-lagged panel model.

Synchronous Effects

Synchronous effects are paths between variables measured at the same occasion. Up until this point, we have discussed only non-directional relationships (i.e., covariances or correlations) between variables or disturbances at a given occasion. Structural models can certainly incorporate unidirectional paths that are synchronous as well. Generally unidirectional paths are included in only one direction when there is a theoretical rationale for a particular causal direction. Two directional paths between a pair of variables pointed in opposite directions (e.g., x_t predicts y_t and y_t predicts x_t), known as *non-recursive effects*, are also sometimes proposed, but tend to be difficult to appropriately estimate, because of predictor–disturbance dependence and identification issues. To identify such a model, instrumental variables that are highly related to each of the dependent variables that also are independent of the predictor are needed. As non-recursive models are generally cross-sectional in nature and there are few data sets that can meet these requirements for proper identification, I do not discuss them further (for discussions, see Berry 1984; Heise, 1975; Mulaik, 2009).

Incorporation of unidirectional synchronous paths in the context of longitudinal models is common and may be desired for a variety of reasons. The simple model depicted in Figure 5.4 provides a useful starting place for discussion. In this model, assume that measurement of x_2 is only available at time t, whereas, we have repeated measures of the dependent variable, y_1 and y_2. Thus, there is an autoregressive effect for y, but the effect of x_2 is synchronous rather than lagged in time. How does the effect of the predictor

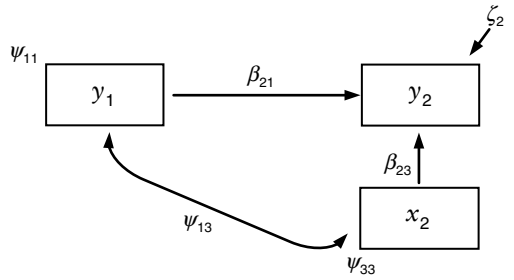

Figure 5.4 Synchronous Covariate at One Time Point.

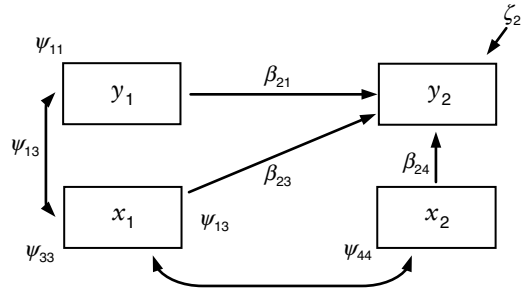

Figure 5.5 Synchronous Covariate at Two Time Points.

x_2 relate to the cross-lagged effect discussed earlier? Although we predict non-stable variance in y, the cross-lagged effect from x_1 in the earlier model holds constant values of y_1 also measured at the earlier time point. The (sometimes) desirable feature in the synchronous path from x_2 is that its instantaneous effect is modeled, but an undesirable feature is that the covariance between x_2 and y_1 may be weaker than the covariance between x_1 and y_1. The additional gap in measurement of y_1 and x_2 in comparison to a model with x_1 and y_1 as predictors may be that each predictor only partially controls for the other.

Now consider a slightly modified model (Figure 5.5) in which x_1, x_2, and y_1 are predictors of y_2. In this model, any stable variance in x is removed, because x_1 and x_2 control for one another, resulting in variance for x_1 and x_2 that is unique to their respective occasions. The stable and unstable effects of x then contribute to the accounted for variance in y_2, and the model estimates the independent effects of the lagged and synchronous predictor. If x is highly stable over time (e.g., $r > .95$), collinearity may be a concern (Cohen, Cohen, West, & Aiken, 2013), but this will generally not be the case. In such instances, the fact that the predictor is highly stable over time would suggest similar results regardless of when it is measured. Relatedly, it should be kept in mind that the more stable the predictor, the more difficult it will be to find significant independent effects of its two measurements even though they may account for substantial total variance in y together.

A model with synchronous effects at each time point has some interesting relationships to other analyses. In the contexts of growth curve models or survival analysis, including an estimate of a synchronous effect of x at each time point is referred to as a *time-varying*

covariate. Consider a modification to the model in Figure 5.5, with x_1 predicting y_1 and x_2 predicting y_2. In equation form, we have two equations:

$$y_1 = \beta_{11}x_1 + \zeta_1$$
$$y_2 = \beta_{22}x_2 + \zeta_2$$

There are two cross-sectional regression models, in which the magnitude of the effects, β_{11} and β_{22}, may differ at each time point. If the two coefficients are constrained to be equal over time, however, assuming that the effect of x on y at each time point is the same, then we can show that this model is related to the difference score model, with $y_2 - y_1$ regressed on $x_2 - x_1$, if we subtract the second synchronous regression equation from the first synchronous regression equation:

$$\begin{aligned} y_1 &= \beta_{(1)}x_1 + \zeta_1 \\ y_2 &= \beta_{(1)}x_2 + \zeta_2 \\ y_2 - y_1 &= \beta_{(1)}(x_2 - x_1) + (\zeta_2 - \zeta_1) \end{aligned} \tag{5.2}$$

Which shows that $\beta_{(1)}$ obtained for a model with synchronous effects at each time point in which the paths are set equal produces the same coefficient that would be obtained for a regression with both predictor and outcome calculated as difference scores. This model is referred to as the *fixed effects regression model* by Allison (2005). The fixed effects regression coefficient can be estimated from a regression analysis where the data set has been restructured to have two records for each case (i.e., a person × period or long-format data set), the format used for multilevel regression modeling. Notice also that, because the regression coefficient represents a simple regression, the standardized coefficient for the relationship between x_t and y_t would be the same regardless of which variable was the predictor and which variable was the outcome. This would not be true about standardized coefficients for the cross-lagged effects obtained in the cross-lagged panel model. So, the correspondence between change in x and change in y in the difference score regression or the fixed effects regression analysis does not investigate the hypothesis about which variable causally precedes the other.

To obtain the equivalent fixed effects analysis in SEM, a model (Figure 5.6) with special latent variable, η_1, can be added with each y_t as its indicators with loadings, λ_{jt}, set equal to 1 (Allison, 2005; Bollen & Brand, 2010; Teachman, Duncan, Yeung, & Levy, 2001). The factor variance is estimated (the factor mean is set equal to 0 if mean structure is included) and the correlations among η_1, x_1, and x_2 must be estimated. The synchronous paths, $\beta_{(1)} = \beta_{31} = \beta_{42}$ and the disturbances, $\zeta_{(1)} = \zeta_3 = \zeta_4$, are constrained to be equal. The latent factor partials stable variance out of the synchronous effects, so that the $\beta_{(1)}$ path estimate represents the effect of $x_2 - x_1$ on $y_2 - y_1$. To include means in the model, the factor mean must be set equal to 0 and the measurement intercepts for y_1 and y_2 must be freely estimated.

Modifications can be made to the fixed effects model to obtain random effects or a combination of random and fixed effects (random effects will be discussed in greater detail in Chapter 7 in connection with latent growth curve models). The extension to three or more waves is a simple matter, with the same specifications for the latent variable and synchronous effects. The random effects variant of the model with three or more waves, allowing for correlations among disturbances over time, tests the same hypotheses as a generalized estimating equations model (GEE; Liang & Zeger, 1986). Although I have illustrated synchronous effects among observed variables, the same principles and analogous model naturally can be extended to synchronous latent variables.

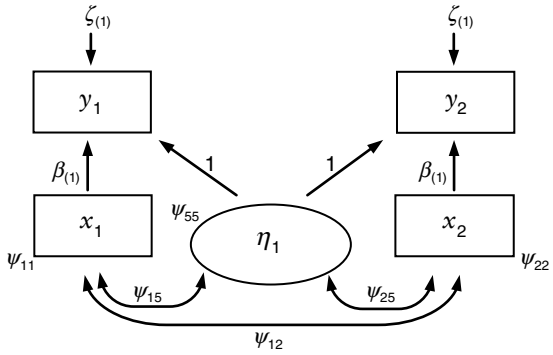

Figure 5.6 Latent Fixed Effects Model.

Example 5.4: Synchronous Covariates

A two-wave model with composite measures of positive affect regressed on unwanted advice at each time point was used to demonstrate the fixed effects analysis. The specifications followed those described above and the model depicted in Figure 5.6. This model had a good fit, with $\chi^2(2) = 3.489$, $p = .175$, CFI = .992, SRMR = .025. The synchronous effect estimate was nonsignificant, $\beta = -.026$, $\beta^* = -.033$, ns, indicating a .026 decrease in positive affect difference score for each unit increase in the unwanted advice difference score. In other words, for those with increases in unwanted advice over the six-month period, there was a slight and nonsignificant decrease in positive affect over the same period. The equivalent fixed effects model can be estimated in several ways, including through a standard OLS model with $y_2 - y_1$ difference scores regressed on $x_2 - x_1$ or by converting the data to a person × period format and regressing y on x and time ($t = 0, 1$) using general linear model procedures that take into account the inflated sample size and observation dependence (Allison, 2005).

More than Two Time Points

Conclusions about the relationships among variables across just two time points may be incorrect when transient factors, such as historical events, changing moods, or cultural zeitgeists, impact the specific observations available. In some cases, the relationship estimated over two particular observed waves may not be representative of a more typical or true relationship that exists more generally. Nothing is known about the variables preceding or following the study. The values observed at the first wave, for instance, are likely to be at least partly determined by transient causes, making them fallible representations of their true values at other points in time. Thus, to the extent that initial values are fallible, estimates of autoregressive or cross-lagged values that attempt to statistically control for them are fallible as well (Rogosa, 1988). With each additional wave of observation, however, greater certainty and reliability of the estimates of stability and change is gained. If any particular wave is affected by transient factors making observations atypical, this will be evident from the data. Analyses of three or more time points therefore have important advantages over analyses of just two time points, and I discuss a number of the issues related to extending models described thus far to three or more time points next.

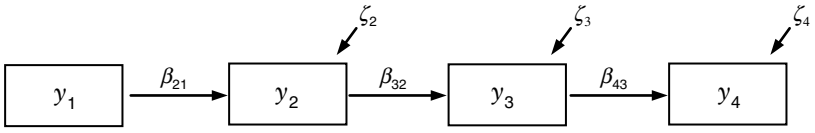

Figure 5.7 Perfect Simplex Model.

Simplex Models

The "Perfect" Simplex Model. The general concept of the simplex model was developed by Guttman (1954) for longitudinal data in the social sciences and was popularized as an analytic method for understanding stages of growth (Jones, 1959). The general concept, which can be applied to any ordered system, is already familiar to us as the autoregressive effect, in which the state at one point in time determines the state at the next point in time.[5] This gives us information about the process of growth over time if many time points can be linked together. A common application is learning, in which the state of knowledge at some initial period impacts learning at the next period. The term *simplex* describes the implied correlation structure for lags greater than 1. For the model pictured in Figure 5.7, the standardized regression coefficients are equal to the correlation coefficient between each adjacent, or lag 1, measurement, $\beta^{*}_{t,t-1} = r_{t,t-1}$. If we use path tracing rules, the structure of the model implies that the lag 2 correlation is the product of the two paths that link the variables (e.g., $r_{13} = \beta^{*}_{21} \times \beta^{*}_{32}$), as with the indirect effect. If we assume each lag 1 correlation is equal across all of the adjacent time points, this implies that the number of lags is exponentially related to the expected correlation implied by the model, so that $\hat{r}_{t,t-\ell} = r^{\ell}_{t,t-1}$, with the symbol ℓ representing the number of lags. For example, if the lag 1 correlation is .6, then the lag 3 correlation will be $r^{3}_{t,t-1} = (.6)^{3} = .216$ if the simplex model holds true (assuming no sampling error). It should be immediately apparent that the expected correlations rapidly decrease with increasing lag lengths.

If the model depicted in Figure 5.7 is correct and the simplex correlation structure is true, then we can also state that the partial regression coefficient for all lags greater than 1 will be equal to 0. In other words, we assume the autoregressive process is only lag 1 and that no remaining variance at time t is accounted for by the variable at $t-2$ or earlier. A useful observation is that a common confirmatory factor model with three or four indicators is empirically indistinguishable from the simplex model. This can be verified by considering the decomposition of the covariances with path tracing rules for both models (Jöreskog & Sörbom, 1979; Marsh, 1993). Rovine and Molenaar (2005) show that the connection also holds beyond just a few variables. The difference theoretically may be a subtle, if not an elusive, one. The autoregression/simplex model posits a causal hypothesis of inertial effects, with one state causing values of the next state, whereas the common factor model suggests a stable underlying latent trait that causes all observations. The latent state-trait-error model used to capture such stable characteristics is described in Chapter 6.

A variation on the model is to set the autoregression coefficient, $\beta_{t,t-1}$, equal to 1 for all intervals. This leads to an estimate of the residual, considering an individual case, that is equal to the difference between the two measurements, $\zeta_{it} = y_{it} - y_{it-1}$. The equivalence of the difference score and the residual from autoregression parameter equal to 1 was discussed in relation to repeated measures models using the latent difference score approach (refer to Equation [4.2]). The simplex model, therefore, can be used to estimate change

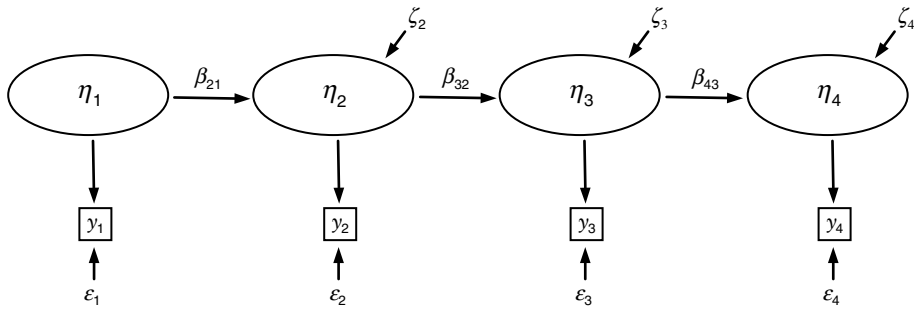

Figure 5.8 Quasi-Simplex Model.

over each interval, one approach to understanding growth. The relation of the difference score model to growth over time will be revisited in Chapter 7.

The Quasi-Simplex Model. The simplex model described above assumes no measurement error, and the model can be modified to estimate measurement error at each time point, even with only one observed variable at each occasion. The measurement error model, depicted in Figure 5.8, is called the *quasi-simplex model* or, sometimes the "quasi-Markov simplex model," and was suggested by Jöreskog (1970). The model is identified if the factor variance at the first occasion is set equal to 1 and the measurement residual variance for the first and last occasion are set equal to 0. The reason these constraints identify the model is that the factor variances are a function of the autoregressive path, the variance of the factor at the prior time point, and the covariance of the adjacent observed variables (Jöreskog & Sörbom,1989). Because not all of this information is available for the first and last time points (i.e., there is factor indeterminancy), constraints are required for identification. The quasi-simplex model must have at least four time points to be identified. An estimate of the measure's reliability and the reliability of differences scores can be computed from the quasi-simplex model (see Rogosa & Willett, 1985; Werts, Linn, & Jöreskog, 1977). Means can be added as well, as noted by Mandys and colleagues (Mandys, Dolan, & Molenaar, 1994), who discuss the relationship to the linear growth curve model. It should be noted that quasi-simplex models can be difficult to estimate because of empirical identification problems (Marsh, 1993) and may be not be indicative solely of the appropriateness of the simplex structure (Rogosa & Willett, 1985).

Binary Variables. A related model that follows the simplex path specification for binary variables derives from the concept of the *Markov chain process*. The concept of the Markov chain is a broad one and one that has been applied in many statistical, physical science, and systems theory applications. The basic model, which can also be referred to as the discrete-time discrete-space Markov model, is for a set of binary repeated observed variables. For a Markov chain model with observed variables and an assumed lag of 1, we can express the joint probability that $y = 1$ across two time points as a product of the probability at time $t-1$, P_{t-1}, and the conditional probability that $y = 1$ at time t, $P_t|P_{t-1}$.

$$P_{t,t-1} = P_{t-1}\left(P_t \mid P_{t-1}\right)$$

This relation parallels the simplex structure for expected correlations under the simplex process. The conditional probabilities, $P_t|P_{t-1}$, contain information about the probability that the outcome will be different at the next occasion, and are therefore referred to as

transition probabilities. We have already seen how the autoregression of a binary variable at a later time point onto the same binary variable measured at a prior time point can be transformed to provide information about odds ratios and how the intercepts can be used to obtain predicted probabilities using the cdf to transform the obtained estimates. The Markov chain model for binary data will be discussed further in Chapter 10 as part of the consideration of latent transition analysis. The simplex correlation structure and the Markov chain process come up again in the discussion of autoregressive process in time series analysis in Chapter 11. The same modeling concepts can also be applied to ordinal variables, where tests of longitudinal invariance of thresholds may be needed (Muthén, 1993).

Multiple Indicators. The simplex model can be extended to latent variables with multiple indicators at each occasion (Marsh, 1993). The primary advantages of simplex models with latent variables is the disattenuation of autoregression effects. Another advantage is to allow for correlated measurement residuals. Sivo and Willson (2000) illustrate that the simplex model can be respecified as a time series model that incorporates correlated measurement residuals over time, at least for the specified lag length (Chapter 11).

Example 5.5: Simplex Models

To illustrate several simplex models, we turn to another data set – the health and aging data set. These data are derived from a national health survey with interviews of individuals aged 50 years and over conducted biannually. More information on this data set is described in Example Data Sets at the beginning of the book. Analyses were of the self-rated health question about overall health collected over six waves. Ratings from this question were from 1 "poor" to 5 "excellent." A perfect simplex model was tested first (as shown in Figure 5.7). This model had a poor fit to the data, $\chi^2(10) = 3{,}564.560$, $p < .001$, CFI = .832, SRMR = .202. Because the simplex model will fit when the observed correlation pattern conforms to the simplex structure, the poor fit of the model suggests that the observed correlations do not decrease exponentially with additional lag lengths. As illustrated by the correlation of the Time 1 with each later time point, $r_{12} = .619$, $r_{13} = .669$, $r_{14} = .626$, $r_{15} = .607$, $r_{16} = .585$, the correlations decline at greater lag lengths though not at an exponential rate. The standardized autoregression coefficients, equal to the observed correlations in this model, were $\beta^*_{21} = .619$, $\beta^*_{32} = .677$, $\beta^*_{43} = .683$, $\beta^*_{54} = .707$, and $\beta^*_{65} = .712$. The increasing values due to cumulative increases in stability (or inertia) estimates are commonly observed in simplex models (Dwyer, 1983).

The quasi-simplex model (see Figure 5.8) fit the data considerably better, $\chi^2(6) = 23.206$, $p < .001$, CFI = .999., SRMR = .005. Estimating measurement error appeared to have improved the fit and may suggest better conformity of the model to the simple structure, though the good fit does not necessarily distinguish this model from alternative models (Rogosa & Willett, 1985). Latent variables at each time point are estimated with a single indicator in the quasi-simplex model, but, for identification purposes, two of the measurement residual variances must be set equal to 0 for two time points (usually the first and last). As a consequence, parameter estimates involving the first and last variable do not take into account measurement error. The estimated correlations among latent variables can be obtained in the output from some programs, and the estimates from this model suggest conformity to the simplex correlation structure. If we examine these correlations for the middle latent variables, we see a close match to the pattern expected from the simplex structure. For example, inspection of latent factor correlations indicates that correlations between Time 2 and Time 4, $\psi_{24} = .906$, and between Time 2 and Time 5, $\psi_{25} = .860$, are

close to the value obtained if the latent variable correlation between Time 2 and Time 3, $\psi_{23} = .955$, is raised to the second and third power, respectively, $\hat{\psi}^2_{24} = (.955)^2 = .912$, and $\hat{\psi}^2_{25} = (.955)^3 = .871$.

The standardized autoregression coefficients from the quasi-simplex model were $\beta^*_{21} = .821$, $\beta^*_{32} = .955$, $\beta^*_{43} = .948$, $\beta^*_{54} = .949$, $\beta^*_{65} = .823$. The first and last autoregression coefficients are considerably lower than the middle autoregression coefficients, a result of the assumption of no measurement error assumed at the first and last time points. Notice also that there is little evidence of cumulative increase in the stability coefficient across the middle three autoregression coefficients once measurement error is taken into account in the model.

Cross-Lagged Panel Models

A simple extension of autoregressive models or cross-lagged panel models is the addition of measurement occasions beyond two time points (Figure 5.9). For the cross-lagged panel model of observed variables measured at three or more time points, the model becomes overidentified, allowing for additional parameters. In addition to the autoregressive or cross-lagged paths between adjacent time points (i.e., lag 1 or *first-order autoregression*), paths that span two or more time points become possible to estimate (shown with dotted lines). The lag 2 (second-order) autoregressive path in a three-wave model, for example, represents the effect of the variable across two units of time over and above the effect of the variable at lag 1. Analogous concepts can be applied to cross-lagged paths with any lag greater than 1. One can think of such paths as representing hypothesized delayed causal effects. For example, exposure to a cold virus will not be manifested as symptoms until a week or more after exposure. If one lag length is shorter than this period, the association between exposure and symptomatology will not be evident until subsequent lag lengths are considered.

With three or more waves, equality constraints on autoregressive, cross-lagged, or disturbance covariances can also be tested. That is, for the model illustrated in Figure 5.9, the model can be estimated assuming $\beta_{21} = \beta_{32}$, $\beta_{54} = \beta_{65}$, $\beta_{24} = \beta_{35}$, $\beta_{51} = \beta_{62}$, $\psi_{25} = \psi_{36}$. Nested model comparisons to determine whether these parameters are equal over time can be conducted, either tested individually or tested with several simultaneous constraints. With an increasing number of time points, the potential for causes specific to an occasion may increase, resulting in parameter estimates that differ over time. When parameters are equal, they are said to exhibit *stationarity*. Stationarity can be distinguished from stability. Stationarity involves equality or invariance in relationships among variables over time (autoregression, cross-lagged, or covariances), whereas stability involves correspondence in the values of variables over time (i.e., a difference score near 0 or a large autocorrelation). When parameters differ over time, it is possible to allow for unique estimates at each time point. There is no requirement, as such, that equality constraints be imposed on structural parameters over time. Equality constraints are not appropriate for unequally spaced intervals, but nonlinear constraints can be incorporated for unequal intervals (Arminger, 1987).

In many instances, finding that cross-lagged paths or autoregressive paths differ over time may not have a substantial impact on interpretation of results. For example, if for the model in Figure 5.9, the cross-lagged path at the first interval, β_{24}, differs from the cross-lagged path between the subsequent interval, β_{35}, in magnitude, yet both are significant, this may not affect the conclusions about the direction of the cross-lagged effect that is of interest. On the other hand, if β_{24} was significant but β_{35} was not, it may be difficult to conclude that the results unequivocally support this direction of causation.

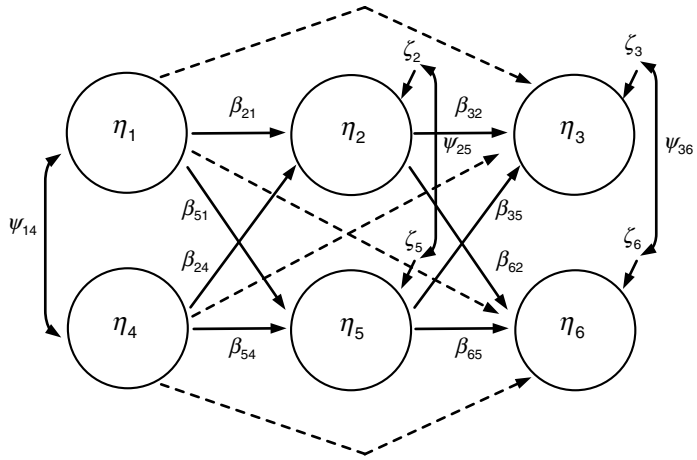

Figure 5.9 Cross-Lagged Panel Model with Three Time Points.

Differences in estimates of the covariances among disturbances over time may be the least consequential (Dwyer, 1983), and thus it is rare to impose equality constraints across time points on disturbance covariances. Although such differences may suggest that external occasion-specific variables, such as life events, moods, or temporary economic changes, may be informative about the relationship at a certain time point, this may have little bearing on the hypotheses of interest in the model. Differences in autoregressive parameters over time may be more troubling. When the autoregressive effect for a certain interval is inflated by a confounding external factor, the cross-lagged estimate for the interval will be diminished. The potential interpretational complexity that arises when results are not consistent over time can be vexing, but the advantage to having additional time points is that a more accurate picture of the lagged effects is possible.

Temporal System Concepts. Although additional time points in a cross-lagged model may aid the formulation of a more accurate picture of the true association between two variables over time, it is important to be aware that such models constitute a dynamic system that are also sensitive to complex mathematical properties. Borrowing from physics, Dwyer (1983) refers to the tendency for a variable to remain stable over time, to exhibit *temporal inertia*, which is estimated with autoregressive (stability) paths in the cross-lagged panel model. In physics, an opposing force to stability is the tendency for disorder to increase over time, referred to as *entropic decay*. The concept of entropic decay has been used as an extrapolation of regression toward the mean observed over a longer series. A system is said to have reached *equilibrium* when the values of a variable are unchanging over time and said to be invariant when the parameters are unchanging over time. These concepts are helpful for understanding some of the phenomena often observed in models across many waves.

With a longer series, increasing (or, less commonly, decreasing) tends in parameter estimates can be observed that result from cumulative effects in the system over time. Dwyer (1983), for example, demonstrates that even for a bivariate cross-lagged system that is truly in equilibrium, autocorrelations, cross-lagged correlations, and synchronous correlations tend to increase over time. Typically, the increases taper off in later waves and the system stabilizes. It is important to be aware of the cumulative increases in a system,

because such increases may affect conclusions about stationarity of the parameters and potentially generate misleading conclusions from the data. A highly stable variable in a two-variable system, for example, can appear to have a larger magnitude cross-lagged effect when predicting a less stable variable. This phenomenon occurs as a result of the differential stability in the two variables and differential rates of increasing stability. Cumulative effects also occur in shorter time series, but they are likely to be less evident.

Example 5.6: Three-Wave Cross-Lagged Panel Model

A cross-lagged panel model between positive affect and unwanted advice extending Example 5.2 to three waves was tested using latent variables for each measure. The initial model examined only lag 1 autoregression and cross-lagged panel effects. Results indicated the model did not entirely reach conventional recommendations for fit, $\chi^2(230) = 575.159$, $p < .001$, CFI = .936, SRMR = .056. The cross-lagged effects for unwanted advice predicting positive affect was nonsignificant at the first interval, $\beta = .000$, $\beta^* = .000$, ns, but was significant at the second interval, $\beta = -.094$, $\beta^* = -.107$, $p < .05$. The effect of positive affect predicting unwanted advice was significant at the first interval, $\beta = -.244$, $\beta^* = -.198$, $p < .001$, but not significant at the second interval, $\beta = -.031$, $\beta^* = -.027$, ns. A subsequent model constrained the autoregressive and cross-lagged effects each to be equal over time, $\chi^2(234) = 590.202$, $p < .001$, CFI = .934, SRMR = .059, which indicated a significantly poorer fit, $\Delta\chi^2(4) = 15.043$, $p < .001$. The significant difference suggested that the autoregressive and/or the cross-lagged effects were not equal over time, although the magnitude of this difference was relatively small, $w = .081$, $\Delta Mc = .007$. With the assumption of invariant autoregressive and cross-lagged paths, only the effect from positive affect to unwanted advice remained significant, $\beta = -.128$, $\beta^* = -.098$, $p < .001$.

The poor fit of the first model may be a result of assuming no autoregressive or cross-lagged paths between the first and third waves. Allowing these additional paths resulted in a considerably smaller chi-square indicating better model fit, $\chi^2(226) = 501.337$, $p < .001$, CFI = .949, SRMR = .046. Inclusion of these additional paths did not change the general pattern of the results for the lag 1 interval cross-lagged effects. Positive affect at Time 1 was a significant predictor of unwanted advice at Time 2, $\beta = -.239$, $\beta^* = -.192$, $p < .001$, but positive affect at Time 2 did not significantly predict unwanted advice at Time 3, $\beta = -.005$, $\beta^* = -.004$, ns. Unwanted advice at Time 1 did not predict positive affect at Time 2, $\beta = -.015$, $\beta^* = -.019$, ns, but unwanted advice at Time 2 was a marginally significant predictor of positive affect at Time 3, $\beta = -.075$, $\beta^* = -.086$, $p = .084$. The autoregressive paths from Wave 1 to Wave 3 were significant for both variables, suggesting that both positive affect and unwanted advice may have an inertial effect over a one-year period that is over and above the inertial effect over a six-month period. Although the cross-lagged effect of unwanted advice at Time 1 on positive affect at Time 3 was not significant, $\beta = -.041$, $\beta^* = -.033$, ns, the effect of positive affect at Time 1 had a marginally significant positive effect on unwanted advice at Time 3, $\beta = .072$, $\beta^* = .088$, $p = .061$. The latter path, although not quite significant, is suggestive of suppression, because the negative relationship is reversed after controlling for the effects of Time 2 variables.

Common Factors and Omitted Variables

Inclusion of covariates is one strategy for addressing the potentially biasing effects of external variables that can affect stability or cross-lagged estimates. Available data never contain all possible factors that may have confounding effects, however, so any model is likely to be subject to omitted variable bias. With three or more time points, it is possible

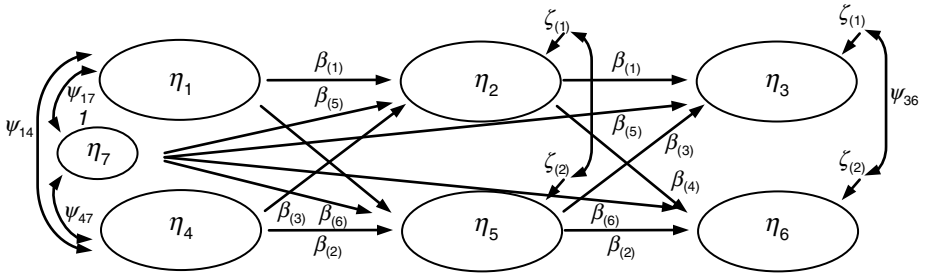

Figure 5.10 Cross-Lagged Panel Model with Static Unmeasured "Phantom" Variable.

to test for spurious associations in the cross-lagged panel model that are due to omitted variables. There are two general strategies. One is to model a static "phantom" variable that represents all stable unmeasured covariates. The other is specification of a common factor that accounts for observations at each time point.

The first strategy (Dwyer, 1983) tests for omitted variables by specifying a static unmeasured, or "phantom," variable that predicts both cross-lagged variables at each time point (see Figure 5.10). To identify the phantom factor, which has no indicators, several of the paths must be set equal. In particular, it must be assumed that the phantom variable, η_7, has a variance of 1 and has the same effect on x and y at each time point (e.g., $\beta_{(5)}$ and $\beta_{(6)}$). In addition, autoregressive paths ($\beta_{(1)}$: $\beta_{21} = \beta_{32}$ and $\beta_{(2)}$: $\beta_{54} = \beta_{65}$), cross-lagged paths ($\beta_{(3)}$: $\beta_{24} = \beta_{35}$ and $\beta_{(4)}$: $\beta_{51} = \beta_{62}$), and variances for the disturbances must all be set equal ($\zeta_{(1)}$: $\psi_{22} = \psi_{33}$, and $\zeta_{(2)}$: $\psi_{55} = \psi_{66}$). Although this model can be specified with single observed variables at each time point, a model with latent variables at each occasion will have superior estimates of the autoregressive and cross-lagged effects as well as more accurate estimates of the effects of the phantom variable. The specification of the phantom variable is similar to the estimation of a second-order factor model. Even with these constraints, a model with single observed variables at each wave may have empirical identification problems, particularly with only three waves. Moreover, the assumptions required to identify the model are not testable and may not be valid.

Figure 5.11 illustrates the second strategy that incorporates a synchronous common factor (Finkel, 1995). The purpose is to test whether a common underlying factor might explain the relationship between the two variables. With single observed variables for x and y at each time point, a latent variable is specified at each wave. No autoregressive or cross-lagged paths are specified. Note that this model is really just a longitudinal factor model with two indicators for each factor, which will typically be underidentified. The model can be identified assuming longitudinal equality constraints and three or more waves of data, however. If both measures assess the same underlying construct, then the cross-lagged model would be erroneous. The latent variable for each wave is conceptualized as a representation of all omitted covariates at each time point. The general structure of the model can be extended to a model with latent variables instead of observed variables, which would help eliminate some restrictions needed for identification of the observed variable model. With multiple indicators, the common factor hypothesis also could be addressed with confirmatory factor analysis of cross-sectional models as a separate step prior the longitudinal analysis.

A hybrid of the common factor model and the static unobserved model is possible by adding the lagged effects of a common factor at each wave to the synchronous common covariate model. General specifications are shown in Figure 5.12. This model examines

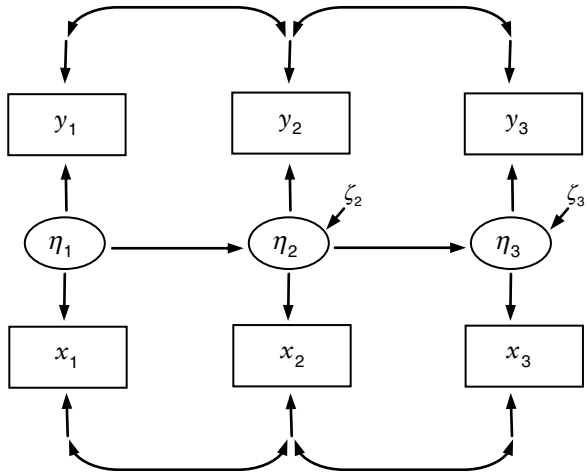

Figure 5.11 Synchronous Common Factor Model.

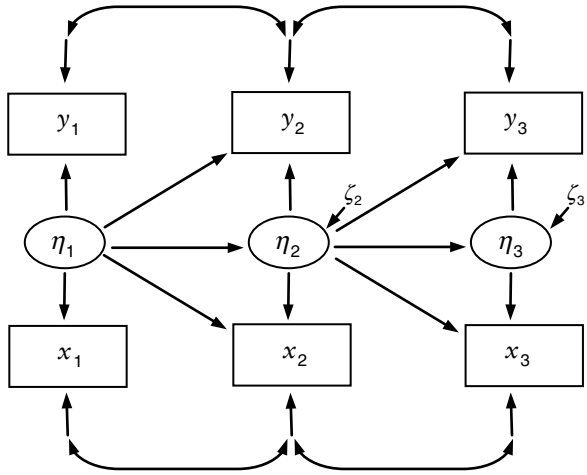

Figure 5.12 Unmeasured Common and Synchronous Factor Hybrid Model.

the possibility that omitted variables might have an effect on measures at each subsequent occasion.

Additional Variables

The basic cross-lagged panel model for two variables, either with single observed variables at each occasion or latent variables at each occasion, can be expanded to examine the relations among three or more variables. An illustration is shown in Figure 5.13 for three variables. All of the basic concepts for the two-variable model apply, including imposing equality constraints and correlated measurement residuals over time. The three-variable model introduces additional possible cross-lagged paths, with η_2 and η_4 both regressed on η_5 and η_6 regressed on both η_1 and η_3. Such a model provides a natural extension of the

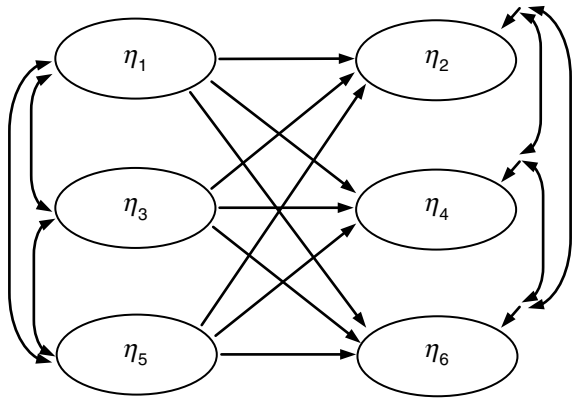

Figure 5.13 Three-Variable Cross-Lagged Panel Model.

model with a static or time-varying covariate to a model that investigates the cross-lagged effects among the three variables. The inclusion of the additional variable as a predictor at each wave means that the two-variable cross-lagged model among η_1, η_2, η_3, and η_4 takes into account (or removes) the cross-lagged effects of η_5. The model in Figure 5.13 clearly shows that additional variables add considerably to the complexity of the model, and theory will be needed to guide choices about whether to include covariates as static predictors, time-varying covariates, or variables with a full set of cross-lagged effects.

Mediation

An additional benefit to including a third variable in some of the models discussed thus far is the ability to investigate mediational pathways. These hypotheses concern indirect effects, that some variable *x* causes *y* through the effects of a mediator, *m*. In practice, mediational effects are often investigated using cross-sectional data, but, because the hypothesized model is one of a chain of causal effects, the analysis would benefit if elements of longitudinal models could be included. Modern methods of mediation analysis generally involve tests of the indirect effect of *x* on *y* and have largely replaced, or at least been used to supplement, older stage-based sequences of analyses (Baron & Kenny, 1987). The indirect effect can be conceptually described as the change in *y* for each unit change in *x* as mediated by variable *m*. The indirect effect can be mathematically described as the product of the two direct paths (from *x* to *m* and from *m* to *y*) or as the difference between the direct effect of *x* on *y* and the effect of *x* on *y* once *m* has been controlled in the model. The estimation of standard errors and significance tests for indirect effect coefficients requires special computations, which are not always available in SEM software programs. For mediation with observed variables, some special considerations are required for indirect tests with binary and ordinal variables (MacKinnon, Lockwood, Brown, Wang, & Hoffman, 2007). I assume the reader has had a general introduction to the concepts of mediation and indirect effects in path models, so I refer to other sources if further review is needed (Iacobucci, 2008; MacKinnon, 2008; MacKinnon, Fairchild, & Fritz, 2007).

Even though the mediational model involves three stages in the causal chain, there is a clear advantage over cross-sectional models even when there are only two waves of data available. With two waves of data, lagged effects for each stage of the mediational relationship cannot all be modeled across different waves. Ideally, we would prefer to have

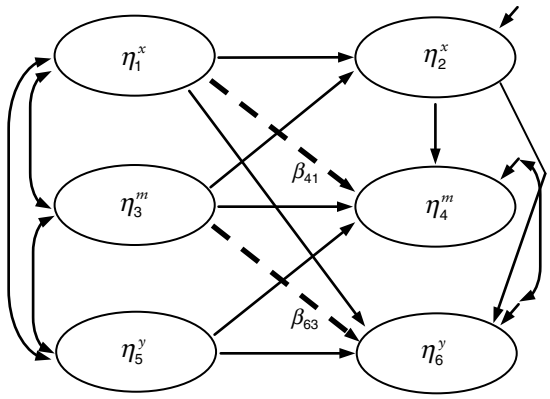

Figure 5.14 Longitudinal Mediation Model with Two Waves.

different temporal lags for the *x–m* path and the *m–y* path, which a three-wave design would provide. If the predictor, mediator, and outcome are measured at both time points, however, a longitudinal lagged effect can be modeled for each stage using only two waves of data (Cole & Maxwell, 2003; see also MacKinnon, 2008). Figure 5.14 illustrates one approach to the problem. In the figure, the latent variables η_1^x, η_3^m, and η_5^y represent the predictor, mediator, and outcome measured at time $t-1$ and η_2^x, η_4^m, and η_6^y represent the predictor, mediator, and outcome measured at time t. Notice that the model is simply a variant of a three-variable cross-lagged panel model. The product of two cross-lagged paths, between the predictor and the mediator, β_{41}, and between the mediator and the outcome, β_{63}, form the indirect effect. The path between the mediator and the outcome, β_{63}, must take into account the lagged effect of the predictor for the correct computation of the indirect effect, so the path from η_2^x to η_6^y must be included.

The model presented here differs slightly from the one proposed by Cole and Maxwell in that directional paths between the predictor and mediator measured at $t-1$ and the predictor measured at t are used instead of non-directional covariances. The modified model produces equivalent parameter estimates for the indirect effect while also generating additional cross-lagged effects on the predictor variable to examine bidirectional relationships. Not all possible paths are included in the model and lack of fit may indicate that direct effect of the outcome on the mediator variable, β_{61}, may be significant if included. The model is illustrated with latent variables, but could be estimated with observed variables instead. The calculation of the indirect effect and optimal significance test must be computed and tested manually (Preacher & Hayes, 2008), because the two constituent paths are not linked in the model. Alternatively, a model with synchronous directional paths for either the *x–m* or the *m–y* could be chosen if a more immediate effect is expected for one of the direct effects. This model would represent only a partial longitudinal mediation model and theory would be needed to guide the choice of longitudinal and cross-sectional paths. Software programs with features that include indirect coefficients estimates and significance tests could be used for the indirect effect in this case. An important advantage of estimating mediation with latent variables is that measurement error can be taken into account, allowing for more accurate estimates of mediational paths (Cole & Preacher, 2014).

For three waves of data and three variables, specification of the model for the indirect effects is more clear-cut. Figure 5.15 illustrates the model. The path from η_2^x to η_9^y

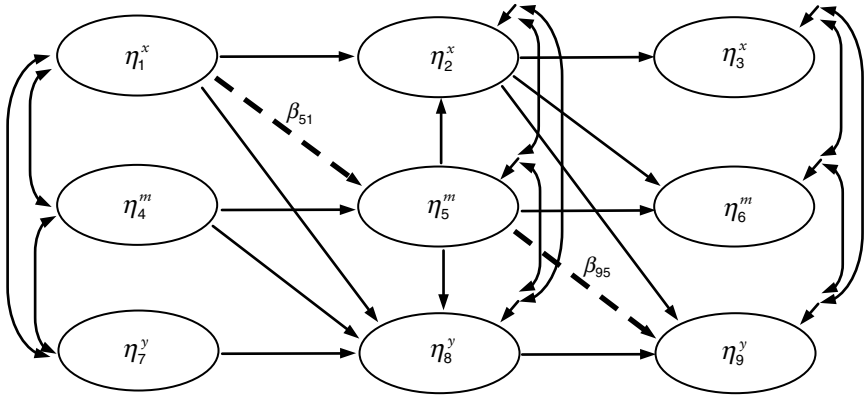

Figure 5.15 Longitudinal Mediation Model with Three Waves.

must be included so that the effect of the mediator on the outcome takes the effect of the predictor into account. The product of β_{51} and β_{95} can then be multiplied for the indirect effect over the three waves.

Example 5.7: Mediation Models

We return to the hypothesis raised in response to the results from Example 5.3 that depressed individuals may receive less support because they are less affable. Another way to phrase this hypothesis is to ask whether the effect of depression on companionship support is mediated by positive affect, or depression → positive affect → companionship support. The mediational model was first tested as if there were only two waves of data available for these variables, following the model specifications shown in Figure 5.14 with depression as the predictor, η_t^x, positive affect as mediator, η_t^m, and companionship as the outcome, η_t^y. From this model, the estimates for positive affect regressed on depression, $\beta_{41} = .121$, $\beta_{41}^* = .134$, $p < .05$, and the estimates for companionship regressed on positive affect, $\beta_{63} = .454$, $\beta_{63}^* = .242$, $p < .001$, were both significant. The indirect coefficient can be computed from these values, $\beta_{\text{indirect}} = .121 \times .454 = .055$, $\beta^*_{\text{indirect}} = .134 \times .242 = .032$. The unstandardized coefficients and their standard errors were used to compute the Sobel test of significance for the indirect effect (Preacher & Hayes, 2008).[6] The test of the indirect effect was marginally significant, $z = 1.918$, $p = .055$. Note that the sign of the path from depression to positive affect was opposite the predicted direction, suggesting that those with higher depression levels initially were higher on positive affect.

The same mediational hypothesis was tested with three waves of data (following Figure 5.15). This model employed effects coding identification, equal loadings over time, and correlated measurement residuals as in previous models. The cross-lagged and autoregressive paths were not constrained to be equal, given earlier results that suggested differences in some parameters across waves. The two direct effects of interest (β_{51} and β_{95} in Figure 5.15) for depression at Time 1 predicting positive affect at Time 2, and for positive affect at Time 2 predicting companionship at Time 3, were both significant, $\beta_{51} = .854$, $\beta_{51}^* = .817$, $p < .01$ and $\beta_{95} = -.463$, $\beta_{95}^* = -.236$, $p < .05$, respectively. Neither of these coefficients were in the direction expected, and the indirect coefficient was not significant, $\beta_{\text{indirect}} = -.396$, $\beta^*_{\text{indirect}} = -.193$, ns.

Continuous Time Models

The autoregression and cross-lagged panel models discussed thus far have assumed equally spaced discrete time intervals (e.g., 6 months, 2 years). The available data may not conform to equally spaced intervals, and, more generally, it can be argued that more precise estimates of change are possible if we consider effects over all possible intervals. Use of an incorrect temporal lag or assuming equally spaced intervals when they are unequal can lead to biased estimates (Gollob & Reichardt, 1987; Lorenz, Conger, Simons, & Whitbeck, 1995; Voelkle & Oud, 2013). With a more complex mathematical representation of time, longitudinal models can be generalized to estimate change as a function of infinitesimally small increments, addressing potential biases resulting from discrete temporal intervals.

Consider a unidirectional cross-lagged model of continuous changes in a variable y_t predicted by a prior measurement $y_{t-\Delta t}$ and a time invariant covariate, $x_{t-\Delta t}$, also measured at some prior time point. Here, continuous increments in time of any unspecified value greater than 0, are symbolized by Δt. For a study involving discrete time intervals, with $\Delta t = 1$, the prior time point, $t - \Delta t$, reduces to just $t - 1$. A model with a cross-lagged effect in only one direction can be stated in terms of derivatives.

$$\frac{dy_t}{d\Delta t} = \alpha_1 + \beta_{23} x_{t-\Delta t} + \beta_{21} y_{t-\Delta t} \tag{5.3}$$

The left side of the equation is read as the derivative of y with respect to Δt, which represents change in y over any time increment greater than 0. A brief primer on derivatives can be found in Appendix B. The derivative equation above can then be restated in terms of the exponential function of three continuous change coefficients, c_0, c_1, c_2 (Coleman, 1968; Finkel, 1995).

$$y_t = \frac{c_0}{c_2}\left(e^{c_2 \Delta t} - 1\right) + \frac{c_1}{c_2}\left(e^{c_2 \Delta t} - 1\right) x_{t-\Delta t} + e^{c_2 \Delta t} y_{t-\Delta t} \tag{5.4}$$

Using the complementary natural log function, ln, each of these continuous change coefficients can be derived from the intercept, the cross-lagged coefficient, and the autoregressive coefficient.

$$\begin{aligned} c_0 &= \beta_0 \frac{\ln \beta_{21}}{\Delta t\left(\beta_{21} - 1\right)} \\ c_1 &= \beta_1 \frac{\ln \beta_{21}}{\Delta t\left(\beta_{21} - 1\right)} \\ c_2 &= \frac{\ln \beta_{21}}{\Delta t} \end{aligned} \tag{5.5}$$

The coefficients represent a continuous change in y_t per unit of time, Δt, and c_0 represents the change in y for each unit increment of time if x and y are zero. The increment in time can be any desired interval, and serves to scale the coefficients in terms of the time scale desired (number of months, number of years, etc.). In other words, any time interval whether shorter or longer than the intervals observed in the study can be chosen as a basis for the interpretation of the unit of change. Derived from the cross-lagged effect, c_1 represents the change in y per unit of time for each increment in $x_{t-\Delta t}$. Derived from the autoregression effect, c_2 represents the change in y for each unit increase in the initial value of $y_{t-\Delta t}$. In simpler terms, c_2 represents absolute values of change regressed on y and reduces to the difference score, $y_2 - y_1$ regressed on y_1, in the two-wave design. It is apparent then

that the model in Equation (5.3) parallels the cross-lagged regression model in which the difference score is predicted by a covariate x while also controlling for initial levels of y.

As in the previous discussion of difference scores with two waves, whenever there is regression toward the mean (entropy) and the autoregressive coefficient is positive, c_2 will be negative. The negative coefficient in this context indicates the eventual arrival at equilibrium in the system. Sometimes the association between initial values and change scores is described as *negative feedback*. Negative feedback indicated by a negative c_2 coefficient can be construed as the speed at which the system will reach equilibrium. The larger the magnitude of the negative coefficient, the faster the system is expected to reach equilibrium. The point at which the system arrives at equilibrium for a case with a particular starting value on $x_{i,t-\Delta t}$ can be estimated from the continuous change coefficients by $(-c_0 + c_1 x_{i,t-\Delta t})/c_2$. The continuous change coefficients and equilibrium estimate can be computed manually from the cross-lagged panel coefficients in the SEM program output or they may be obtained by setting complex equality constraints in programs that have this feature.

The above derivations are sometimes described as an indirect approach to continuous time modeling (Hamerle, Nagl, & Singer, 1991). The estimates from the indirect method should be considered approximations and can only be calculated from these formulas if a constant time increment is assumed. More recently, a method for more direct estimation of continuous autoregressive and cross-lagged effects has been proposed by using complex equality constraints (Oud, 2012; Voelkle & Oud, 2013; Voelkle, Oud, Davidov, & Schmidt, 2012). The direct method involves a transformation of the autoregression and cross-lagged coefficients much as in Equation (5.4), but it also includes the correct constraints on the factor variances and structural disturbances in the stochastic differential equation that includes the variance component. The specification of the direct continuous time model is not possible in most SEM programs,[7] because the constraints require identification of the time interval from the data as well as integration matrix algebra manipulations (see Voelkle et al., 2012, and Voelkle & Oud, 2013, for details).

Example 5.8: Computation of Continuous Time Coefficients

To illustrate the computation of continuous time coefficients, results from the cross-lagged panel model from Example 5.2 were used. The model was retested with mean structures to obtained intercept values for the cross-lagged equations. Because the interval between Time 1 and Time 2 was six months in the discrete panel study, continuous time coefficients were computed based on six one-month increments. The path estimates from the model were inserted into each of the formulas in Equation (5.5). Using the unstandardized values from the effect of Time 1 unwanted advice on Time 2 positive affect, the intercept was $\alpha = 1.132$, the cross-lagged effect was $\beta = -0.015$, and the autoregressive effect was $\beta = 0.601$.

$$c_0 = \beta_0 (\ln \beta_{21}) / \left[\Delta t (\beta_{21} - 1)\right] = 1.132 \cdot \ln(.601) / \left[6(.601 - 1)\right] = .241$$

$$c_1 = \beta_1 (\ln \beta_{21}) / \left[\Delta t (\beta_{21} - 1)\right] = -.015 \cdot \ln(.601) / \left[6 \cdot (.601 - 1)\right] = -.003$$

$$c_2 = (\ln \beta_{21}) / \Delta t = \ln(.601) / 6 = -.085$$

The c_0 coefficient estimates that there will be a .241 increase per month in positive affect when positive affect and unwanted advice are 0. The c_1 coefficient suggests that there will be a .003 continuous decline in positive affect per unit increase in unwanted advice in one month. The c_2 coefficient suggests that there will be a –.085 continuous change in positive affect for each unit increase in positive affect one month earlier. The

estimate of equilibrium for the cross-lagged effect for an individual with a high initial value on unwanted advice, say 3, is

$$\left(-c_0 + c_1 x_{i,t-\Delta t}\right) / c_2 = \left[-.241 + \left(-.003 \cdot 3\right)\right] / \left(-.085\right) = 2.875$$

This result suggests that, at the point of equilibrium, an individual with a very high initial level of unwanted advice will have a score of 2.875 on positive affect, a slightly higher than average value on the affect scale.

The continuous coefficients can also be computed for the effect of positive affect on unwanted advice. The unstandardized values for the effect of Time 1 positive affect on Time 2 unwanted advice were intercept $\alpha = 1.67$, cross-lagged coefficient $\beta = -.227$, and autoregressive coefficient $\beta = .601$. And the continuous time coefficients are therefore

$$c_0 = \beta_0 \left(\ln \beta_{21}\right) / \left[\Delta t \left(\beta_{21} - 1\right)\right] = 1.898 \cdot \ln(.391) / \left[6(.391 - 1)\right] = .488$$

$$c_1 = \beta_1 \left(\ln \beta_{21}\right) / \left[\Delta t \left(\beta_{21} - 1\right)\right] = -.227 \cdot \ln(.391) / \left[6 \cdot (.391 - 1)\right] = -.058$$

$$c_2 = \left(\ln \beta_{21}\right) / \Delta t = \ln(.391) / 6 = -.157$$

An estimate of the value of unwanted advice given a very low initial value of positive affect (1.00) is a very high value on unwanted advice, 3.489, which is well above the average of the scale:

$$\left(-c_0 + c_1 x_{i,t-\Delta t}\right) / c_2 = \left[-.488 + \left(-.058 \cdot 1\right)\right] / \left(-.157\right) = 3.489$$

Multigroup Models

Any of the models described previously can be extended to involve multigroup comparisons. Such models allow for comparisons across treatment groups, gender, cultural groups, or age cohorts, creating an abundance of possible models that may be of interest to researchers. Multigroup comparisons should include tests of measurement invariance across groups as an initial analysis to ensure that group differences in measurement properties are not confused with group differences. Because the procedures for testing such models follow the general steps described in introductory SEM texts and the extensions of the models described in this chapter do not involve substantially new longitudinal concepts, I leave it to other authors to discuss the relevant issues.

Comments

The primary goal of cross-lagged panel analyses is to investigate questions about causal precedence. These analyses are generally conducted with passive observational designs with no particular design elements to address threats to interval validity. Experimental designs are preferable for investigating questions about causal precedence and isolation of the effects of the independent variable. But true experiments are not always feasible, ethical, or applicable to real world settings. Quasi-experimental designs often have advantages over passive observations longitudinal designs as well, and many of the analysis approaches described here could be used in conjunction with quasi-experimental design elements. We should be candid in acknowledging that for whatever the statistical rigor we can gain from cross-lagged panel models of passive observational data, cross-lagged panel models are imperfect tools for determining causality and causal directionality with certainty. Due to the availability of observations over time, cross-lagged analyses undoubtedly hold a number of advantages over any analyses derived from cross-sectional data, however. There is really no hope for

disentangling causal precedence with passive cross-sectional observations. Coupled with the benefits derived from removing measurement error with latent variables and estimating correlated measurement residuals over time, cross-lagged panel models offer considerable strength for investigating longitudinal hypotheses about causal directionality.

Without losing sight of the strengths, however, there are several points that should be kept in view. We can only assume the effects we estimate with discrete time longitudinal designs are unbiased to the extent that the time lag is correct. If the time length is different from what is measured, the relationship estimated may be larger or smaller than the true relationship (or relationships expected for a given time lag). The incorrect lag can affect the autoregression or the cross-lagged estimates. I have also confined the discussion to linear relationships. One obvious approach to extending these models is to test nonlinear relationships using conventional transformations of independent variables. Other methods, as in the growth curve model, are adaptable to the investigation of questions about the nonlinear changes (see Chapter 8).

Researchers applying cross-lagged panel models frequently forget the potential role of covariates. Initial states, autoregressive effects, or cross-lagged effects may be confounded with other variables omitted from the model. Even though many theoretically relevant external variables may be available, they are often left out of cross-lagged panel models, perhaps because their inclusion requires added complexity. This should not diminish their importance, however. Dwyer (1983, p. 327) summarizes this role of covariates nicely in stating: "In the end, the statement that 'X shapes Y to a certain extent' must be tempered to 'X or some factor closely related with X shapes Y to a certain extent.' "

Recommended Readings

There are a number of good sources on cross-lagged panel models and related topics. The best general resource is Finkel (1995), who provides an accessible overview with many political science and sociological examples. A classic treatment of analysis of panel data is by Kessler and Greenberg (1969). Although there is little discussion of many of the issues related to latent variables, this text covers many relevant topics from fundamental concepts of change to the effects of measurement error. An in-depth and balanced discussion of the interpretation of lagged effects and difference scores can also be found in Dwyer (1983, Chapter 11), with a superb discussion of fundamental concepts of change over many waves. A didactic illustration of a cross-lagged panel analysis with SEM is given by Burkholder and Harlow (2009). Two writings by Jöreskog and colleagues provide reasonably accessible introductions to simplex models with illustrations (Jöreskog & Sörbom, 1989, Chapter 6; Werts, Linn, & Jöreskog, 1977). Despite the causal nature of mediational hypotheses, few authors have addressed issues involved in mediation testing with longitudinal data. Three exceptions provide detailed discussions of mediation for longitudinal data (Cole & Maxwell, 1993; MacKinnon, 2008; Roth & MacKinnon, 2012). The least technical introductions to Oud's recent developments of continuous time analyses with SEM cross-lagged panel models can be found in Oud and Delsing (2010) and Voelkle, Oud, Davidov, and Schmidt (2012).

Notes

1 The cross-lagged panel correlation analysis was subsequently critiqued in part for examination of simple bivariate correlations without adequately taking into account earlier values of dependent variable and for several other reasons (e.g., Dwyer, 1983; Rogosa, 1980).

2 It is also possible to count the df in this special case using the alternative formula, $df = [J(J-1)/2] - q$, in which the variances are not counted as free parameters.

3 This is generally the case, but I overstate the point somewhat. An examination of the formula for the standardized partial regression coefficient will reveal that even if the correlation between the covariate and the outcome is zero, the correlation between the covariate and the predictor can affect the magnitude of the partial regression coefficient.
4 Weighted root mean square residual values less than 1.0 are considered indicative of good fit according to Yu and Muthén (2002).
5 The general simplex or autoregressive model is sometimes referred to as the *Wiener model*, because it relates to the stochastic process that Wiener outlined to help describe the Brownian motion in particle physics (Wiener, 1923). The simplex model and Wiener process are also described as a *random walk* process, which will be discussed in greater detail in Chapter 11.
6 Although this test is not optimal for small samples sizes, it converges with other tests in larger sample sizes. There are other preferable tests (MacKinnon, 2008). When each direct effect forms a compound path, as in the three-wave case, several SEM software programs have special features that employ more optimal indirect tests.
7 At this time, only Mx and MECOSA have these computational capabilities.

References

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary, NC: SAS Institute Inc.

Arminger, G. (1987). Misspecificaiotn, asymptotic stability, and ordinal variables in the analysis of panel data. *Sociological Methods & Research, 15*, 336–348.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603–617.

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.

Berry, W. D. (1984) *Nonrecursive causal models*. Beverly Hills, CA: Sage.

Bollen, K. A., & Brand, J. E. (2010). A general panel model with random and fixed effects: A structural equations approach. *Social Forces, 89*, 1–34.

Burkholder, G. J., & Harlow, L. (2009). An illustration of a longitudinal cross-lagged design for larger structural equation models. *Structural Equation Modeling, 10*, 465–486.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research on teaching*. Dallas, TX: Houghton Mifflin.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Erlbaum.

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*, 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*, 300–315.

Coleman, J. S. (1968). The mathematical study of change. In H. M. Blalock, Jr. & A. B. Blalock, *Methodology in social research* (pp. 428–478). New York: McGraw-Hill.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, 105, 317–327.

Deshon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3*, 412–423.

Duncan, O. D. (1969). Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, 72, 177–182.

Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences*. Oxford, UK: Oxford University Press.

Finkel, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage.

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development* [Special section on structural equation modeling], *58*, 80–92.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*, 424–438.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. R. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (pp. 258–348). Glencoe, IL: Free Press.

Hamerle, A., Nagl, W., & Singer, H. (1991). Problems with the estimation of stochastic differential equations using structural equation models. *Journal of Mathematical Sociology, 16*, 201–220.

Heise, D. R. (1975) *Causal analysis*. New York: Wiley.

Iacobucci, D. (2008) *Mediation analysis*. Newbury Park, NJ: Sage.

Jones, M. B. (1959). *Simplex theory.* Monograph series no. 3. Pensacola, FL: U.S. Naval School of Aviation Medicine.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23*, 121–145.

Jöreskog, K. G., & Sörbom, D. (1979). Basic ideas of factor and component analysis In J. Magidson (Ed.), *Advances in factor analysis and structural equation models*. (pp. 5–20). Cambridge, MA: Abt Books.

Jöreskog, K. G., & Sörbom, D. (1989). *Lisrel 7: A guide to the program and applications*. Chicago, IL: SPSS Inc.

Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: Models of quantitative change*. New York: Academic Press.

Kim, J., & Feree, G. (1981). Standardization in causal analysis. *Sociological Methods & Research, 10*, 187–210.

Lazarsfeld, P. F. (1940). "Panel" studies. *Public Opinion Quarterly, 4*, 122–128.

Lazarsfeld, P. F., & Fiske, M. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly, 2*, 596–612.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika, 73*, 13–22.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13*, 59–72.

Lorenz, F. O., Conger, R. D., Simons, R. L., & Whitbeck, L. B. (1995). The effects of unequal variances and covariances in simultaneous inference: The case of hostility and marital quality. *Journal of Marriage and the Family*, *574*, 1049–1064.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614.

Mandys, F., Dolan, C. V., & Molenaar, P. C. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics*, *19*, 201–215.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement, 30*, 157–183.

McDonald, R. P. (1996). Path analysis with composite variables. *Multivariate Behavioral Research, 31*, 239–270.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman & Hall/CRC.

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Newsom, J.T., Nishishiba, M., Morgan, D. L., & Rook, K. S. (2003). The relative importance of three domains of positive and negative social exchanges: A longitudinal model with comparable measures. *Psychology and Aging, 18*, 746–754.

Oud, J. H. L. (2012). Continuous time modeling of the cross-lagged panel design. *Kwantitatieve Methoden, 69*, 1–27.

Oud, J. H. L., & Delsing, M. J. M. H. (2010). Continuous time modeling of panel data by means of SEM. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 201–244). New York: Springer.

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*, 879–891.

Rigdon, E. E. (1994). Calculating degrees of freedom for a structural equation model. *Structural Equation Modeling*, *1*, 274–278.

Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*, 245–258.

Rogosa, D. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–210). New York: Springer.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228.

Roth, D. L., & MacKinnon, D. P. (2012). Mediation analysis with longitudinal data. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 181–216). New York: Routledge.

Rovine, M. J., & Molenaar, P. C. M. (2005). Relating factor models for longitudinal data to quasi-simplex and NARMA models. *Multivariate Behavioral Research, 40*, 83–114.

Sivo, S. A., & Willson, V. L. (2000). Modeling causal error structures in longitudinal panel data: A Monte Carlo study. *Structural Equation Modeling*, *7*, 174–205.

Teachman, J., Duncan, G. J., Yeung, W. J., & Levy, D. (2001). Covariance structure models for mixed and random effects. *Sociological Methods & Research, 30*, 271–288.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83–91.

Voelkle, M. C., & Oud, J. H. L. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *British Journal of Mathematical and Statistical Psychology*, *66*, 103–126.

Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods, 17*, 176–192.

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1977). A simplex model for analyzing academic growth. *Educational and Psychological Measurement 37*, 745–756.

Wiener, N. (1923). Differentiable space. *Journal of Mathematics and Physics, 2*, 131–174.

Yu, C. Y., & Muthén, B. (2002, April). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the *Annual Meeting of the American Educational Research Association,* New Orleans, LA.

Keywords
cross-lagged panel models, autoregression, change, Granger causality, mediation, simplex models

6 Latent State-Trait Models

A set of related models, which I will refer to collectively as latent state-trait models, are designed to model stable aspects of a construct over time. In these models, the stable variance, referred to as trait variance, is separated from systematic variance that changes over time, referred to as state variance, and non-systematic variance, referred to as error variance. Although to date latent state-trait models have largely been of interest to personality psychologists, the term "trait" is simply a convenient label in this context for consistency in a construct over time and could be fruitfully applied to many other domains. The concept of stability that this class of models is concerned with is autocorrelation rather than level stability, so that consistency is defined as the relationship of a single case over time relative to other individuals in the data set. That means a variable might be considered "stable" even if there are increases or decreases in the mean or individual values over time. With similar generality, the term "state" does not need to apply to a psychological state, but can represent any systematic variance that is not stable over time. Trends, cycles, or less regular changes over time that are systematic (i.e., not random or stochastic) are all grouped together as state or occasion-specific variability.

Many disciplines make useful distinctions between consistent, stable aspects of a variable and more temporally changing aspects. Within personality psychology, for example, the hypothesized trait of neuroticism tends to show high levels of stability over years (Costa & McCrae, 1988), and may be distinguishable from more temporary fluctuations in moods or emotions that occur on a daily, weekly, or monthly basis.[1] Within economics, fundamental trends in the economy are distinguished from seasonal changes in employment (Kropf & Hudson, 2012). In political science, ideological undercurrents are distinguished from popular political trends (Adams, Clark, Ezrow, & Glasgow, 2004). In sociology, distinctions are made between institutional stability and external stresses on social systems (Mahoney & Thelen, 2010).

Steyer and colleagues (Steyer, 1987; Steyer & Schmitt, 1990; Steyer, Majcen, Schwenkmezger, & Buchner, 1989; Schmitt & Steyer, 1993) originally proposed a confirmatory factor model formulation of the *latent state-trait model* based on the idea of partitioning systematic variance from classical test theory (Tack, 1980). A major alternative was later proposed by Kenny and Zautra (1995, 2001) using a combination of confirmatory factor analysis and autoregressive models. Several other variants of either model have been proposed since, incorporating elements of multitrait-method matrix models, second-order factor models, and time series analysis (e.g., Cole, Martin, & Steiger, 2005; Geiser & Lockhart, 2012; Hamaker, Nesselroade, & Molenaar, 2007). Although some of these models could use just two waves of data in theory, they are generally applied with three or more waves, with some requiring more. This chapter provides a general overview and introduction to this class of structural equation models that enables investigation of a range of hypotheses about stable and occasion-specific aspects of some phenomenon.

Latent State-Trait Model

Specification

The concept of the latent state-trait model proposed by Steyer and colleagues (Schmitt & Steyer, 1993; Steyer, 1987) further decomposes the variance from classical test theory, $\mathrm{Var}(y_{jt}) = \mathrm{Var}(T_{jt}) + \mathrm{Var}(e_{jt})$, into two components of true score variance. The model, depicted in Figure 6.1, can be expressed with the following equation:

$$\mathrm{Var}(y_{jt}) = \beta_t^2 \lambda_{jt}^2 \mathrm{Var}(\eta) + \lambda_{jt}^2 \mathrm{Var}(\zeta_t) + \mathrm{Var}(\varepsilon_{jt}) \tag{6.1}$$

The first term on the right of the equation represents trait variance, the second term represents state variance, and the last term represents error. The observed variance of a variable j measured at time t is a function of stable variance of trait factor η, occasion-specific residual variance of factor η_t at each time t (i.e., the disturbance variance $\mathrm{Var}(\zeta_t) = \psi_{tt}$), and unaccounted-for variance in the observed variable, ε_{jt} (i.e., $\mathrm{Var}(\varepsilon_{jt}) = \theta_{jj}$). The occasion-specific factors η_t are latent variables with multiple indicators having loadings λ_{jt}, and the trait factor is a second-order latent variable η (see Loehlin, 2004, for a review of second-order factor models). As usual, variance components imply variation across individual cases in the data set, but the index i is omitted from each of the relevant terms to minimize subscripts. Because this model is a second-order factor model, a minimum of three time points would usually be recommended to avoid empirical underidentification problems unless additional constraints are used. The loadings (or regression paths) for the second-order factor, β_t, are generally set to 1, giving equal contribution to the factor from each occasion-specific factor.

Other variations for identifying the factor are possible, such as setting loadings equal (and fixing the factor variance to 1) or allowing the second-order factor loadings to be freely estimated at each time point. The equal weighting of each time-specific factor is congruent with the "trait" notion of a constant effect from the underlying stable construct. The model can be specified with two time points if factor loadings are set to specific values or constrained to be equal, but we will assume three or more time points in practice. Each occasion-specific factor, η_t, is defined by the same indicators at each time point (and generally should include equality constraints on loadings), and represents the construct at each occasion. The residual variance for the occasion-specific latent variable, $\mathrm{Var}(\zeta_t)$, is conditional on the trait factor term, ψ_{tt}, and represents "state" variance that remains after the stable variance of the trait factor has been removed. These disturbance terms are allowed to vary over time, reflecting any variance from unknown factors that cause systematic variation specific to an occasion. Such factors might involve any unstable factors that can potentially be measured, including moods, seasonal employment, or weather. The unstable causal factors do not need to be measured in the data set or included in the model but could be incorporated as time-varying covariates to account for the state variance.

Variance Decomposition

Building upon the parallels to classical test theory, the ratio of true score variance to total observed variance that quantifies reliability, $\rho = \mathrm{Var}(T) / \mathrm{Var}(y)$, can be applied to each of the systematic components as well. Steyer and colleagues refer to the proportion of variance due to stable or trait variance as the *common consistency coefficient*:

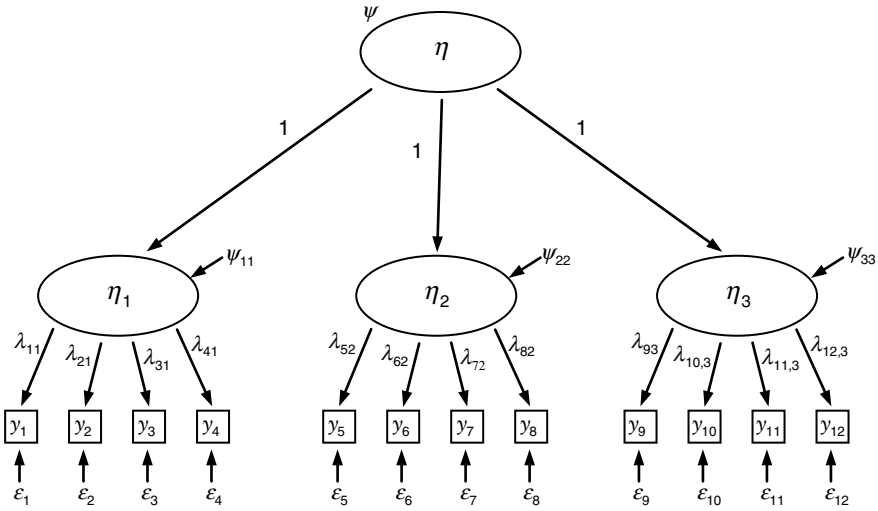

Figure 6.1 Latent State-Trait Model.

$$\begin{aligned}\text{common consistency coefficient} &= \frac{\beta_t^2 \lambda_{jt}^2 \operatorname{Var}(\eta)}{\operatorname{Var}(y_{jt})} \\ &= \frac{\beta_t^2 \lambda_{jt}^2 \operatorname{Var}(\eta)}{\beta_t^2 \lambda_{jt}^2 \operatorname{Var}(\eta) + \lambda_{jt}^2 \operatorname{Var}(\zeta_t) + \operatorname{Var}(\varepsilon_{jt})}\end{aligned} \tag{6.2}$$

And the proportion of state variances is gauged by the *occasion specificity coefficient*:

$$\begin{aligned}\text{occasion specificity coefficient} &= \frac{\lambda_{jt}^2 \operatorname{Var}(\zeta_t)}{\operatorname{Var}(y_{jt})} \\ &= \frac{\lambda_{jt}^2 \operatorname{Var}(\zeta_t)}{\beta_t^2 \lambda_{jt}^2 \operatorname{Var}(\eta) + \lambda_{jt}^2 \operatorname{Var}(\zeta_t) + \operatorname{Var}(\varepsilon_{jt})}\end{aligned} \tag{6.3}$$

In terms of psychological interpretation, the occasion-specific coefficient represents the proportion of total variance due to situational factors as well as the person × situation interaction. The second-order factor loadings are usually set equal to 1, so β_t^2 would be dropped from each of the equations. Although not shown in Figure 6.1, measurement invariance constraints, such as setting factor loadings for repeated indicators equal over time, would typically be incorporated.

The common consistency and the occasion specificity coefficients represent partitioning of true score variance and with error variance they sum to the total variance of y_{jt}. As defined, there is a common consistency and occasion specificity coefficient value computed for each observed variable, where the coefficients provide estimates of the portion of the observed variance of a particular variable that is due to stable and unstable factors. Researchers may be interested in reporting a single value for the average estimate of the variance due to trait and state variance. This could be accomplished in several ways, including computing the average of the variable specific coefficients, retesting the model with equality constraints on all state loadings, state disturbances, and measurement

residual variances (e.g., Schmitt & Steyer, 1993), or computing the coefficients based on an average of the state loadings, state-disturbances, and measurement residual variances. Use of the average of the variable specific coefficients is the most straightforward approach, but the three methods should produce similar values under most common conditions.

Method Variance

The basic latent state-trait model can be expanded to include method factors (depicted in Figure 6.2). This approach applies if there are multiple factors per occasion or multiple methods of measurement per occasion. An example might be an inventory with several subfactors, as illustrated by Schmitt and Steyer (1993) using two parallel measures of social desirability. In this expanded model, the true variance can be further decomposed into trait, state, and method variance.

$$\mathrm{Var}\left(y_{jt}\right) = \beta_t^2 \lambda_{jt}^2 \,\mathrm{Var}\left(\eta\right) + \lambda_{jt}^2 \,\mathrm{Var}\left(\zeta_t\right) + \lambda_{jk}^2 \,\mathrm{Var}\left(\eta_k\right) + \mathrm{Var}\left(\varepsilon_{jt}\right) \tag{6.4}$$

Equation (6.4) is the same as Equation (6.1), except that the third term, $\lambda_{jk}^2 \,\mathrm{Var}\left(\eta_k\right)$, representing method variance for each of the K method factors, has been added. Each indicator measured using a particular method (e.g., belonging to a particular subscale) loads on one of the method factors. Method factors are assumed independent of the trait factor. A common approach also assumes the method factors to be independent of one another to partition method variance into separate components due to each of the methods. As with the trait variance and the occasion-specific variance, the method variance can be similarly quantified by a ratio of the method variance to the total variance.

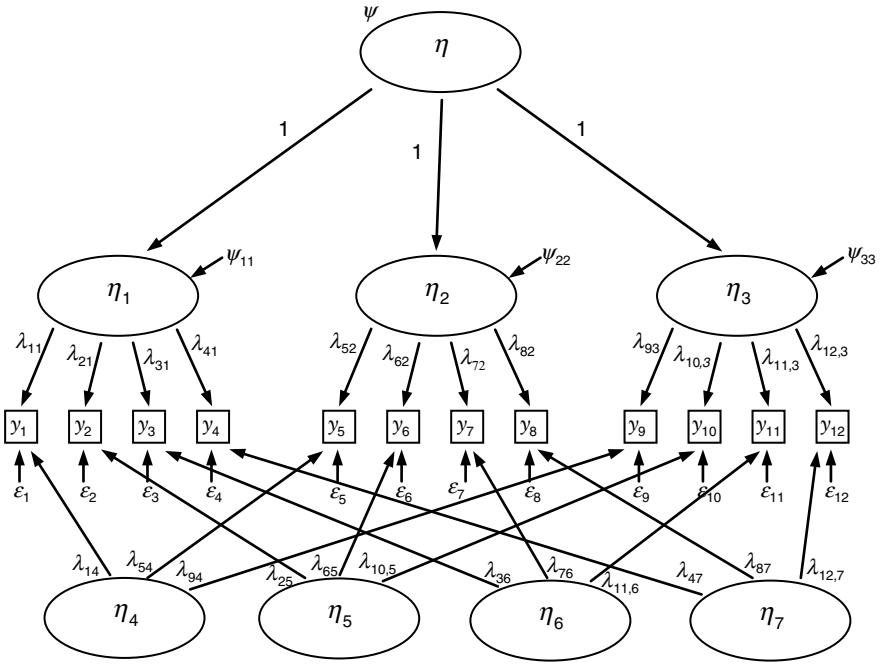

Figure 6.2 Latent State-Trait Model with Method Factors.

$$\text{method specificity coefficient} = \frac{\lambda_{jk}^2 \operatorname{Var}(\eta_k)}{\operatorname{Var}(y_{jt})} = \frac{\lambda_{jt}^2 Var(\zeta_t)}{\beta_t^2 \lambda_{jt}^2 \operatorname{Var}(\eta) + \lambda_{jt}^2 \operatorname{Var}(\zeta_t) + \lambda_{jk}^2 \operatorname{Var}(\eta_k) + \operatorname{Var}(\varepsilon_{jt})} \tag{6.5}$$

Marsh (1993) suggests that the latent state-trait model, whether including method factors or not, be compared to a first-order factor model with correlated factors. The ratio of the chi-square for the fit of the correlated-factors model to the fit of the latent state-trait model will always fit more poorly because it is more restrictive. The ratio of chi-squares, referred to as the *target coefficient*, is recommended to be close to 1.0, indicating a small decrement in fit by imposing the additional constraints, such as adding a higher-order factor. The criterion for practical difference in fit is arbitrary, of course, but Marsh and Hocevar (1985) concluded that a target coefficient of .95 computed from their data indicated a small practical difference. Smaller target coefficients with latent state-trait models will result from the assumption of equal trait loadings, independence of method factors (if applicable), or independence of correlated disturbances. Depending on the number of available occasions and other factors that may affect empirical identification (e.g., sample size), these assumptions may be tested and relaxed. The relative validity of the second-order factor model compared to the first-order factor model is also often evaluated with the Schmid–Leiman solution (Schmid & Leiman, 1957; see Wolff & Preising, 2005 for computational details), which may provide a more precise comparison.

The latent state-trait model can also be specified with correlated measurement residuals, or "correlated uniquenesses," for repeated measurements of each indicator (as in Figure 2.2). The pattern of correlations parallels the loadings on a particular method factor, so the correlated uniqueness model implies the same decomposition of effects as the method factor model if the method factors are orthogonal. Notice that in either case the pattern of relationships corresponds to autocorrelations of residuals used in many longitudinal models.

Another variant on the method factor specification is to allow method factors to be correlated. In many instances, correlated method factors make considerable theoretical sense, but they can be difficult to estimate (Tomás, Hontanges, & Oliver, 2000). A general solution to estimation difficulties with correlated method factors is to use one fewer method factors than the number of indicators for each occasion factor (or "$M - 1$" method factors; Eid, 2000), referred to as the *residual method factor approach*. Assuming one identification constraint for each factor, there will be one fewer possible method factors than loadings on each factor. Referent indicator or effect coding can be used to identify the method factors. Simulation work (Geiser & Lockhart, 2012) comparing these several approaches suggests that the correlated uniqueness approach and the orthogonal method factor approach may have biased parameter estimates even with a fairly low proportion of variance accounted for by the method factors (e.g., 5–10%). The residual method factor approach seems to be a viable approach to correlated method factors, performing well in terms of parameter bias and convergence when sample sizes are 300 or larger.

Identification of State Factors

The unstandardized estimates and thus the decomposition are dependent on the identification approach used to scale the factor variance. Either referent indicator or effects coding identification of the state factors are possible (see Chapter 1). Because the

state factors are endogenous, the factor variance or the single occasion identification approaches are not possible for scaling the state factors. Although the fit of the model will not be affected by identification method, some of the unstandardized parameter estimates will depend on the identification method, leading to potential differences in the estimates of the proportion of state and trait variance. The first-order loadings and state disturbance variances will differ for the two identification approaches, although the measurement residual variances will be the same. With referent identification, the scaling of each occasion-specific factor (state) variance is a function of the observed variance of the indicator used as the referent. With effects coding identification, the scaling of each variance is a weighted function of all of the loadings and observed variances of the indicators. The covariances among the occasion-specific factor variances will therefore differ in the two identification approaches, leading to differences in the trait factor variance. As a consequence of a combination of the different parameter estimates, the values for the consistency and specificity coefficients may differ. Effects coding identification has the advantage of variance estimates that are a weighted function of all of the observed indicators, thus avoiding any dilemmas about the selection of the most appropriate referent variable.

Example 6.1: Latent State-Trait Models

Several latent state-trait models (e.g., Steyer et al., 1989) were tested using the depression measure from the health and aging data set over six biennial waves. Syntax and data sets used in the examples are available at the website for the book. There were three subscale scores, negative affect, positive affect (reverse scored), and somatic symptoms, derived from the eight-item depression scale, used as indicators at each wave.[2] The first model only distinguished between stable trait variance and unstable state true score variance (i.e., no method factors). The occasion-specific (first-order) factors were identified using the effects coding approach, and longitudinal equality constraints were used for the loadings for repeated indicators over time. The single latent trait factor was constructed by setting the (second-order) factor loadings from each occasion equal to 1 for equal weighting. The disturbances for each first-order factor were allowed to differ across occasions. This model did not fit the data well, $\chi^2(143) = 7{,}109.912$, $p < .001$, CFI = .816, SRMR = .072. The common consistency coefficient and occasion-specificity coefficients using Equation (6.2) and Equation (6.3), estimated the proportion of variance due to the trait factor and the state factors, respectively. To estimate the coefficients, I used an average of the values for all loadings and variance estimates for convenience, and other methods of averaging may produce slightly different results. The average common consistency coefficient was .304, suggesting approximately 30% of the true score variance was stable over time. The average occasion specificity coefficient was .168, suggesting approximately 17% of the true score variance was unstable. The remaining 53% of the variance is considered error in this model.

These estimates are likely to be biased if specific variance for each indicator is correlated over time. Correlations among measurement residuals if unaccounted for in the model are likely to lead to over-estimates of stable variance, as some of the stability is simply due to common method variance. A retest of the latent state-trait above using the same specifications but adding correlated measurement residuals among the negative affect indicators, the positive affect indicators, and the somatic indicators, fit the data considerably better, $\chi^2(98) = 750.140$, $p < .001$, CFI = .983, SRMR = .055, suggesting that there was important common specific variance that was due to associations among measurement residuals.

Alternatively, the common specific variance can be modeled with method factors. This model (Figure 6.2) used effects coding for each of the method factors (negative affect, positive affect, somatic symptoms). When method factors were assumed to be orthogonal and uncorrelated with the trait factor, this model did not converge due to a negative variance estimate of the negative affect method factor variance. Inspection of the correlations among measurement residuals for this factor obtained with the previous model showed nonsignificant correlations and nearly all were considerably smaller in magnitude than the correlations among residuals related to the other two factors.

The residual method factor model with two correlated factors using the $M-1$ method converged and fit the data well, $\chi^2(130) = 1{,}005.360$, $p < .001$, CFI = .977, SRMR = .056. Negative affect was the comparison method factor and effects coding identification was used for identification of the remaining two factors. Standardized loadings of the method factors ranged between .307 and .546, so were of substantial magnitude. Correlation among the two factors was small but significant, .053, $p < .05$. The common consistency, occasion specificity, and *method specificity coefficients* were computed for the model using the averaging method, resulting in values equal to .210, .175, and .267. The remaining proportion was considered to be due to error, equaling approximately 35%. These results suggest that a considerable portion of the trait variance was due to method-specific variance. The percentage of stable trait variance was reduced by approximately one-third compared to the model that did not include method factors.

Trait-State-Error Models

Another modeling strategy for partitioning true score variance was proposed by Kenny and Zautra (1995, 2001) and is depicted in Figure 6.3. Initially called the *trait-state-error model*, the general modeling strategy was later renamed as the stable trait autoregressive trait and state model (START). This approach to partitioning trait and state variance incorporates an autoregressive component and can be used when only a single indicator at each occasion is available. Although the goal of separating stable and occasion-specific variance is similar, the autoregressive and single-indicator features contrast with the latent state-trait model of Steyer and colleagues.

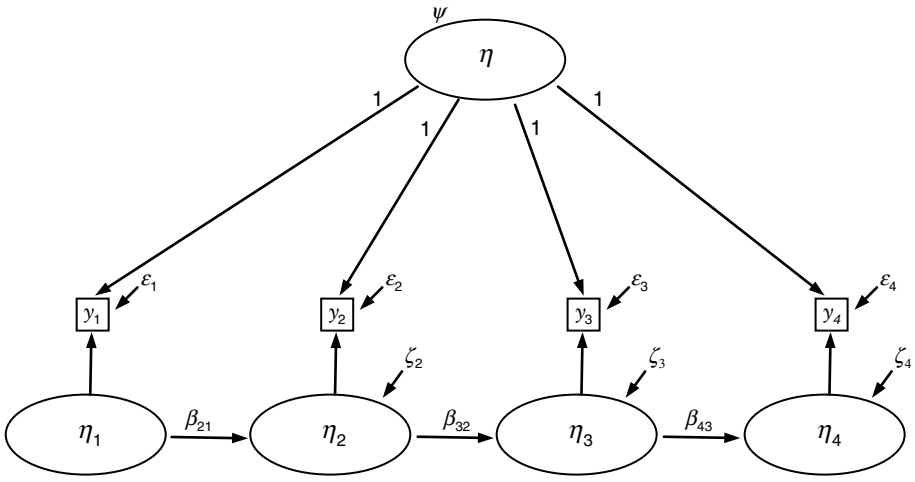

Figure 6.3 Trait-State-Error Model.

Specification

The model specifies a second-order factor to model the stable variance with each occasion-specific loading set equal to 1. Two basic equations can be used to describe the model, with each occasion-specific observed variable a function of trait, state, and error variance,

$$\mathrm{Var}\left(y_{jt}\right) = \lambda_{jt}^{2}\left(\eta_{t}\right) + \mathrm{Var}\left(\eta\right) + \mathrm{Var}\left(\varepsilon_{jt}\right) \tag{6.6}$$

and the variance of each occasion-specific state factor as a function of the prior occasion and the unaccounted for variance,

$$\mathrm{Var}\left(\eta_{t}\right) = \beta_{t,t-1}^{2}\,\mathrm{Var}\left(\eta_{t-1}\right) + \mathrm{Var}\left(\zeta_{t}\right)$$

Identification

To identify the model, stationarity of all parameters is assumed, so that the autoregression coefficient, the disturbance variance, and the measurement residual variance are set equal across waves. Estimation of the trait-state-error model with single observed variables at each occasion requires four or more waves (Kenny & Zautra, 1995). Even with this minimum requirement, difficulties still may be encountered when estimating the model. Empirical identification issues, such as failure to converge, autoregressive paths that exceed 1.0, or negative error variances, appear to arise fairly readily and additional cases or waves may be needed to avoid estimation difficulties. The authors acknowledge these potential limitations and state, "We think that it is understandable that the published applications of the model all have many waves of data (e.g., 10). It seems advisable to use this technique only with many waves of data" (Kenny & Zautra, 2001, p. 263).

If there are model estimation problems due to empirical underidentifcation, Kenny and Zautra (1995) suggest setting its measurement residual variance equal to 0 and placing a constraint on the variance of the first state (exogenous) factor by initially setting it equal to 1 and then using a manual iterative process of retesting the model with this variance set at a new value. On the second and each subsequent run, the state variance is replaced with the value $\mathrm{Var}\left(\zeta_{t}\right) / \left(1 - \beta_{t,t-1}^{2}\right)$, computed from the unstandardized estimates of the disturbance variance, $\mathrm{Var}(\zeta_{t}) = \psi_{tt}$, and the autoregression coefficient from the model from the prior run. This process is repeated until the value is trivially changed (below a minimum value of .005). Alternatively, a complex constraint can be used, if software allows, that requires the first state variance to be equal to $\mathrm{Var}\left(\zeta_{t}\right) / \left(1 - \beta_{t,t-1}^{2}\right)$.

Empirical underidentification issues are not unique to the trait-state-error model and also appear to be an issue with quasi-simplex models, which closely resemble the trait-state-error model. Marsh (1993), for example, found six out of seven quasi-simplex models had improper solutions even when the models were based on eight waves.[3] As with empirical underidentification in other contexts, a larger sample size may be helpful for preventing out of bounds estimates. Based on simulation results, Cole and colleagues (2005) suggest that a sample size of 500 or more may be needed to consistently obtain proper solutions. Smaller sample sizes may not be the only cause of estimation difficulties. Autoregressive coefficients near 1 or 0, when variables are highly stable or highly unstable, may be more difficult to model as well (Cole et al., 2005; Kenny & Zautra, 1995).

Variance Decomposition

Similar to the latent trait-state modeling approach, the unstandardized estimates of each variance component can be used to estimate the proportion of variance attributed to the trait, state, and error components, by dividing each estimate by the sum of the three components. Because loadings for each state factor and the trait factor are all set to 1 in this model and because equality constraints are used, the proportion of variance is simply computed by dividing each variance estimate by the sum of the three estimates, $\text{Var}(\eta) + \text{Var}\left(\zeta_{(t)}\right) + \text{Var}\left(\varepsilon_{(t)}\right)$.

Example 6.2: Trait-State-Error Models

A trait-state-error model (Kenny & Zautra, 1995) was fit to the depression scale data for six waves. For a model that is generally comparable to the latent state-trait model in Example 6.1, a composite variable averaging the three indicators (negative affect, positive affect, and somatic symptoms) was formed for each occasion. The model follows that shown in Figure 6.3 and, specifying a single trait factor with the loading for each indicator set equal to 1 and a latent state factor with a single loading set equal to 1 at each occasion. Each state factor was regressed on the state factor from the prior time point. Several longitudinal equality constraints were imposed (stationarity): all autoregressive coefficients were set equal, all state factor disturbances were set equal, and all measurement residuals were set equal. The trait variance and the state factor variance at Time 1 were free to be estimated and assumed independent. The model fit well, $\chi^2(16) = 81.074$, $p < .001$, CFI = .995, SRMR = .030, with no estimation difficulties. There was no need, therefore, to place a constraint on the variance of the first state factor.[4] The good fit suggests that the equality constraints for the model were not highly detrimental to the fit and that stationarity of these parameters held at least approximately. The unstandardized autoregressive parameter was .778 with standardized estimates ranging from .748 to .774. The standardized estimates increased steadily from the first to the last interval, which is common in simplex models. The proportion of trait, state, and error variance can be computed by dividing each estimate by the sum of the three, with trait = .091/(.091 + .038 + .098) = .401, state = .038/(.091 + .038 + .098) = .167, and error = .098/(.091 + .038 + .098) = .432. These estimates suggest a larger proportion of trait variance than the estimates obtained from the latent trait-state-error model without method factors (approximately 40% vs. approximately 28%), a result that might be expected given that all common variance, including method and some error variance, was treated as true score variance by computing a single composite variable of the three factors at each time point.

Extensions

The analysis can be extended in several ways. The model described above is for a lag 1 autoregressive structure with only autoregressive paths between immediately adjacent time points. Higher-order lags (lag 2 or higher) and nonlinear constraints can be modeled as well. Parallel process trait-state-error models of two or more constructs permits the examination of cross-lagged estimates of the effects of one state on another (Kenny & Zautra, 1995). Cross-lagged effects between state factors would represent occasion-specific effects after removing stable variance of the two variables. Given that the cross-lagged panel model takes into account prior values of both variables, however, one would not expect stable variance to impact the cross-lagged effects. An additional extension is the use of

latent variables with indicators at each time point, allowing for the possibility of correlated errors or method factors (Cole, Martin, & Steiger, 2005). The use of latent variables in the trait-state-error models will be discussed further in the next section.

Trait-State-Occasion Model

An extension of the trait-state-error model combines elements of the latent state-trait model with the trait-state-error model using latent variables with multiple indicators at each occasion (Figure 6.4). This model, first employed by Ormel and Schaufeli (1991) and called the *trait-state-occasion model* by Cole and colleagues (2005), specifies latent variables at each time point (state factors) that serve as second-order indicators of a single, stable trait factor. The autoregressive structure is imposed on an additional second-order latent factor specified at each time point (occasion factors). Multiple indicators are not required, but, as this is a major difference from the trait-state-error model (Kenny & Zautra, 1995), I will not focus on this variant. With single indicators at each wave, the model has a close resemblance to the state-trait-error model and is identical to a model described by Jöreskog and Sörbom (1977, Model 9b)

The variance components for the trait-state-occasion model are the same as in Equation (6.6), except that there is more than one observed variable, y_{jt}, at each occasion to define the latent state variable, η_{St}, at each time point. The occasion factors, η_{Ot}, can be viewed as phantom variables that represent residual variation in the state factors once stable variance from the trait factor is removed.[5] These phantom occasion factors can be identified by a single loading for each state factor set equal to 1, the disturbance for each state factor set equal to 0, and the variance at the first occasion set equal to 1. The autoregressive process consequently estimates the effects of the conditional variance at each time point rather than the total variance.

Multiple indicators appear to improve admissibility and there are fewer empirical identification problems (Cole et al., 2005). The additional parameters and improved estimation provide flexibility for testing stationarity assumptions. Autoregressive parameters, occasion-specific variances, and measurement residuals need not be set equal over time.

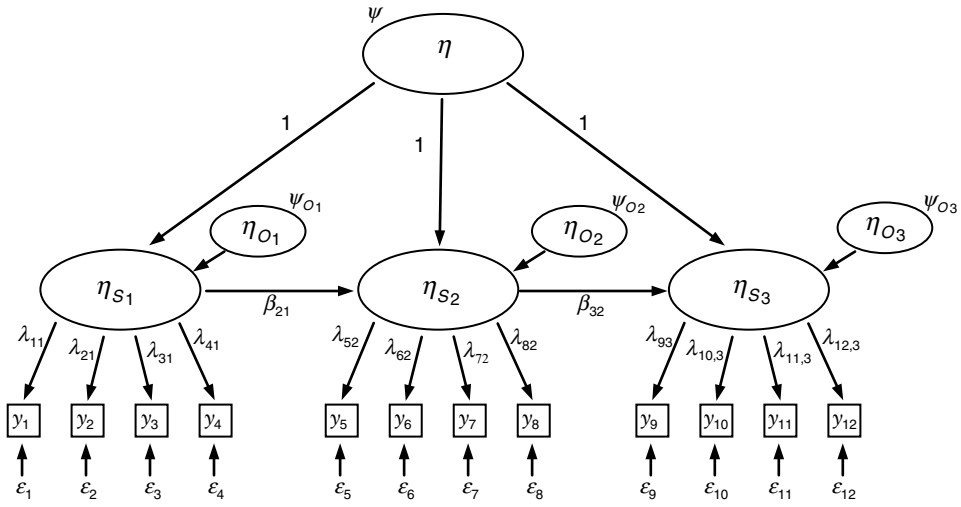

Figure 6.4 Trait-State-Occasion Model with Multiple Indicators.

The use of multiple indicators also permits correlated measurement residuals across waves or method factors to be specified. LaGrange and Cole (2008) investigated the performance of various approaches to accounting for method variance in the trait-state-occasion model and concluded that correlated measurement residuals and method factors performed similarly under conditions in which methods were truly independent. In addition to these extensions, latent means can be added to trait-state-occasion models conveniently (Cole, 2012).

Example 6.3: Trait-State-Occasion Models

A trait-state-occasion model (Cole et al., 2005) of the depression measure was tested, following the model shown in Figure 6.4. The three subscale scores, negative affect, positive affect, and somatic symptoms, were used as observed indicators of a state factor for each of the six waves. The state factors are indicators of the latent trait factor, with loadings all set to 1 (assuming equal contribution by each state factor) and disturbances set to 0. The model specifies a separate phantom occasion factor which models the residual variance remaining after the trait factor variance is accounted for and an autoregressive structure is imposed. The variance for the first phantom occasion factor and the variance of the trait factor can be freely estimated, but they are assumed to be orthogonal. Equality constraints on the autoregressive paths and disturbances for the occasion factors are not required but can be tested through nested model comparisons.

The fit of the trait-state-occasion model was poor, $\chi^2(138) = 6{,}926.700$, $p < .001$, CFI = .821, SRMR = .070, but this model did not include longitudinal measurement residual correlations or method factors to account for common variance over time. The autoregressive paths linking consecutive occasion factors were significant and increased over time, with unstandardized estimates ranging from .103 at Wave 1 to .338 at Wave 6. Standardized estimates of the autoregressive paths suggested a small to moderate effect, ranging from .108 at Wave 1 to .312 at Wave 6. The estimate of the proportion of trait, state, and error variances can be computed similarly to the trait-state-error model. The trait factor accounted for about 28% of the total variance in this model, .103/(.103 + .077 + .191) = .278. The occasion-specific variance was approximately 21% of the variance, .077/(.103 + .077 + .191) = .208. Based on the average measurement residual variance, error comprised the remaining 51% of the variance, .191/(.103 + .077 + .191) = .514.

Tests of stationarity by comparing nested models suggested that neither the autoregressive paths nor the disturbances for the occasion factors (i.e., unaccounted for variance from the autoregression) were invariant. Although the decrement in fit after constraining the disturbances to be equal, $\chi^2(142) = 6{,}950.708$, $\Delta\chi^2(4) = 24.008$, $p < .001$, $w = .102$, $\Delta Mc = .000$, was larger than the decrement in fit after constraining the autoregressive paths, $\chi^2(142) = 6{,}949.111$, $\Delta\chi^2(4) = 22.411$, $p < .001$, $w = .099$, $\Delta Mc = .000$, both effects were of small magnitude.

A modification of the trait-state-occasion model allowing for correlated measurement residuals among indicators for each subscale (negative affect, positive affect, somatic symptoms) greatly improved the fit of the model, $\chi^2(93) = 640.889$, $p < .001$, CFI = .986, SRMR = .054. The inclusion of correlated measurement residuals had very little impact on proportion of variance accounted for by each component. The trait factor accounted for about 23% of the total variance in this model, .084/(.084 + .086 + .189) = .234. The occasion-specific variance was approximately 24% of the variance, .086/(.084 + .086 + .189) = .240. Based on the average measurement residual variance, error comprised the remaining 53% of the variance, .189/(.084 + .086 + .189) = .530. The reason that including correlated measurement residuals appears to have less impact on these estimates than seen

with the latent state-trait model is that the autoregressive component takes into account specific variance that is stable over time. This is also evident in the smaller standardized autoregressive paths in the model that include correlated measurement residuals, which ranged from .093 to .209.

The trait-state-occasion was also tested specifying three method factors accounting for specific variance in the subscale scores of negative affect, positive affect, and somatic symptoms. The $M-1$ residual factor approach was used to estimate two correlated method factors, both identified using effects coding. This model also fit better than the original trait-state-occasion model that did not account for any specific variance, $\chi^2(123) = 826.322$, $p < .001$, CFI = .981, SRMR = .052, but not quite as well as the model with correlated measurement residuals. The estimates of the autoregressive paths and the factor variances were similar to the estimates obtained with the correlated measurement residual model.

State-Trait Models for Binary and Ordinal Data

There are some special considerations for binary and ordinal measured variables in estimating latent state-trait models. Naturally, any of the models discussed thus far can be estimated using full ML or WLSMV estimators, but the error component cannot be estimated for ML or if the delta method is used with WLSMV.

The Latent State-Trait Model

Eid (1996; Eid & Hoffman, 1998; see also Steyer & Partchev, 2001) discusses an alternative modeling approach, called the multistate-multitrait model, that extends the latent state-trait model to binary and ordinal variables while also showing the relationship of the multistate-multitrait model to the graded response and Rasch IRT psychometric models (Samejima, 1969; see also Chapter 1). The structure of the multistate-multitrait model is modified compared with the structure of the latent state-trait model, however, where the observed variables serve as direct indicators of both trait and state factors (Figure 6.5). In the figure, the trait factor representing stable or consistent variance is designated by η, and the state factor, representing occasion-specific variance, is designated by η_t. Only one trait factor is shown, but more than one trait factor can be specified. Although the model

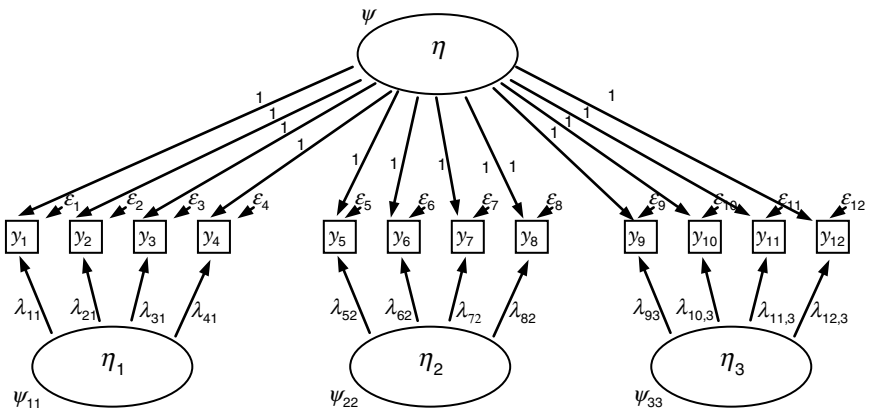

Figure 6.5 Multistate-Multitrait Model (with Single Trait).

is constructed differently from that proposed by Steyer and colleagues, the variance is still portioned into stable variance and unstable variance, because the loadings and variance estimates for the occasion-specific factors take into account (hold constant) the variance contributed by the trait factor. Part of the rationale for the multistate-multitrait model (Eid, 1996; Eid & Hoffman, 1998) seems to be the use of binary and ordinal variables. Eid (1996, p. 79) argues that models using the latent-state specification (Steyer, 1987; Schmitt & Steyer, 1993) "cannot be defined as stochastic measurement models." In fact, with equal weighting of indicators (e.g., all state loadings equal to 1), the multistate-multitrait model and the latent state-trait model are equivalent. In addition, it is fairly easy to see that the multistate-multitrait model is the same as the trait-state-occasion model (Cole, 2005) if there are no autoregression parameters.

Eid (1996) proposes decomposing the variance into only occasion-specific and consistent variance, omitting the error variance component. The rationale for this difference concerns the constraints required for the y^* distribution and the nonlinear link function. "With polytomous response variables the consistency and specificity coefficients cannot be defined in this way, as the manifest variables are not linear-additive functions of the latent trait variables, the state residuals and the error variables" (Eid, 1996, p. 71). In the multistate-multitrait formulation, the consistency coefficient and the specificity coefficient are thus computed and conceptualized in a manner that differs from Equation (6.2) and Equation (6.3), with the variance due to trait and state components summing to 1.0 (Eid, 1996). The denominator of the equation for the two coefficients is no longer the variance of Y_{jt}, but the sum of the variances of the trait and state components.

$$\text{common consistency coefficient} = \frac{\lambda_{jk}^2 \text{Var}(\eta_k)}{\lambda_{jk}^2 \text{Var}(\eta_k) + \lambda_{jt}^2 \text{Var}(\eta_t)} \tag{6.7}$$

$$\text{occasion specificity coefficient} = \frac{\lambda_{jt}^2 \text{Var}(\eta_t)}{\lambda_{jk}^2 \text{Var}(\eta_k) + \lambda_{jt}^2 \text{Var}(\eta_t)} \tag{6.8}$$

The denominator common to both equations, $\lambda_{jk}^2 \text{Var}(\eta_k) + \lambda_{jt}^2 \text{Var}(\eta_t)$, can be considered an estimate of the total variance of y^*, the unobserved continuous variable that underlies the observed discrete variable. With this formulation, the computations of the consistency and specificity coefficients do not take into account the measurement residual variances as they do in the continuous case. An alternative method would be to use the value of the residual variance y^* appropriate to the estimation method and parameterization, as described in Chapter 1, inserted into the denominator of Equation (6.7) or Equation (6.8).

$$\text{common consistency coefficient} = \frac{\lambda_{jk}^2 \text{Var}(\eta_k)}{\lambda_{jk}^2 \text{Var}(\eta_k) + \lambda_{jt}^2 \text{Var}(\eta_t) + \text{Var}(\varepsilon_{jt}^*)}$$

$$\text{occasion specificity coefficient} = \frac{\lambda_{jt}^2 \text{Var}(\eta_t)}{\lambda_{jk}^2 \text{Var}(\eta_k) + \lambda_{jt}^2 \text{Var}(\eta_t) + \text{Var}(\varepsilon_{jt}^*)}$$

The State-Trait-Error Model

The state-trait-error (or START) model cannot be conveniently estimated with binary or ordinal variables using full ML or delta-parameterization, given that the error variance component relies on measurement residual variance estimates. The theta parameterization

can be used, on the other hand, and the model can be specified in the same form depicted in Figure 6.3. As the model appears to be sensitive to empirical underidentification problems (Cole et al., 2005) and the minimum suggested sample size with WLSMV estimation is generally higher than required by standard ML with multinormal continuous variables, the state-trait-error model with binary or ordinal variables is likely to require additional sample size and time points.

The Trait-State-Occasion Model

The trait-state-occasion model described by Cole and colleagues generally presumes multiple indicators of each state factor. Under the delta parameterization the variance of these state factors are estimated with the assumption that the measurement residual variance is equal to 1 for each indicator and, this assumption, affects the scaling of the factor variance. Because the occasion-specific factors (i.e., the phantom disturbance factor) and the trait factor are both based on the variance of the same state factor as defined at each time point by its indicators, the ratio of variance of the occasion-specific and trait variance will be comparable. The delta method, however, does not partition out the error variance, suggesting that the theta parameterization approach would provide a method of estimating these models that is more comparable with the approach used for continuous variables.

Example 6.4: The Latent State-Trait Model with Binary Variables

Rather than illustrating all of the modeling approaches with binary variables, I tested only two – the latent state-trait model (Steyer, 1987; Schmitt & Steyer, 1993) and the multitrait-multistate model (Eid, 1996; Eid & Hoffman, 1998). Stability was examined for reported unwanted advice over three waves using artificially dichotomized items, coding each item as 0 for no unwanted advice and 1 for all other frequency ratings. Dichotomization was used purely for didactic reasons and I do not recommended it as a general practice. The model was specified exactly as the initial model in Example 6.1 with no method factors and each of the state factors identified using effects coding. Estimation was WLSMV with theta parameterization, but delta parameterization or full ML would also be possible. The model fit the data well, $\chi^2(30) = 30.736$, $p = .429$, CFI = .998, WRMR = .862. Using Equations (6.7) and (6.8), the relative proportion of stable and unstable variance was computed by averaging the occasion-specific factor variances. Using the estimate of the residual variance, $\text{Var}(\varepsilon_{jj}^*)$, as unity in the common consistency coefficient computation suggested that the percentage of variance due to the trait factor was approximately 32% (.319), the percentage due to state factors on average was approximately 34% (.338), and the percentage due to error was approximately 34% (.343). These results suggested approximately equal proportions of variance due to stable and unstable variance.

A second model using the same data illustrated the multistate-multitrait model as shown in Figure 6.5 (Eid, 1996; Eid & Hoffman, 1998), also estimated with WLSMV theta parameterization. This model also fit the data well. The chi-square value that was nearly identical to that obtained with the latent state-trait model, $\chi^2(32) = 32.092$, $p = .462$, CFI = 1.000, WRMR = .884. The coefficients computed using the average estimate of $\text{Var}(\varepsilon_{jt}^*)$ indicated that the percentages due to trait, state, and error were approximately 32% (.320), 34% (.340), and 34% (.340), respectively.

Extensions

The state-trait models introduced in this chapter can be extended in a variety of ways (Steyer, Geiser, & Fiege, 2012 review some of the possible models). Instead of a single stable

trait, the latent trait-state model can be modified to include multiple traits (Eid, 1996; Eid & Hoffman, 1998; Steyer & Partchev, 2001), the *multiple trait-multiple state model*. This model suggests a somewhat fine line between interpretation of method factors and multiple traits. To the extent that only a single trait is specified and method factors are included, the model can be interpreted as partitioning total variance into a variance due to a single trait and multiple methods (or subfactors). With multiple traits specified but no additional method factors specified, there is no ability to differentiate between method factors and unique trait factors. If data that can be used to model both multiple traits and multiple methods are available, partitioning of stable variance due to multiple stable constructs and multiple methods is possible, making it possible to separate state and error variance.

In a related approach, distinct subfactors can be modeled with their own latent state-trait models and then combined in a hierarchical factor model. Schermelleh-Engel and colleagues (Schermelleh-Engel, Keith, Moosbrugger, & Hodapp, 2004) describe how third-order trait factors can be specified using traits from separate latent state-trait models as indicators. In addition to the use of correlated measurement residuals, method factors, and multitrait- multistate models, this strategy represents yet another option for conceptualizing common method variance that might derive from subfactors of a scale or methods of measurement.

Latent-state models can also be extended to accommodate more complex longitudinal hypotheses. The available time series can be divided up meaningfully into segments to examine before and after type changes in trait or state variance. Eid and Hoffman (1998) illustrate models comparing trait variance in student interest in nuclear issues prior to and after the Chernobyl incident. The general concept of change in levels of trait variance can be further generalized to show the relationship between growth models and latent state-trait models (Geiser, Lockhart, & Keller, 2013; Tisak & Tisak, 2000). Chapter 7 revisits the relation between latent state-trait models and growth curve models.

Development in this area is likely to emphasize modeling intraindividual variation in stability. Hamaker and colleagues (Hamaker, Nesselroade, & Molenaar, 2007) propose an *integrated trait-state model* designed to model individual state factors based on the *p*-factor technique allowing for the potential variability in state estimates across individuals. This model requires specialized software to estimate. In an attempt to describe such individual differences, Courvoisier and colleagues discuss latent classes of latent trait-state models to identify meaningful population subgroups (Courvoisier, Eid, & Nussbeck, 2007).

Comments

The primary difference between the latent state-trait model of Steyer and colleagues and the trait-state-error model of Kenny and Zautra involves the inclusion of an autoregressive component for the trait variance. Cole and colleagues (Cole et al., 2005; Cole, 2012) propose a useful framework for integrating the two approaches with the flexibility of including or not the autoregressive component within the same general modeling approach. The inclusion of an autoregressive structure on state factors may require careful theoretical consideration, because autoregression estimates for state or occasion factors involve an association that remains after accounting for the common trait.

The underlying similarity between common traits and autoregression processes (Jöreskog & Sörbom, 1979; Marsh, 1993) suggests that researchers must carefully distinguish between more labile, local stability and more permanent, global stability. The autoregression process reflects a kind of local stability in the variable because the values covary across adjacent time points for less than the full time series. Inclusion of autoregression paths of higher order beyond lag 1 would further blur the lines between local and

global stability. Stability, whether local or global, needs to be interpreted in the context of the theorized processes and the interval length used in the study.

Modeling consistent variance in latent state-trait models is based on the autocorrelation notion of stability rather than the absolute level notion of stability. It is possible for the value of a trait to increase or decrease over time without having any bearing on its stability operationalized as autocorrelation. The models and examples in this chapter have emphasized relative amounts of stable and unstable variance in aggregate. It also is quite reasonable and often desirable to examine predictors or consequences of trait variance, and, in so doing, we examine interindividual differences of the trait. Nesselroade and colleagues (Burr & Nesselroade,1990; Hertzog & Nesselroade, 1987) distinguish between interindividual variability and intraindividual variability. The latter involves changes across time within an individual that are entailed when examining fluctuations in state variance or growth. These authors also posit a distinction between intraindividual changes in states and traits, with changes in state variance involving change of a less developmental nature (e.g., changes in mood states) and changes trait variance of a more long-term developmental nature (e.g., verbal ability). The distinction between stable trait variance and changing trait variance, such as developmental increases or decreases in a characteristic, may be a subtle, if not elusive, one. Such a distinction may be more theoretical than mathematical, hinging on whether the time frame under study is long enough to be considered developmental.

The autoregressive process for states is likely to be inflated to the extent that specific variance common across repeated measures is not taken into account (Davey, 2001). To address this issue, correlated measurement residuals can be estimated when multiple indicators are available. Most discussions of state-trait models have only addressed correlated specific variance in the context of method factors. Where method factors are not relevant, the topic of modeling correlated specific variance over time has often been overlooked. Autoregressive paths for trait-state-error models for single indicators at each occasion confound any correlated measurement residuals, and the use of multiple indicators at each occasion, as implemented in the trait-state occasion model (Cole et al., 2005), has the flexibility to include this component.

It is worthwhile to reiterate that "states" and "traits" are merely convenient labels for any type of temporarily fluctuating or unstable variance and any type of consistent or stable variance. Application of the state and state distinction and the state-trait models to other fields beyond personality research, be it social systems, economic factors, or biological processes, will no doubt be highly fruitful for generation of knowledge and theory in many disciplines.

Recommended Readings

Two recent reviews describe some of the modeling approaches to partitioning state and trait variance (Cole, 2012; Steyer et al., 2012). Some interesting theoretical papers from psychology (Fraley & Roberts, 2005; Roberts & DelVecchio, 2000), sociology (Geels, 2004), and macroeconomics (Ahmed, Levin, & Wilson, 2014)) provide useful context on how stability and change are conceptualized in these disciplines. More understanding of the basics of trait and state models can be gained by comparing them to autoregressive and common factor models (Jöreskog & Sörbom 1977; Kenny & Campbell, 1989). Multimethod-multimatrix models are commonly integrated into trait-state error models and there are a number of complex issues related to how they are best specified (Geiser & Lockhart, 2012; Marsh, 1989: Marsh & Grayson, 1995; Tomás, Hontanges & Oliver, 2000). Finally, more can be gleaned from considering how concepts of stability can be contrasted with the concept of growth (Tisak & Tisak, 2000), the topic that is the focus of the next chapter.

Notes

1 There has been long-standing debate in personality psychology about what constitutes a trait, whether traits exist, and how much they drive human behavior (e.g., Mischel & Shoda, 2010; Sherman, Nave, & Funder, 2010), which I will not enter into. Clearly, some behavioral, cognitive, social, or attitudinal characteristics are more stable over time than others and, for many researchers, tools for understanding the factors predictive of or predicted by stable versus more labile aspects of a construct will be considered valuable.

2 Each observed indicator in this model can be said to be an "item parcel," but item parcels have a number of potential pitfalls (e.g., Bandalos, 2002). They are used here out of pedagogical convenience to avoid a more complicated third-order factor model. Parcels are generally less problematic if their components can be shown to be from a common factor with a strong relationship to the factor. Confirmatory factor models of Center for Epidemiologic Studies-Depression measure on these data and in other studies (Schafer, 2006) have often supported the existence of these three factors (and less often supported a fourth factor made up of interpersonal judgment items).

3 The resemblance of the two models suggests some minor alternative specifications of the model that may improve estimation. In the quasi-simplex, the first factor variance is set equal to 1 and the measurement residual is set equal to 0 for the first and last time point. Kenny and Zautra (1995, 2001) only mention constraints on the first measurement residual, and an additional constrain on the final time would presumably improve estimation. These two constraints would also likely necessitate an alternative specification of the first and last loading on the trait factor so their loadings are not set to be equal to the loadings on the indicators at the middle time points.

4 Placing a nonlinear constraint on the variance to $\mathrm{Var}(\zeta_t)/(1-\beta^2_{t,t-1})$ in this case resulted in a very similar value to that obtained when the variance was freely estimated.

5 The terminology regarding "states" and "occasions" can be a bit confusing. The Kenny and Zautra trait-state-error model estimates autoregressive paths between what they refer to as "state" factors. Their "state" factors, however, can be viewed as the same as the "occasion" factors referred to by Cole and colleagues (Cole et al., 2005; Cole, 2012). In either case, however, the phantom variable is essentially just a strategy for modeling the conditional variance at each time point after removing the trait variance. Without the autoregressive component, there is really no difference between a single longitudinal factor model that assumes equal loadings or stable variance. Such a model would provide equivalent estimates to the variance of the occasion-specific disturbance variance from the latent state-trait model of Steyer and colleagues.

References

Adams, J., Clark, M., Ezrow, L., & Glasgow, G. (2004). Understanding change and stability in party ideologies: Do parties respond to public opinion or to past election results? *British Journal of Political Science*, *34*, 589–610.

Ahmed, S., Levin, A., & Wilson, B. A. (2014). Recent US macroeconomic stability: good policies, good practices, or good luck? *Review of Economics and Statistics*, *86*, 824–832.

Burr, J., & Nesselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research. Vol. 1: Principles and structuring change* (pp. 3–34). New York: Academic Press.

Cole, D. A. (2012). Latent trait-state models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 585–600). New York: Guilford.

Cole, D. A., Martin, N. M., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait–state models: Introducing a trait–state–occasion model. *Psychological Methods*, *10*, 3–20.

Costa, P. T., & McCrae, R. R. (1988). Personality in adulthood: a six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology*, *54*, 853–863.

Courvoisier, D. S., Eid, M., & Nussbeck, F. W. (2007). Mixture distribution state–trait-models: Basic ideas and applications. *Psychological Methods*, *12*, 80–104.

Davey, A. (2001). The trait-state distinction and its dialectic balance. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp.265–269). Washington, DC: American Psychological Association.

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research Online*, *1*, 65–85.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241–261.

Eid, M., & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics*, *23*, 193–215.

Fraley, R. C., & Roberts, B. W. (2005). Patterns of continuity: a dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60–74.

Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy*, *33*, 897–920.

Geiser, C., Keller, B. T., & Lockhart, G. (2013) First- versus second-order latent growth curve models: Some insights from latent state-trait theory, *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 479–503.

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, *17*, 255–283.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait-state model. *Journal of Research in Personality*, *41*, 295–315.

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development*, 93–109.

Jöreskog, K. G., & Sörbom, D. (1977). Statistical models and methods for analysis of longitudinal data. In D. J. Aigner & A. S. Goldberger (Eds.), *Latent variables in socio-economic models* (pp. 285–325). Amsterdam: North Holland.

Jöreskog, K. G., & Sörbom, D. (1979). Basic ideas of factor and component analysis. In J. Magidson (Ed.), *Advances in factor analysis and structural equation models*. (pp. 5-20) Cambridge, MA: Abt Books.

Kenny, D. A, & Campbell, D. T. (1989). On the Measurement of Stability in Over-Time Data. *Journal of personality*, *57*, 445–481.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*, 52–59.

Kenny, D. A., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change. Decade of behavior* (pp. 243–263). Washington, DC: American Psychological Association.

Kropf, J., and Hudson, N. (2012). Current Employment Statistics seasonal adjustment and the 2007–2009 recession. *Monthly Labor Review*, *135*, 42–53.

LaGrange, B., & Cole, D. A. (2008). An expansion of the trait-state-occasion model: accounting for shared method variance. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 241–271.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural analysis* (4th ed.). Hillsdale, NJ: Laurence Erlbaum Associates.

Mahoney, J., & Thelen, K. (2010). *Explaining institutional change: Ambiguity, agency, and power.* Cambridge, UK: Cambridge University Press.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*, 335–361.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement*, *30*, 157–183.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 177–198). Thousand Oaks, CA: Sage.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. *Psychological Bulletin*, *97*, 562–582.

Mischel, W., & Shoda, Y. (2010). "The situated person." In B. Mesquita, L. Feldman Barrett, & E. R. Smith (Eds.), *The mind in context* (pp. 149–173). New York, London: Guilford Press.

Ormel, J., & Schaufeli, W. B. (1991). Stability and change in psychological distress and their relationship with self-esteem and locus of control: a dynamic equilibrium model. *Journal of Personality and Social Psychology*, *60*, 288–299.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3–25.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Iowa City: Psychometric Society.

Schmid, J., & Leiman, J. N. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53–61.

Schermelleh-Engel, K., Keith, N., Moosbrugger, H., & Hodapp, V. (2004). Decomposing person and occasion-specific effects: An extension of latent state-trait (LST) theory to hierarchical LST models. *Psychological Methods*, *9*, 198–219.

Schmitt, M. J., & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, *14*, 519–529.

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*, 123–146.

Sherman, R. A., Nave, C. S., & Funder D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology*, *99*, 330–343.

Steyer, R. (1987). Konsistenz und Spezifitat: Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihder Identifikation [Consistency and specificity: Definition of two central concepts of differential psychology and a simple model for their determination]. *Zeitschrift for Differentielle und Diagnostische Psychologie*, *8*, 245–258.

Steyer, R., Geiser, C., & Fiege, C. (2012). Latent state-trait models. In H. Cooper (Ed.), *Handbook of research methods in psychology: Vol. 3. Data analysis and research publication* (pp. 291–308). Washington: American Psychological Association.

Steyer, R., Majcen, A.-M., Schwenkmezger, P., & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research*, *1*, 281–299.

Steyer, R., & Partchev, I. (2001). Latent state-trait modeling with logistic item response models. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 481–520). Lincolnwood, IL: Scientific Software International.

Steyer, R., & Schmitt, M. (1990). Latent state-trait models in attitude research. *Quality & Quantity*, *24*, 427–445.

Tack, W. H. (1980). Zur Theorie psychometrischer Verfahren – Formalisierung der Erfassung von Situationsabhängigkeit und Veränderung. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *1*, 87–106.

Tisak, J., & Tisak, M. S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychological Methods*, *5*, 175–198.

Tomás, J. M., Hontangas P. M., & Amparo, O. (2000). Linear confirmatory factor models to evaluate multitrait-multimethod matrices: The effects of number of indicators and correlation among methods. *Multivariate Behavioral Research*, *35*, 469–499.

Wolff, H -G., & Preissing, K. (2005). Exploring item and higher order factor structure with the Schmid–Leiman solution: Syntax codes for SPSS and SAS. *Behavior Research Methods*, *37*, 48–58.

Keywords
state-trait models, latent trait-state-error models, method factors, trait-state occasion model, state-trait-error model

7 Linear Latent Growth Curve Models

Growth curve models investigate level change in a variable over time. The two general approaches, growth curve analysis using multilevel regression (hierarchical linear modeling) and latent growth curve models using SEM, have undoubtedly become the most popular method of analyzing longitudinal data. Growth curve models, in general, represent a very flexible approach to modeling change that allows for investigation of linear and nonlinear trends and individual differences in these trends. For basic models, the two methods of estimating them are equivalent, but latent growth curve models offer some additional flexibility as well as the potential for incorporating growth curves into the general SEM framework, making it possible to investigate the relationship between latent predictor variables on growth, the effects of growth on other factors, mediational hypotheses, and modeling of parallel growth curves. This chapter provides a general introduction to linear latent growth curve models, and the following chapter focuses on nonlinear models.

Meredith and Tisak (1984, 1990) are generally credited with the inception of modern latent growth curve analysis by formalizing earlier work on exploratory factor analysis of growth (e.g., Baker, 1954; Rao, 1958; Tucker, 1958). They proposed latent variables with repeated measures as indicators, with and without special constraints on the loadings, in order to account for change over time. The concurrent development of this approach with the multilevel regression approach to growth (Bock, 1983; Bryk & Raudenbush, 1987; Laird & Ware, 1982) has resulted in an explosion of application of these strategies to investigating change in social sciences and other disciplines (for more on the history of latent growth curve models, see Bollen, 2007; Bollen & Curran, 2006). Latent growth curve models exploit the measurement portion of the structural equation model to estimate the variable of interest as some function of time.

The chapter begins with an introduction of the general concepts of linear growth models and then explains the estimation of latent growth curve models within the structural equation modeling framework. In addition to these fundamental concepts, a variety of other essential issues from model identification to analysis of discrete variables to second-order models are described.

Latent Growth Curve Models with Observed Continuous Variables

General Concepts

The goal of latent growth curve modeling is to describe the relationship between repeated measurement of a variable and some metric of time. Although estimated by a set of simultaneous equations, a simple way to begin is to consider one part of the model, a regression for an individual case in a data set.

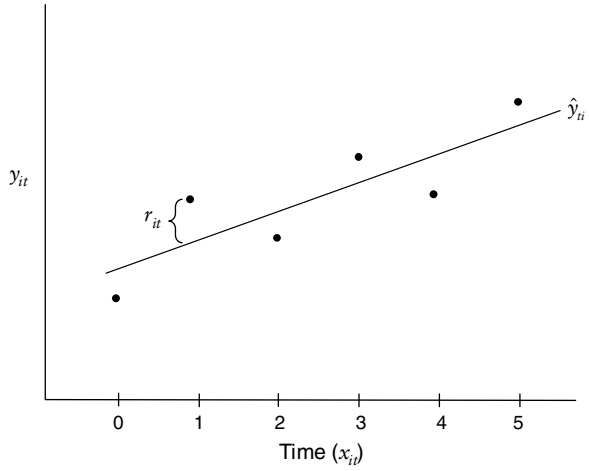

Figure 7.1 Hypothetical Linear Growth for a Single Case.

$$y_{ti} = \beta_{0i} + \beta_{1i} x_{ti} + r_{ti} \tag{7.1}$$

The subscript *ti* is used to represent a score at time point *t* for some individual case *i*, emphasizing that the regression model is over time for one case. The equation includes an observed value for the case at each time point, y_{ti}, and two constants, β_{0i} and β_{1i}, representing the *y*-intercept and slope, respectively. If x_{ti} is a variable that is some function of time, whether it is time codes representing the waves of the study (e.g., 0,1,2, …), age, grade, years, or some other metric of time, then the regression line describes the change in y_{ti} for each increment in time. The slope for such a regression represents an increase in the level of the dependent variable over time if it is positive, a decrease over time if it is negative, or no change if the slope is equal to 0. Across various authors and contexts, the slope in these models is referred to interchangeably as the rate of change, the trajectory, or growth.

Figure 7.1 depicts a hypothetical relationship between x_{ti} and y_{ti} and illustrates that the observed points do not all fall exactly on the line, and, thus, there is some error in prediction, r_{ti}. Because the line represents predicted values, $\hat{y}_{ti}$, for a particular value of x_{ti}, the residual is the deviation of the observed score from the predicted value, $r_{ti} = y_{ti} - \hat{y}_{ti}$. The variance of these residuals can be used to quantify the degree to which there is variance within each case that is unexplained by the time variable, often described as within-person variance when cases are people. The intercept, β_{0i}, is the value of y_{ti} when x_{ti} equals 0. If x_{ti} is the wave number starting with 0 (0, 1, 2, …T – 1), then the intercept represents the expected value of y_{ti} at the first wave of the study ($t - 1 = 0$).

To describe the entire data set, we estimate the expected value of y_{ti} at the beginning of the study, by taking the average of all the intercepts, γ_{00}, where the intercept for any given case may deviate from that average by u_0.

$$\beta_{0i} = \gamma_{00} + u_{0i} \tag{7.2}$$

The variance of the residual, Var(u_{0i}), gives the variability across cases of the expected value of y_{ti} at the beginning of the study assuming the above-mentioned coding of the time

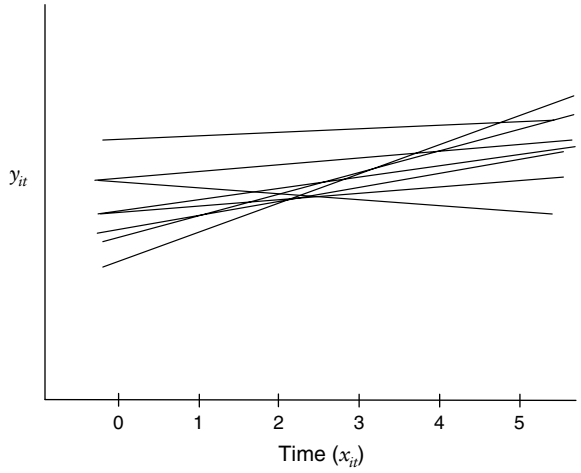

Figure 7.2 Hypothetical Linear Growth for Several Cases.

variable. Similarly, we could estimate the average slope by using γ_{10} and the deviation of any individual's slope from that average by u_{1i}, having variance Var(u_{1i}).

$$\beta_{1i} = \gamma_{10} + u_{1i} \tag{7.3}$$

The variability of the intercepts and the slopes are illustrated in Figure 7.2, a plot of a set of hypothetical linear growth curves. The variability of the intercepts can be seen by the differences among the starting points of each predicted line at the left and the variability of the growth can be seen by the different angle of the slopes for each individual case.

Inserting Equations (7.2) and (7.3) into Equation (7.1) gives a single multilevel equation for time points nested within individuals.

$$y_{ti} = (\gamma_{00} + u_{0i}) + (\gamma_{10} + u_{1i})x_{ti} + r_{ti}$$
$$y_{ti} = \gamma_{00} + \gamma_{10}x_{ti} + u_{1i}x_{ti} + u_{0i} + r_{ti}$$

The final equation above can be viewed as a linear regression model representing a line with the first two terms on the right, $\gamma_{00} + \gamma_{10}x_{ti}$, resembling a standard linear regression model, and the remaining terms representing three components of error, $u_{1i}x_{ti} + u_{0i} + r_{ti}$ instead of just a single error term. This is the multilevel regression model or hierarchical linear model applied to longitudinal data, with time points nested within cases. The term u_{0i} represents the extent to which the intercept value for any given case deviates from the average intercept for the sample. The variance of the residual, r_{ti} is sometimes characterized as *intraindividual variation*, whereas the variance of the residual u_{0i} is sometimes characterized as *interindividual differences*. The residual u_{1i} provides information about interindividual differences in change, or the extent to which the slope for any given case deviates from the average slope.

Latent Growth Curve Model

SEM can be used to construct equations parallel to the two sets of equations for the multilevel regression growth curve analysis by using special constraints on the measurement

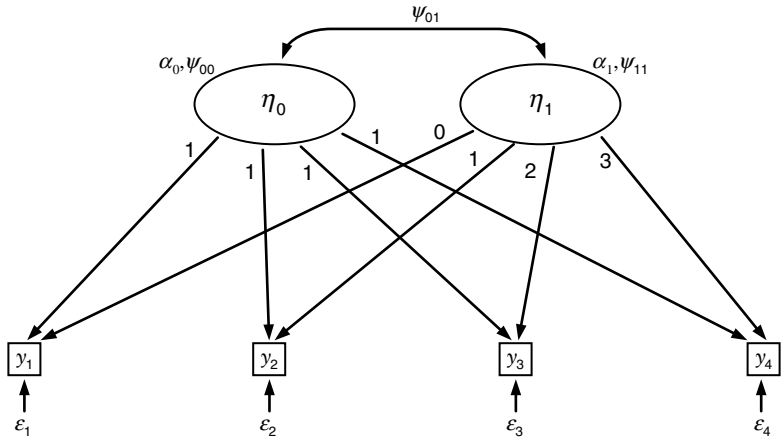

Figure 7.3 Latent Growth Curve Model.

portion of the model (refer to Figure 7.3). The case-level regression is modeled by specifying two latent variables, one latent variable, η_{0i}, representing the intercept and the other latent variable, η_{1i}, representing the slope.[1]

$$y_{ti} = \lambda_{t0}\eta_{0i} + \lambda_{t1}\eta_{1i} + \varepsilon_{ti}$$

If loadings for the intercept factor are set equal to 1 for all time points, then λ_{t0} is a constant and the term $\lambda_{t0}\eta_{0i}$ reduces to the intercept α_{0i}.

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\eta_{1i} + \varepsilon_{ti} \tag{7.4}$$

It then becomes clear that Equation (7.4) is directly parallel to Equation (7.1) for the special case in which the loadings λ_{t1} are given values that represent time codes, such as 0, 1, 2, 3, …T – 1. The latent growth curve model is therefore conceptually equivalent to the growth curve model with multilevel regression (Chou, Bentler, & Pentz, 1998).[2] Estimation of the mean and variance of the two factors gives equations directly parallel to Equations (7.2) and (7.3), with the means of η_0 and η_1 represented by α_0 and α_1, respectively, and the deviation of each individual value of the latent variable from each of their respective means represented by ζ_{0i} and ζ_{1i}.

$$\eta_{0i} = \alpha_0 + \zeta_{0i} \tag{7.5}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i} \tag{7.6}$$

To identify the factor means, the measurement intercepts at each time point, ν_t, are generally set equal to 0. Variances of the deviations can be estimated in the structural model by $\mathrm{Var}(\zeta_{0i}) = \psi_{00}$ and $\mathrm{Var}(\zeta_{1i}) = \psi_{11}$. Equations (7.5) and (7.6) can be substituted into Equation (7.4) to create a single multilevel equation.

$$\begin{aligned} y_{ti} &= (\alpha_0 + \zeta_{0i}) + \lambda_{t1}(\alpha_1 + \zeta_{1i}) + \varepsilon_{ti} \\ y_{ti} &= \alpha_0 + \alpha_1\lambda_{t1} + \zeta_{0i} + \zeta_{1i}\lambda_{t1} + \varepsilon_{ti} \end{aligned} \tag{7.7}$$

The last three terms, $\zeta_{0i} + \zeta_{1i}\lambda_{t1} + \varepsilon_{ti}$, constitute the multilevel variance components, with $\mathrm{Var}(\zeta_{0i}) = \psi_{00}$ and $\mathrm{Var}(\zeta_{1i}) = \psi_{11}$ providing information about the level-2 variance and

$\text{Var}(\varepsilon_{ti}) = \theta_{tti}$ providing information about the level-1 variance. The latent growth curve model can then be expressed in terms of the common structural equation model:

$$y_{ti} = \eta_0 + \eta_1 \lambda_{t1} + \varepsilon_{ti},$$

where loadings are values associated with a time metric (e.g., 0, 1, 2, …$T-1$) and the means and variances of the intercept and slope factors are estimated.

Interpretation of Coefficients

The average slope, represented by α_1, indicates whether there is a linear increase or decrease over time on average. Significance of the slope indicates that, on average, there is an increase or decrease in the level of the dependent variable over time that is statistically different from 0. The variance of the slope, ψ_{11}, represents the individual differences in growth, with a significant effect suggesting that the slopes for individual cases are not parallel (as depicted in Figure 7.3). If the slope variance is zero, it indicates that the slopes for each individual case in the data set are parallel. Rarely would the variance be exactly zero, but a nonsignificant result would suggest that the slopes are not significantly different across cases in the data set.[3]

The average intercept and slope are referred to as *fixed effects* and their variances are referred to as *random effects* in multilevel parlance, because the estimation of variances of the intercepts and slopes assumes these quantities are random. Accordingly, multilevel or growth curve models are often referred to as random effects models, or random coefficient models, or mixed models (i.e., a mix of random and fixed effects). Either variance estimate can be set to 0 or another value to create a non-varying parameter, although this is rare in practice unless empirical underidentification issues are encountered.

As depicted in Figure 7.3, the correlation between the intercept and slope factors is typically estimated. Without estimating this correlation, the interpretation of the regression model changes, because the "effects" of the two latent variables on the observed variables are no longer independent. Assuming the correlation between intercept and slope factors is zero would change the interpretation of the intercept factor, where it can no longer be seen as the value of y_{ti} when x_{ti} is equal to 0 (e.g., the average of the dependent variable at the first time point). The correlation between the intercept and slope factors can be complicated to interpret. Simply put, the correlation represents the association between an individual's score at baseline (assuming the common coding scheme for the slope loadings) and change over time. A positive correlation indicates higher baseline scores are associated with a higher slope value. A "higher slope value" is a less negative slope or a larger positive slope. Plots of a subset of predicted slopes can be helpful in understanding the results.

Some reflection on whether the average slope value is positive or negative also can be helpful for understanding the intercept-slope relationship. Figure 7.4 illustrates four hypothetical relationships for a growth model of perceived economic security. If an average increase in perceived security is obtained, then a positive correlation between intercept and slope would suggest that higher perceived security at the beginning of the study would be associated with greater increases in perceived security over time (Figure 7.4a). With an average increase in growth in perceived security but a negative correlation between intercept and slope, higher perceived security initially would be associated with lower rate of increase in security or a decline in security over time (Figure 7.4b). If security declines on average, a positive correlation between intercept and slope would suggest that higher initial perceived security was associated with a slower rate of decline in security over time or an increase in security over time (Figure 7.4c). If security declines on average but there is a negative correlation between intercept and slope, then those individuals with a high initial

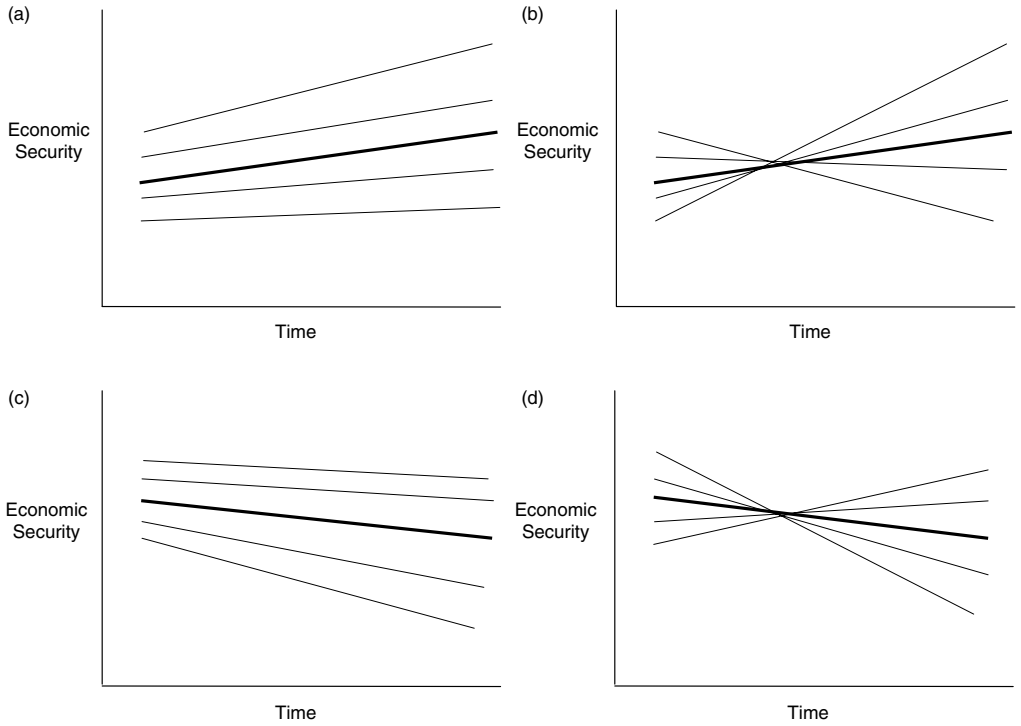

Figure 7.4 Four Possible Hypothetical Relationships between Intercept and Slope for a Growth Model of Perceived Economic Security. (a) average increase with positive correlation between intercept and slope; (b) average increase with negative correlation between intercept and slope; (c) average decrease with positive correlation between intercept and slope; (d) average decrease with negative correlation between intercept and slope.

level of perceived security will to tend to decrease more rapidly or increase less rapidly in their perceived security over time (Figure 7.4d).

Time Coding

Because we can interpret Equation (7.4) as any other regression, the intercept is the expected value for y_{ti}, when x_{ti} equals 0. If the loadings for η_1 are specified so that they begin with 0, then the intercept mean, α_0, can be interpreted as the expected value or sample average of y_{ti} at the first time point. By far, 0, 1, 2, 3, ….$T-1$ is the most commonly used coding scheme for the slope factor loadings, although other coding schemes are certainly possible. In some instances, a researcher may be interested in an intercept that can be interpreted as the middle time point (e.g., $\lambda_{t1} = -2, -1, 0, 1, 2$) or the last time point (e.g., $\lambda_{t1} = -4, -3, -2, -1, 0$). Alternative coding schemes can also be used to accommodate unequal spacing between waves or nonlinear trends. Nonlinear change is the focus of Chapter 8.

Assuming a linear coding scheme with increments of one unit, changes in the loadings for the slope factor will leave the mean of the slope factor unaltered from the estimates obtained with values of 0, 1, 2, 3, …$T-1$. The intercept, however, will be changed by altering the coding scheme and this will also affect its variance. Imagine for a moment

moving the *y*-axis so that the value of the intercept (where x_{ti} equals 0) is modified. It can be seen from Figure 7.4a how such a change in the scaling of the *x*-axis would affect the estimate of the variance of the intercept and its associated significance test. Moving the axis to the right would no doubt increase the variability of the intercepts. The relationship between the intercept and slope factors or the relationship of the intercept factor with a covariate will also be affected by such a coding change. It may be troublesome that an arbitrary scaling change could affect the estimates and their significance, but the scaling change modifies how the intercept is defined and, therefore, the analytic question. Neither scaling of the intercept is correct nor incorrect, but care must be taken to be certain that the conclusions properly match the intercept interpretation in light of the coding scheme used for the slope factor loadings.

Relation to Other Models

It is interesting to note that if the slope factor is omitted and only a single "intercept" factor is specified with loadings equal to 1, then the model is the same as a trait factor concept used in latent state-trait models discussed in Chapter 6. Without the slope factor, the single latent variable is equally weighted by all the repeated measures, and, therefore, represents a sole underlying, stable cause of responses at every time point. Any occasion-specific variance is estimated in the measurement residual for each time point. The relation of the latent growth curve model to latent state-trait models suggests several extensions, such as incorporating method factors, multiple indicators at each occasion, and autoregressive processes.

If the latent growth curve model is specified for two time points, with loadings of the slope factor set equal to 0 and 1 for the two time points, respectively, it is equivalent to the contrast coding model to repeated measures ANOVA discussed in Chapter 3. With only two time points, a constraint on the intercept factor variance, the slope factor variance, or the measurement residual variances will be needed for identification. The connection between these two models exemplifies several useful points. One crucial point is that the random effects ANOVA model is a special case of the multilevel and growth curve models (Newsom, 2002; Raudenbush, 1993). Because the repeated measures ANOVA and paired *t*-test are also tests of the average difference score (see Chapter 3), the latent growth curve model is a generalization of the analysis of difference scores. In fact, the test of the linear trend over three time points is contained in the test of the average slope from the latent growth curve model if loadings of the slope factor are set to –1, 0, and 1. The advantage of the growth curve model, however, is that, with an increasing number of time points, the estimate of the rate of change over time becomes less influenced by errors or occasion-specific fluctuations at any given time point and therefore becomes a more reliable estimate of level change.

The link between the growth model slope and the difference score analysis can be seen by examination of the simple formula for calculating the slope for the growth model in Equation (7.1) (Rogosa, Brandt, & Zimowski, 1982).

$$\beta_{1i} = \frac{y_{ti} - y_{t-1,i}}{x_{ti} - x_{t-1,i}}$$

The numerator and denominator represent difference scores calculated from two repeated measures, where x_{ti} and $x_{t-1,i}$ are scores representing a time metric. Most simply, the *x* scores would be 0 and 1, just as in the first two loadings of the slope factor in the latent growth curve model using the common coding scheme. In this simplified case, the

denominator becomes 1 and the equation reduces to $\beta_{1i} = y_{ti} - y_{t-1,i}$, indicating that the individual slope is just the difference between the two dependent variable scores.

Taking the equivalence to difference scores a bit further is useful for understanding the nature of the intercept and slope. As we know that the expected value of the difference of two scores is equal to the difference of the expected values of each score, we also can restate the average slope, α_1, as the difference between two means.

$$\alpha_1 = E\left(y_{ti} - y_{t-1,i}\right) = E\left(y_{ti}\right) - E\left(y_{t-1,i}\right)$$

With the 0 and 1 codes for the time variable, we know that the average intercept is the expected value of the dependent variable at the first time point, so we can substitute the average intercept into the equation above.

$$\alpha_1 = E\left(y_{ti}\right) - \alpha_0 \tag{7.8}$$

This relation then implies that the expected value at the second time point, y_{ti}, is equal to the sum of the intercept and slope means for a latent growth curve model with two time points (Bollen & Curran, 2006).

$$E\left(y_{ti}\right) = \alpha_0 + \alpha_1$$

In other words, we can arrive at the mean of the second time point from the mean of the first time point by an increment equal to the average value of the slope. The general point of all this is that the linear growth curve model is closely related to the difference score analysis. Stating that the dependent variable is a function of a unit increment in time is equivalent to assessing the difference between two time points. With more than two time points, a linear slope represents the average change per unit increase in time.

The correlation between intercept and slope is estimated in common applications, but it is interesting to know that a latent growth curve model with no correlation between intercept and slope factor is equivalent to the simplex model with autoregressive paths set equal to 1 and disturbances set equal across time points (Bianconcini, 2012; Raykov, 1998). This particular specification of the simplex model also closely resembles the latent difference score model (McArdle, 1988). This should not seem too surprising if you recall from Chapter 4 that a simple autoregression between two time points is equal to the difference score computed from the two scores if the autoregression coefficient is set equal to 1. Latent difference score models that generalize this analysis to more than two time points are the focus of Chapter 9.

Intraclass Correlation Coefficient

Consider for a moment an intercept only model with no growth factor. For such a model, the variance of the intercept represents the interindividual or between-person variance in the dependent variable. The variance of the residuals represents intraindividual or within-person variance. The proportion of the total variance of y_{ti} due to these two components is often quantified by the *intraclass correlation coefficient* (ICC). The ICC is the proportion of between-group variance relative to total variance and can be calculated by a ratio of the variance of the intercept factor, $\text{Var}(\eta_0) = \psi_{00}$, to total variance.[4]

$$\rho = \frac{\psi_{00}}{\text{Var}\left(y_{ti}\right)} = \frac{\psi_{00}}{\psi_{00} + \theta_{(tt)}} \tag{7.9}$$

Here, I assume for convenience that the measurement residual variances are set equal across waves, so that there is a single value for the residual, with $\text{Var}(\varepsilon_{ti}) = \theta_{(tt)}$. The ICC

is the proportion of total variance due to interindividual or between-person variance and can be used to gauge the extent to which scores are a function of variance between cases compared to variance within cases. Noting the similarity to the partitioning of variance for the latent state-trait model, the ICC can be interpreted as the proportion of consistent vs. occasion-specific variance. Using the same formula, the ICC can be computed for a model with the growth factor included, which can be called a *conditional intraclass correlation coefficient*. The latter can be interpreted as the proportion of variance between cases in the intercept (e.g., scores at baseline) relative to the total variance of y_{ti} after taking into account variance due to linear change over time.

Reliability

Another useful index that is conceptually distinguished from ICC but is closely related mathematically is the *reliability coefficient*. Reliability, in this case, does not refer to psychometric properties (although see Kreft 1997 for a discussion of psychometric analyses with multilevel regression), but refers to the stability of the estimate of either the intercept or the slope. With higher reliability of the estimate, the standard error for the estimate will be smaller, indicating more precise population inferences (Raudenbush, 1988). Below is the reliability coefficient formula for the intercept.

$$\rho = \frac{\psi_{00}}{\psi_{00} + \theta_{(tt)} / T} \tag{7.10}$$

In this version of the reliability coefficient (Raudenbush & Bryk, 1986, 2002), where $\mathrm{Var}(\eta_0) = \psi_{00}$ and $\mathrm{Var}(\varepsilon_{ti}) = \theta_{(tt)}$, the only difference from the formula for ICC is that the number of time points, T, is taken into account. In general, reliability will be higher to the extent that there is more variability between cases in the data set, as represented by larger values of ψ_{00}, as compared with variability within cases, as represented by $\theta_{(tt)}$. The equation also suggests that, as the number of time points is increased, the quantity $\theta_{(tt)}/T$ in the denominator decreases, resulting in higher reliability.

A parallel formula for the reliability of the slope is perhaps of more interest for growth curve analysis.

$$\rho_{RB} = \frac{\psi_{11}}{\psi_{11} + \theta_{(tt)} / T} \tag{7.11}$$

The slope factor variance estimate, $\mathrm{Var}(\eta_1) = \psi_{11}$, is used instead of the intercept factor variance estimate. With greater variability of the slope between cases (i.e., larger ψ_{11}), reliability increases. Complementarily, smaller variability within cases and more time points leads to higher reliability of the slope estimate as well. It makes sense that both less within-case variability and more time points would improve reliability of the slope estimate. Smaller measurement residuals, on average, indicate that the observed points are closer to the growth regression line, and, therefore, the slope is a more accurate predictor of observed scores. As more time points are added, we should also expect that regression line to be a more accurate estimate of true change over time.

In practice, the ICC and the reliability coefficients must be computed manually. This is most likely why these quantities are rarely reported for latent growth curve models.[5] The general formula for the reliability coefficients described above is only one possible approach that has been proposed. McArdle and Epstein (1987) proposed an index of reliability that combines the estimates of variability of the intercepts, ψ_{00}, slopes, ψ_{11}, and their covariance, ψ_{01}. It is an occasion-specific index which can be computed using the following formula:

$$\rho_{ME} = \frac{\psi_{00} + 2\lambda_{t1}\psi_{01} + \lambda_{t1}^2\psi_{11}}{\psi_{00} + 2\lambda_{t1}\psi_{01} + \lambda_{t1}^2\psi_{11} + \theta_{(tt)}}$$

The index is specific to the time point where the value of the slope loading, λ_{t1}, for a particular time point must be inserted. The numerator and denominator are the same except that the denominator contains information about the within-case variability of scores, $\theta_{(tt)}$.

Another alternative reliability index, proposed by Willett (1989), is specific to intercepts or slopes but takes into account the spacing of the time metric, based on the rationale that greater variance of the independent variable should increase reliability and power. Instead of dividing the intraindividual variance, $\theta_{(tt)}$, by the number of time points, T, as in Equation (7.11), T is replaced by the sum of squared deviations from the mean time point, $SST = \sum(t - \bar{t})^2$.

$$\rho_W = \frac{\psi_{11}}{\psi_{11} + \theta_{(tt)}/SST} \tag{7.12}$$

Comments

In a simulation, Rast and Hofer (2014) examine power comparing ρ_{ME} and ρ_W, and they concluded that ρ_W, which is based only on slope variability, is a more appropriate indicator of power to detect variability in change than ρ_{ME}. There is less to go on in terms of choosing between ρ_W and ρ_{RB}. Use of *SST* in the denominator gives exponentially increasing weight to the number of waves and the length of interval between waves (e.g., 1 year vs. 2 years) as compared with just using *T*. With equal spacing, researchers typically use arbitrary time scores for loadings of the growth curve factor rather than coding based on the original interval spacing, so it is important to take this into account if using the reliability computation based on *SST*.

The difference between ρ_W and ρ_{RB} is that ρ_{RB} is unconcerned with the actual time interval underlying the occasions of measurement, whereas ρ_W gives heavy weight to the actual time interval. In other words, with ρ_{RB}, three waves are treated the same in terms of reliability whether waves are semiannual, annual, bienniel, or once every 10 years. The ρ_W index, on the other hand, will show dramatically different values for reliability when the actual time interval differs (Rast & Hofer, 2014, figure 1). Underlying the Willett's conceptualization of ρ_W is the implication that larger increments in actual time should afford greater sensitivity for detecting change. Although this conceptualization makes sense in terms of optimal design of the variance of the predictor in a regression analysis, in the context of growth curve analysis, the conceptualization assumes that greater passage of actual time will necessarily lead to greater sensitivity for detecting change. There may be some theoretical contexts where giving greater weight to longer intervals (e.g., increasing body weight with age, early childhood vocabulary acquisition) makes good sense and others where it does not (e.g, mood, consumer confidence).

Model Fit

The degrees of freedom for a linear growth model are computed in the same way as usual for a model with mean structure. The number of freely estimated parameters subtracted from the number of unique elements in the variance–covariance matrix.

$$df = \frac{J(J+1)}{2} + J - q$$

The number of observed variables is J, which will be equal to T in the basic model, and q is the number of parameters to be estimated. For a latent growth curve model, the typical model would estimate the variance for the intercept factor, ψ_{00}, the variance for the slope factor, ψ_{11}, the covariance between the intercept factor and slope factor, ψ_{01}, the mean of the intercept factor, α_0, the mean of the slope factor, α_1, and for each of the observed variables, J. The measurement intercepts are usually constrained to be equal to 0 to identify the factor means, so this involves no additional parameters. Thus, for this typical specification, the number of free parameters estimated will be equal to 5.

Because a latent growth curve model with observed variables will always have $J = T$, we can restate the number of free parameters, q, as $T + 5$ (Bollen & Curran, 2006). This implies that there must be at least three time points if all of the above stated parameters are to be estimated. With only two time points, there are $T + 5 = 7$ parameters estimated but only $J(J+1)/2 + 2 = 2(2+1)/2 + 2 = 5$ available. For two time points, then, constraints must be placed on some of the parameters. The necessity of such a constraint was seen with the contrast coding repeated measures ANOVA model for two time points discussed in Chapter 3, the measurement residual variances were constrained to 0 to identify the model.

Measurement residuals are usually freely estimated at each time point, allowing for *heterogeneity of variance*. By default, multilevel regression programs nearly universally assume that variances at each time point are equal. This *homogeneity of variance* assumption can be lifted in most multilevel software programs, however, given a sufficient number of variance–covariance elements to free parameters.[6] By default, measurement residual variances in SEM software programs will estimate each measurement residual variance individually, so that heterogeneity of variances is specified by default. A test of the homogeneity of variance is possible for either approach (Willett & Sayer, 1994). Within the multilevel regression framework, there are several possible approaches, where the Bartlett test is the most commonly employed (Bartlett & Kendall, 1946; Raudenbush & Bryk, 2002; see Mendes,& Özcaya Turhan, & Gürbüz, 2006 for a review). With latent growth curve models, the assumption can be tested using a likelihood ratio test that compares the chi-square of the model allowing measurement residual variances to differ from a model that constrains them to be equal.

The fit of a latent growth curve model is determined by the discrepancy between the implied covariance matrix and the obtained covariance matrix. In the context of a linear latent growth curve model, the lack of fit is a function of the average deviation of observed values from the linear slope as illustrated in Figure 7.1. Variance of the measurement residuals in this context is due to several factors (Bollen, 2007; Wu, West, & Taylor, 2009), including random measurement error in the observed variable, *occasion-specific systematic variance*, *occasion-specific nonsystematic variance*, and the correctness of the functional form (i.e., linear in the present model). Occasion-specific systematic variance is explainable by variables omitted from the model, whereas occasion-specific nonsystematic variance is stochastic or not explainable by other factors. Because poor fit may be due to any of these factors, the fit of the model should not be routinely taken as an indication of the appropriateness of the specified functional form. Nor should model fit be viewed as a gauge of the magnitude of the slope. Imagine, for instance, a flat slope indicating no linear change with occasion-specific residuals that are very close to the predicted line. Such a model would fit quite well but would not indicate any change over time.

Although chi-square will reflect lack of fit for the model, with larger sample size, it may tend to reject models with small discrepancies from perfect fit. Researchers commonly rely on alternative fit indices, such as the SRMR, the RMSEA, the CFI, or the WRMR (for binary an ordinal data), but these indices may not detect fit equally well for latent growth curve models. The SRMR appears to have insufficient power for identifying growth models with the incorrect functional form compared with the TLI, CFI, and RMSEA using standard cutoffs when there were few time points (e.g., five; Leite & Stapleton, 2011; Wu & West, 2010; Yu, 2002). With more time points, the SRMR may have better power for rejection, however (Yu, 2002). The WRMR tended to reject too many models with eight and more time points. Leite and Stapleton recommend specific LR tests for detecting incorrect functional form instead of global fit indices. This point can be broadened, and I would suggest that researchers should investigate specific sources of model misfit individually by investigating alternative function forms, tests of homogeneity of variance, and exploring omitted variables. Wu and colleagues (2009) review some additional issues related to fit of growth curve models.

Correlated Measurement Residuals

There may be theoretical or empirical reasons for taking into account residual autocorrelation over time when estimating latent growth curve models. With a single observed variable at each time point, correlated measurement residuals are typically not specified, however. In general, there are $J-1$ additional parameters required to estimate a lag 1 autocorrelation structure and $J(J-1)$ additional parameters required to estimate correlations among all possible measurement residuals. With three time points, there are insufficient elements available in the variance–covariance matrix for estimating either of these structures, for instance, at least given the standard specifications as described for the model in Figure 7.3.

Inclusion of an autoregression correlation structure will generally have little impact on the estimates of the average intercept or average slope, because the estimate of the mean at baseline (and any other chosen referent point for the intercept) or level changes over time are based on observed means rather than covariances. Variance estimates for the intercept and slope factors and their covariance may be affected by inclusion of correlated measurement residuals, however. In particular, because the estimates of the variances of the factors and their covariances are a function of the covariances among the observed variables, any estimated covariance between measurement residuals of any two observed variables will absorb some of the covariance between the growth factors.

As one simple illustration, the observed covariance between two variables in Figure 7.3, say y_2 and y_3, can be decomposed into the variance of the intercept factor, ψ_{00}, the variance of the slope factor, ψ_{11}, the covariance between the intercept and slope factors, ψ_{01}, and the measurement residual covariance, $\text{Cov}(y_1,y_2)=\theta_{12}$.

$$\text{Cov}\left(y_1,y_2\right)=\psi_{00}+\lambda_{11}\lambda_{21}\psi_{11}+\lambda_{11}\psi_{01}+\lambda_{21}\psi_{01}+\text{Cov}\left(y_1,y_2\right) \tag{7.13}$$

The equation follows the traditional path-analytic decomposition of the observed covariance in which the covariances are a function of the product of the compound paths connecting the two variables (e.g., Bollen, 1989). Because each loading for the intercept factor, η_0, is set to 1, these loadings are dropped from the equation. Loadings λ_{21} and λ_{31} connect y_2 and y_3 through the slope factor. An additional path, the covariance between the measurement residuals, θ_{23}, is not depicted in Figure 7.3 but, if added, would be an additional source of covariance. When the measurement residual covariance value is different

from 0, the other estimated values (variances or covariances) must be modified given the same value for $\text{Cov}(y_2,y_3)$. If the correlated measurement residual is positive, an initially negative covariance between the intercept and slope factor will be moved in the positive direction and an initially positive covariance between intercept and slope will be moved in the negative direction.

Example 7.1. Growth Curve Model with Observed Variables

A latent growth curve model as depicted in Figure 7.3 was estimated to investigate changes in body weight over 12 years (six waves of data separated by two years each) using the health and aging data set ($N = 5{,}335$). Body weight was measured by the body mass index (BMI), which is a ratio of weight to square of height (kg/m^2). Syntax and data sets used in the examples are available at the website for the book. The model set each intercept factor loading equal to 1 and the slope factor loadings equal to 0, 1, 2, 3, 4, and 5. Correlated measurement residuals were not included in the initial model. The model fit the data well according to the relative fit indices, with $\chi^2(16) = 623.877$, CFI = .990, SRMR = .031, RMSEA = .084. The mean of the intercept factor was 27.211, which is nearly identical to the observed mean of the same for the first wave (27.176). Although this value is significant, the test merely indicates the value is greater than zero, so its significance is usually a trivial matter. The mean of approximately 27 suggests that, at the beginning of the study, the average respondent was in the overweight category. The mean of the slope factor was .150, $p < .001$, indicating that there was a significant increase of approximately .15 points on the BMI score every two years.

An alternative coding scheme for time could have been used for the loadings, with 0, 2, 4, 6, 8, and 10. These values would have produced a slope half the magnitude, indicating a .150/2 = .075 increase in BMI per year. The standardized estimate of the slope mean was .337, suggesting a moderate-sized effect on average and an approximate increase of .3 standard deviation units on BMI per standard deviation increase in time. The variance of the intercept, 23.377, and the slope, .197, were both significant, $p < .001$, showing significant between-person variance of the initial BMI score and the slope. The latter result indicates that some individuals increase at greater or lesser rates over time. Figure 7.5 is a plot of predicted slopes from the model for a random sample of 20 cases.[7] The figure shows variability in the intercepts as well as slopes. A heavier type line is drawn for the average slope and suggests a slight linear increase in BMI over time. The average intercept and slope values in this plot differ slightly from the predicted values from the equation because of the particular sample of cases. The covariance between the intercept and slope factors, –.057, was nonsignificant, with the standardized value (correlation) equal to –.026. Although not different from what would be expected by chance, the slight negative correlation would suggest that higher baseline BMI tended to be associated with less increase in BMI over time.

To illustrate the impact on model parameters, a second model replicated the first but added autocorrelations among adjacent time points of measurement residuals (ε_1 with ε_2, ε_2 with ε_3, etc.). This model had a considerably better fit, with $\chi^2(11) = 187.542$, CFI = .997, SRMR = .031, RMSEA = .055. The estimates of the average intercept and slope were virtually unchanged, 27.214 and .151, but there were substantial changes in the estimates of variances and covariances. The intercept variance was 22.783, $p < .001$, the slope variance was .110, $p < .001$, and the covariance between the intercept and slope factors was .150, $p < .001$. The covariance, which was not significant in the first model, indicated a positive correlation (.095) between BMI at baseline and changes in BMI over time after

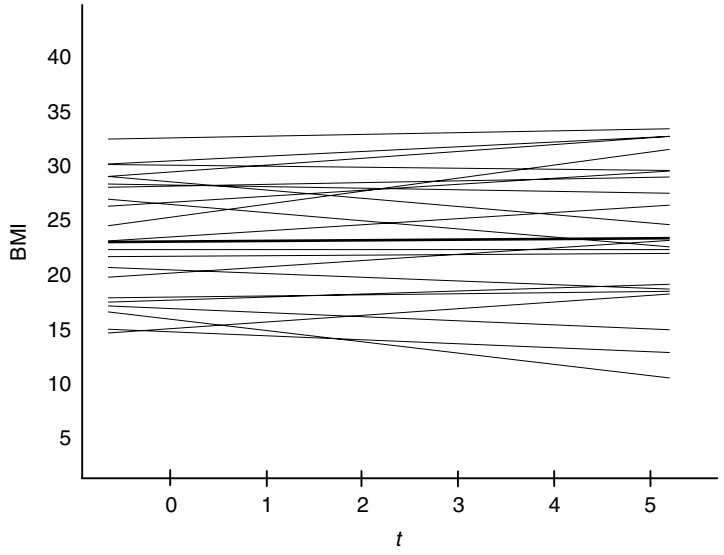

Figure 7.5 Sample of 20 Predicted Growth Curves for Change in BMI Over 10 years.

incorporating correlated measurement residuals. This positive covariance between intercept and slope factors suggested that those who weighed more initially gained significantly more weight over time.

Homogeneity of variance was tested in a subsequent model. Correlated measurement residuals were removed so that this model could be compared to the initial model depicted in Figure 7.3 that allowed for heterogeneous variance. The model constraining measurement residual variances to be equal over time had a substantially higher chi-square than the first model, with $\chi^2(21) = 969.387$, CFI = .984, SRMR = .026, RMSEA = .092, although this model still had an acceptable fit according to several of the alternative fit indices. The difference in chi-square from the model with heterogeneous variance was significant and of moderate magnitude, with $\Delta\chi^2(5) = 345.510$, $p < .001$, $w = .347$, Δ Mc= .151. The results from this comparison and from the autocorrelation model suggest that the linear model may not be appropriate or that there are important omitted variables. Further investigation of the trajectories for this variable will be conducted in the next chapter on nonlinear growth curve models.

Comments

There are several features of the latent growth curve model that distinguish it from other longitudinal analysis methods. First, compared to trend analysis with repeated measures ANOVA, the growth curve model provides additional information. Not only do growth curve models provide information about average increase or decrease on the level of a variable over time, they also provide information about individual variation in changes. This is important for identifying the cases that are more likely to increase or decrease or change at different rates. Second, although the growth curve model is an extension of difference scores derived from two time points, change estimates based on three or more time points provide greater precision in estimating individual change than difference scores. Concerns about imprecision due to unreliability of difference scores or fallibility of individual scores become increasingly remote with more time points. Third, even

considering only the unconditional latent growth curve models, there are advantages over the multilevel growth curve model in the ability to estimate a variety of error structures and assess these specifications with nested tests.

Latent Growth Curves with Binary and Ordinal Observed Variables

Growth curve models of binary or ordinal variables are also possible. The individual growth curve from Equation (7.4) is altered to be a logit or probit regression, depending on the estimation approach (see Chapter 1 for a review of these concepts). Taking the simple case of a binary outcome and the logit model, the predicted growth curve for an individual case is a logistic regression model.

$$\ln\left[\frac{P(y_{ti}=1)}{P(y_{ti}=0)}\right] = \eta_{0i} + \lambda_{t1}\eta_{1i}$$

It may seem odd, however, to contemplate individual growth over a series of 0s and 1s, so it is often easier to imagine an increasing or decreasing propensity for an individual's observed score to be equal to 1 given what we know about the other cases in the data set for any given time point. More generally, then, we can consider the individual linearized growth curve in terms of the underlying continuous variable y^*.

$$y^*_{ti} = \eta_{0i} + \lambda_{t1}\eta_{1i} + \varepsilon_{ti}$$

Growth can then be interpreted as an increase in the underlying continuous y^* for every unit increment in time.

With the y^* conceptualization, the logistic regression with binary variables is easily generalized to any observed variables with two or more ordinal values. Latent growth models with ordinal variables differ from latent growth continuous variables in that a scaling assumption is needed to map the observed discrete variable onto the underlying continuous variable, a step not necessary with continuous variables. With binary variables, the y^* metric is usually established by setting the single threshold equal to 0. To appropriately gauge the level increase or decrease for ordinal variables with more than two categories, thresholds should be constrained in some way to establish a common metric for y^* across time. One approach is to set the first threshold for each variable to be equal to 0 and set all other thresholds equal over time whenever there are three or more ordinal categories. Without equivalent thresholds, level changes in the outcome may be as a result of changing metrics of the dependent variable.[8] An initial step, of course, can be used to test whether the equal threshold assumption is viable (see Chapter 2 for more detail). With two ordinal categories, only a single threshold, usually assumed 0, is needed, and such tests of threshold invariance are not possible.

The identification constraints necessary for the distribution of y^* (either delta or theta parameterization), also implicitly assume homogeneity of variance. With delta parameterization all measurement residuals, $\theta_{tt} = \mathrm{Var}(\varepsilon_t)$, are assumed to be equal to 1. Even with theta parameterization, because the variance of θ_{tt} requires the standardization assumption about the y^* distribution, equality constraints would be problematic for correct interpretation of the factor variances (Muthén & Asparouhov, 2002) and would not be appropriate. Thus, there is no feasible direct test of homogeneity of variance assumption for single indicator ordinal variables. Autocorrelated measurement residuals are possible to include in the ordinal growth models, assuming sufficient identification conditions.

For binary variables using full ML for binary variables (logistic model), the log odds of the threshold can be represented by the difference between the threshold for the first time

point and the intercept factor mean. Conceptually, at the individual case level, the probability that $y_{ti} = 1$ is an exponential function of the case-level intercept.

$$P(y_{ti} = 1) = \frac{1}{1 + e^{\alpha_{0i} + \lambda_{t1}\alpha_{1i}}}$$

For WLSMV estimation, the normal cdf transformation is used to find the probability that $y_{ti} = 1$. It is not a simple matter, however, to extend the logistic or normal cdf transformation in order to estimate the expected proportions that $y_t = 1$ at a certain time point or the average expected increase in the probability that $y_t = 1$ over time, because, when data are binary, taking the average of proportional increase as measured by each individual slope does not equal the proportional increase of the average slope (Masyn, Petras, & Liu, 2014). One can only derive these estimates after obtaining the factor score values for the slope and intercept, deriving the proportional increase for each individual case, then averaging the proportions across cases. This method also is needed for plotting individual slopes for the binary or ordinal case.

Although the intraclass correlation coefficient and reliability can be conceptualized in a similar manner to the continuous case, the constraints on the measurement residual variance requires a modification to $\theta_{(tt)}$ in Equation (7.9) or Equation (7.10) to be equal to 1 for delta parameterization or equal to $\theta_{tt} = \Delta^{-2} - \psi_{00} - \lambda_{t1}^2\psi_{11} - 2\lambda_{t1}\psi_{01}$, where Δ is the scaling factor with all loadings from η_0 assumed equal to 1. In practice, it will be simplest to constrain all θ_{tt} to be equal except the first, which is usually set as an identifying constraint, to obtain an estimate of $\mathrm{Var}(\varepsilon_t) = \theta_{(tt)}$ if $\theta_{tt} \neq \theta_{11}$.

Example 7.2: Growth Curves with Binary and Ordinal Variables

As an example of a growth curve using binary variables, I again used the health and aging data set with six waves of biennial assessments. The first model was a linear growth curve (as depicted in Figure 7.3) of self-report of whether the respondent had diabetes or not. The model was estimated with WLSMV and delta parameterization. To identify the model and obtain an estimate of the intercept factor mean, the measurement intercept threshold for the first indicator was set to 0, a scaling constraint of 1 was placed on the y^* distribution for the first indicator, the remaining thresholds for Times 2–6 were set equal, and the intercept factor mean was estimated. The chi-square for this model was significant and the WRMR was higher than recommended, but the CFI and RMSEA suggested the model had a good fit to the data, with weighted $\chi^2(10) = 117.288$, $p < .001$, CFI = .998, RMSEA = .045, WRMR = 1.520. The intercept factor mean estimate was −1.336. As an informal check, I used the normal cdf function of this value. The result was .091, a value that equaled the obtained proportion of the sample reporting diabetes diagnosis at baseline (9.1%). The mean of the slope factor was .081, which was significant, $p < .001$, indicating an increase in the proportion reporting diabetes every two years. Variance of the intercept and slope were both significant, .917, $p < .001$, and .027, $p < .001$, respectively, suggesting the baseline proportions and rate of change in proportions varied across respondents. The reliability of the slope can be computed using the slope variance, the assumed within-person variance of 1 for delta parameterization, and $T = 6$, $\rho_1 = \psi_{11} / (\psi_{11} + 1/6) = .027 / (.027 + 1/6) = .139$. This value suggests a relatively low reliability of the slope coefficient, a value that ranges from 0 for no reliability to 1 for perfect reliability.

To demonstrate a latent growth model for ordinal variables, the BMI index was used to construct four categories according to international conventional cutoffs, underweight (<18.5), normal weight (≥18.5 to < 25.0), overweight (≥25.0 to <30.0), and obese (≥30.0).

The first threshold value for each observation was set equal to 0 and the remaining thresholds were held equal over time to ensure equal metrics. This model used WLSMV with theta parameterization, although delta parameterization or full ML would be acceptable too. The CFI indicated the model fit well, but the WRMR was above the recommended value of 1.0, $\chi^2(26) = 203.148$, CFI = 1.000, WRMR = 1.774. The average intercept was equal to 7.155, $p < .001$ and the average slope was equal to .107, $p < .001$. The unstandardized slope represents an increase in the propensity on the ordinal scale categories across time, with a y^* increase of .107 on the standard normal scale for each two-year interval. The average standardized slope was .284 suggesting a moderate-sized increment. This value is somewhat smaller than the value obtained in Example 7.1 using the continuous BMI measure. The intercept and slope variances were both significant, 16.550, $p < .001$, and .143, $p < .001$, suggesting variability in the initial values of the ordinal weight variable and slopes across respondents.

Time-Invariant Covariates and Cross-Level Interactions

The models discussed thus far are sometimes referred to as unconditional growth curve models, usually representing an initial step to describe changes over time. It is often of interest, however, to attempt to explain between-case variation in trajectories. What accounts for the fact that some individuals may increase weight more rapidly than others? Or why are some individuals able to maintain consistent weight over time, whereas others gain? Inclusion of predictors measured at one time point, or *time-invariant covariates*, are frequently included to account for variation of scores at baseline (or other intercept referent point) and growth. Unless there are strong theoretical reasons, the intercept factor and the slope factor are usually both regressed on the time-invariant covariates. Figure 7.6 illustrates inclusion of a single covariate, x_1. Any number of covariates, observed or latent variables, can be included in the model, however. Although the inclusion of a time-invariant covariate represents a fairly simple extension of the unconditional latent growth model in some ways, there are a number of issues that require consideration.

We can modify the previously unconditional Equations (7.5) and (7.6) to include a predictor.

$$\eta_{0i} = \alpha_0 + \beta_{01} x_{1i} + \zeta_{0i}$$
$$\eta_{1i} = \alpha_1 + \beta_{11} x_{1i} + \zeta_{1i}$$

Note that x_{1i} only has a subscript i for an individual and does not include a subscript t for occasion, as it is only measured once, and β_{01} and β_{11} are constants for the sample. The insertion of these equations into Equation (7.4), shows that prediction of x_{1i} in the level-2 equation implies a *cross-level interaction* in the single multilevel equation,

$$y_{ti} = \left(\alpha_0 + \beta_{01} x_{1i} + \zeta_{0i}\right) + \lambda_{t1}\left(\alpha_1 + \beta_{11} x_{1i} + \zeta_{1i}\right) + \varepsilon_{ti}$$
$$y_{ti} = \alpha_0 + \alpha_1 \lambda_{t1} + \beta_{01} x_{1i} + \beta_{11} \lambda_{t1} x_{1i} + \zeta_{0i} + \zeta_{1i} \lambda_{t1} + \varepsilon_{ti}$$

with the term $\lambda_{ti} x_{1i}$ representing a product of the time score, λ_{ti}, with the independent variable, x_{1i}, and β_{11} representing the coefficient for the interaction. In conceptual terms, this cross-level interaction signifies that the change in y_{ti} over time depends on the independent variable. For example, if age is a significant predictor of the slope in a growth curve model of BMI, then older and younger respondents change at different rates. A significant effect might indicate that age is associated with different rates of increase in weight over

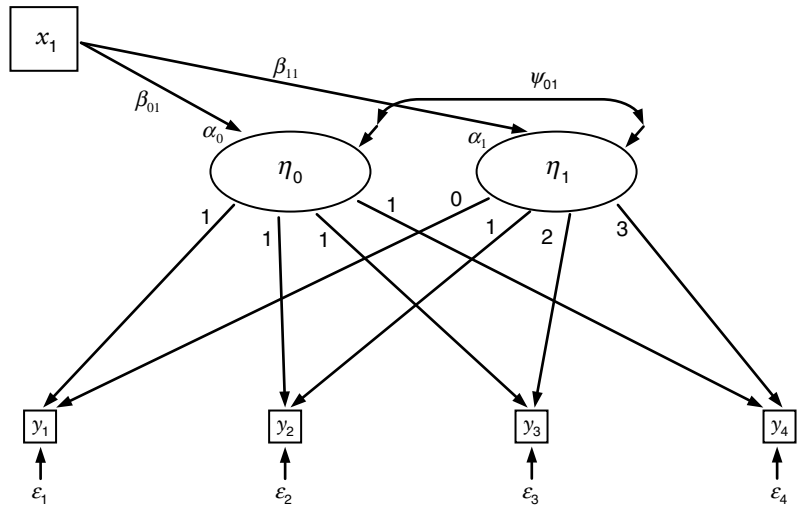

Figure 7.6 Latent Growth Curve Model with Time-Invariant Covariate.

time, different rates of decrease in weight over time, or a mix of increases and decreases in weight over time.

Predictor Variable Scaling

As when testing interactions in regression analysis, the scaling of the independent variables may need to be considered. Although the interaction coefficient (i.e., the highest-order term) or its significance is not impacted by multicollinearity that occurs when including the product variable and the main effect variables in the equation, the standard errors for the main effects will tend to be inflated (Aiken & West, 1991). Centering the predictors (i.e., creating deviation scores by subtracting the mean) will tend to reduce this non-essential multicollinearity. If the predictor variable is latent, centering the indicators is one method of centering the factor mean.

There is another important reason for considering the scaling of the independent variable in a latent growth curve model, and that is the interpretation of the intercept factor of the growth curve. Although slope estimates are usually not impacted by linear transformations of the predictor, the intercept is affected. In many cases, the interpretation of the intercept when the predictor is equal to 0 may not be meaningful whether there is an interaction involved or not. Consider the health and aging weight example. If age is left in its original scaling with the average age at baseline equal to 50, the intercept for BMI can be interpreted as the expected value of BMI at the beginning of the study for anyone of the age 0 (i.e., at birth). In some instances, such an interpretation may be plausible or desirable, but the interpretation of the intercept should never be made without careful consideration of the meaning of the zero-point of the independent variables. In fact, it should be kept in mind that, with several predictors included in a model, the growth curve intercept is the expected value of *y* when all predictors are equal to zero. Without rescaling the predictors, an unreasonable zero-point will commonly occur. Even considering a wide range of variables, from income to population to

positive affect, centering the covariate is likely to be more appropriate as a default than use of the original scaling.

Probing Cross-Level Interactions

It will frequently be of interest to understand the nature of a significant interaction by using plots and simple slopes tests when covariates are used to predict slopes in a latent growth curve model. In general, a simple slope describes the relationship between a predictor and the outcome at particular values of a moderator. In the context of a growth curve model, a simple slope describes the relationship between the time variable and the outcome at particular values of the predictor.[9] In other words, we estimate a simple growth curve to describe change over time for certain values of the covariate. Theoretically chosen values may be used for the covariate, but often the arbitrary points of one standard deviation below the mean, the mean, and one standard deviation above the mean are used. For any predictor variable x, centered at its mean, these values would be $-sd_x$, 0, and sd_x. If the covariate is a latent variable with centered indicators and a mean of 0, the variance of the factor can be used to derive a standard deviation estimate, with $-\sqrt{\psi_{kk}}$, 0, and $\sqrt{\psi_{kk}}$ serving as the particular values of the covariate.

Curran, Bauer, and Willoughby (2004) refer to a simple slope of the predictor variable, x_1, on the slope factor, η_1, as a *simple trajectory*. The simple trajectory can be described in the conditional mean of the latent slope variable, adjusting for the effect of the covariate evaluated at a particular value of the covariate (e.g., $-sd_x$, 0, or sd_x).

$$\alpha_{1|x_1} = \alpha_1 - \beta_{11}x_1 \tag{7.14}$$

In this conditional equation, a particular value of the covariate, x_1, is used to obtain the expected simple trajectory, $\alpha_{1|x1}$, using the overall conditional average of the slope, α_1, and the regression coefficient for the slope regressed on the covariate, β_{11}. The i subscript has been dropped from x_1 in this and subsequent equations but should be understood. The conditional average of the slope $\alpha_{1|x1}$ is the simple trajectory when the covariate is equal to a chosen value.

Simple intercepts can also be calculated, giving the expected value of the intercept (e.g., baseline mean) for particular values of the covariate. Although the simple intercepts may not be of much interest in many studies, their values are needed to plot simple trajectories. The formula for the simple intercept is directly parallel to Equation (7.14), $\alpha_{0|x_1} = \alpha_0 - \beta_{01}x_{1i}$. Here, α_0 is the intercept for the full sample, and β_{01} is the regression coefficient for the intercept factor regressed on the covariate. Once the simple intercepts are obtained, simple trajectories can then be plotted with predicted values as a function of the time scores, λ_{t1}, multiplied by the simple trajectory, $\alpha_{1|x_1}$. Plot points are obtained for each simple trajectory line by calculating the predicted values for each value of λ_{t1}, using the regression equation for the particular simple trajectory, $\hat{y}_{t|x_1} = \alpha_{0|x_1} + \alpha_{1|x_1}\lambda_{t1}$ for each value of x_1 (e.g., $-1\ sd_x$, 0, and $+1\ sd_x$).

In addition to plotting the simple trajectories, each simple trajectory can be tested for significance to determine if it is significantly different from 0. The test is a simple Wald ratio of the coefficient to a standard error estimate.

$$\frac{\alpha_{1|x_1}}{SE_{\alpha_1|x_1}} \tag{7.15}$$

Computation of the standard error for the simple trajectory requires values from the asymptotic variance–covariance matrix of parameter estimates, which is the estimation of the extent to which the slopes vary and covary across the sampling distribution. Note that these values may not be available in some software programs and are usually not included in the default output for others. The standard error for the simple trajectory uses the variance errors of α_1 and β_{11} and the estimated sampling covariance between the two parameters (Curran et al., 2004):

$$SE_{\alpha_1|x_i} = \sqrt{\widehat{\mathrm{Var}(\alpha_1)} + 2x_1\widehat{\mathrm{Cov}(\alpha_1, \beta_{11})} + x_1^2\widehat{\mathrm{Var}(\beta_{11})}} \tag{7.16}$$

Once obtained, the Wald ratio is then evaluated against a standard z-curve for significance (i.e., > 1.96 for .05 significance level). The caret symbol is used to denote the estimated values from the asymptotic variance–covariance matrix of parameter estimates.

Plotting and significance testing of simple trajectories is just one of several ways of exploring the nature of a cross-level interaction. Aside from potentially investigating simple slopes for the predictor at particular values of time, one can compute Johnson–Neyman significance regions or confidence bands for conditional intercepts or slopes (Bollen & Curran, 2006; Curran et al., 2004).

Investigating Group Differences in Trajectories

Researchers are often interested in group differences in trajectories, with the desire to describe which cases increase over time and which cases decrease. There are several ways to approach such hypotheses with latent growth curve models that will be discussed in this section: a MIMIC model, a multigroup model, and trajectory grouping analyses. The MIMIC model and the multigroup model both investigate differences in change among two or more known groups, whereas trajectory grouping analyses derive groups from various patterns of change found in the data.

The MIMIC Approach

We have already explored one such way of explaining individual differences in change, by including a time-invariant covariate to predict slopes. A special case of this analysis involves a binary or categorical covariate represented by $G-1$ dummy variables. Such an analysis is akin to the MIMIC (multiple-indicator multiple-cause) modeling approach discussed in Chapter 3 for mixed factorial ANOVA. In the MIMIC approach to the mixed factorial ANOVA, the path between the binary predictor and the factor representing the repeated measures difference is a mixed between and within interaction. A significant effect of the predictor on the difference factor represents an interaction indicating that the change in the dependent variable is of different magnitude in the two groups represented by the binary predictor. The interaction can be modeled in several ways, using the contrast coding model, the latent difference score model, or the single occasion constraint approach. Including a categorical covariate in the latent growth curve model then represents a type of generalization of the mixed factorial ANOVA, because the effect of the binary variable on the slope factor provides information about the difference in magnitude or direction of change in the two groups.

This special case of latent growth model with a time-invariant covariate yields several meaningful parameters. The effect of the predictor (path β_{11} in Figure 7.6) on the slope factor represents the difference in slopes in the two groups, $\alpha_{1g_1} - \alpha_{1g_0}$, where the

subscripts g_1 and g_0 represent the two groups on x_1 ($x_1 = 0$ and $x_1 = 1$). The two means would represent the average change in each group and are the same as the simple trajectories in Equation (7.14). The intercept for the regression of η_1 on x_1 (shown as α_1 in Figure 7.6) is a conditional mean that represents the average linear slope in the 0 group, α_{1g_0}. The test of significance of the α_1 from the model is the appropriate test of whether the simple trajectory for this group differs from 0, which was shown in Equation (7.15). So, one of the simple trajectories is already tested within the specification of the model. The other simple trajectory, α_{1g_1}, can be tested by reversing the codes for the groups. Be aware that the specific values chosen matter in this case. Coding x_1 as 1 and 2, for example, would produce an estimate of the simple trajectory for a nonsensical value, a 0 group that does not exist.

The Multigroup Approach

A second approach to investigating group differences in growth curve parameters is the multigroup modeling approach. Within this framework, any of the parameters, including means and variances can be compared among groups. This approach represents a fairly straightforward extension of the basic latent growth curve model (Figure 7.3) specified for each of the G groups and then compared via nested tests constraining desired parameters. The multigroup approach offers a few advantages in that variances can be compared among the groups to answer questions about whether variability in baseline values or variability in slopes differs in the two groups, a type of question that cannot be investigated with the MIMIC model. Inclusion of covariates within each of the group-specific models would allow for testing of three-way interactions. That is, assuming a binary predictor, a multigroup comparison of the coefficient β_{11} for the model shown in Figure 7.6 would be the same as a $2 \times 2 \times T$ interaction, where the last factor T is the linear trend. For example, do native language (English vs. Spanish), gender (female vs. male), and changes in perceived economic security interact?

The multigroup analysis of the latent growth curve model also has flexibility in estimation of the error structure. Allowing measurement residuals to differ across groups models heterogeneity of variance, whereas the MIMIC approach must assume homogeneity of variance. This flexibility is useful for exploring whether the functional form of the trajectory is the same across multiple groups.

Exploratory Approach

The third approach to investigating group differences in growth involves deriving a classification of cases from their patterns of change. It may be of interest, for example, to classify individuals into groups who increase, decrease, remain low, or remain high over time on perceived economic security. Once classified, it is then possible to examine the sociodemographic or other characteristics of such groups. The approach allows for subtler and more qualitative distinctions among individual patterns of change that go beyond prediction of increases or decreases. One method of classifying trajectories, sometimes referred to as a pattern-mixture or person-centered approaches (Nagin, 2005), is to save factor scores for the intercepts and slopes, and then these values can be used in graphical explorations, cluster analysis (Dumenci & Windle, 2001), discriminant function analysis, or other comparative analyses classifying trajectories or to learn about associated characteristics. Growth mixture models are discussed in Chapter 10.

Example 7.3: Growth Curve with Time-Invariant Covariate

A conditional growth curve model of BMI was tested using age of the participant at baseline as a time-invariant covariate. This model builds on the unconditional model in Example 7.1. Age was centered by subtracting the sample mean (55.975) from each individual score. Although the chi-square for the model was significant, $\chi^2(20) = 641.026$, $p < .001$, the alternative fit indices suggested a good fitting model, CFI = .989, SRMR = .029, RMSEA = .076. Age was not significantly related to BMI at baseline, $\beta_{01} = -.011$, ns, but age was significantly related to the slope factor, $\beta_{11} = -.015$, $p < .001$, indicating a cross-level interaction. The standardized coefficient suggested a fairly modest effect, $\beta_{11}^{*} = -.141$. The conditional mean for the intercept factor was 27.211, indicating the beginning BMI score for those of average age. The conditional mean for the slope factor indicated an increment of approximately .15 in the BMI score per wave for those of average age, $\alpha_1 = .150$, $\alpha_1{}^{*} = .337$, $p < .001$. Because the unconditional model showed an overall increase in BMI, the negative coefficient for the cross-level interaction suggests that older age at baseline is related to less rapid increases in BMI over time.

The standard deviation for age was 4.095, and this value was used to derive three points on the covariate used for plotting and simple trajectory tests. Plugging in the values –4.095, 0, and 4.095, along with the conditional mean value and covariate path estimate from above, yielded three simple trajectory values:

$$\alpha_1 - x_{1i}\beta_{11} = \alpha_{1|x_1}$$
$$.150 - (-4.095)(-.015) = .211$$
$$.150 - (0)(-.015) = .150$$
$$.150 - (4.095)(-.015) = .089$$

The simple trajectories indicate that for one standard deviation below the mean of age, there is a linear increase of .211 points on the BMI per wave. For average age, the increase is .150, the same as the overall average for the sample. For one standard deviation above the mean of age, there was a smaller linear increase of .089 points on the BMI per wave. The parameter variance and covariance estimates were obtained from the output, and using Equation (7.16), the Wald ratio was then computed for each simple trajectory and indicated that each was significantly greater than 0. For the low, average, and high values of age, the ratios were .211/.010 = 20.000, $p < .001$, .15/.007 = 20.05, $p < .001$, and .089/.010 = 8.373, $p < .001$. Although each of the simple trajectories was significant, the significant cross-level interaction implies that they were of differing magnitudes.

To illustrate a latent growth curve model with a binary covariate, age was artificially dichotomized at age 65 (50–64 vs. ≥65). I first used the MIMIC approach with intercept and slope factors regressed on the binary age variable, a method identical to the model just described with continuous age. The binary age variable did not significantly predict the intercept factor, $\beta_{01} = -.568$, ns, suggesting there was no age difference in BMI at baseline. This nonsignificant difference represents a slightly lower average BMI for the over 65 group at baseline. Age category was a significant predictor of the slope factor, however, suggesting a cross-level interaction, $\beta_{11} = -.157$, $p < .05$. The conditional mean (intercept from covariate-slope factor regression) for the slope was significant, $\alpha_1 = .152$, $p < .001$. Because age categories were 0 and 1, this conditional mean also represents the simple trajectory for those under age 65 (i.e., the 0 group), $\alpha_{1|<65}$. To obtain the other simple trajectory value, I switched the codes for the age category variable so that 0 represented ≥65 and 1 represented <65. The conditional mean from this model was $\alpha_{1|\leq 65} = -.005$, which

was nonsignficant. The results suggest that BMI linearly increases for those beginning the study under age 65 but does not significantly increase for those 65 years of age or older.

The same hypothesis of age differences in trajectories was tested using a multigroup approach, comparing those under 65 to those aged 65 or older. The model fit when the slope factor means were allowed to be estimated differently in the two groups was $\chi^2(34) = 633.350$, $p < .001$, CFI = .990, SRMR = .047, RMSEA = .081.[10] When the conditional means for the slope factor were constrained to be equal in the two groups, the $\chi^2(35) = 639.474$, $p < .001$, which was a significant increase in the chi-square, $\Delta\chi^2(1) = 6.124$, $p < .001$. The difference indicates that the BMI slope for the under 65 group, .152, $p < .001$, and the BMI slope for the 65 and over group, –.009, ns, were significantly different from one another. These coefficient estimates were quite similar to the conditional means estimated with the MIMIC model approach. The intercept factor means, intercept factor variances, slope factor variances, or intercept-slope covariances could be compared across groups in a similar fashion.

Time-Varying Covariates

An alternative way to include additional variables in a latent growth curve model is by the use of *time-varying covariates*, variables that have been measured at each time point. The basic model set up is depicted in Figure 7.7. The primary value of including a time-invariant covariate is to explain variation in the intercept or slope factor, whereas the primary value in including a time-varying covariate is to remove occasion-specific variance from the dependent variable. The covariate may reduce the error in the prediction of the slope factor by accounting for some of the measurement residual variance at each time point. In other words, the deviations of the observed points from the individual slope shown in Figure 7.1 are reduced to the extent that some of the misfit of the slope is due to extraneous factors accounted for by the covariate.

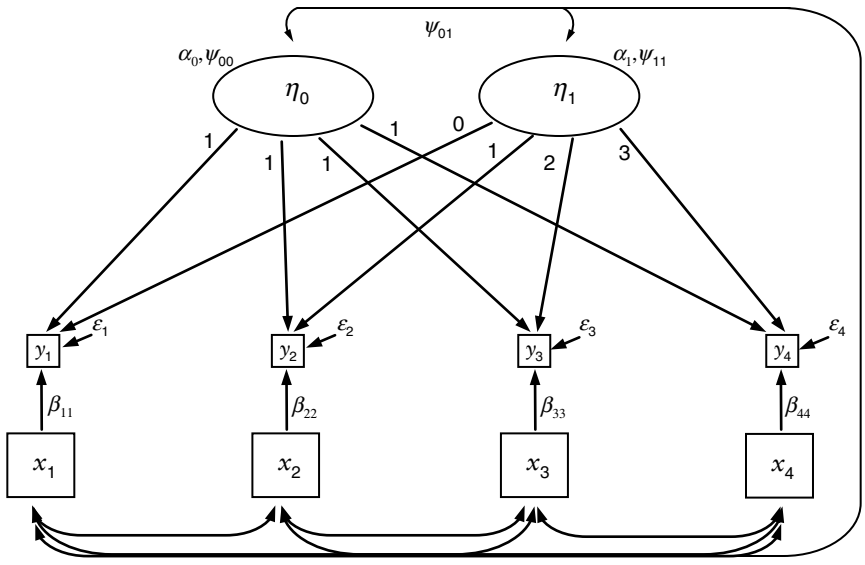

Figure 7.7 Latent Growth Curve Model with Time-Varying Covariate.

By the addition of a time-varying covariate, we modify the level-1 (within-case) equation adding the predictor variable, x_{ti} measured at each wave.

$$y_{ti} = \alpha_0 + \lambda_{t1}\alpha_1 + \lambda_{t1}\zeta_{1i} + \beta_{tti}x_{ti} + \zeta_{0i} + \varepsilon_{ti}$$

The subscript for synchronous path β_{tti} has a double t subscript to represent y_t predicted by x_t at each time point, a value that may vary across cases in the data set. The model assumes each x_{ti} is independent of ε_{ti} as in any regression model. It should also be mentioned that modeling the independent effects of x_{ti}, η_{0i}, and η_{ti} implies estimating the covariances among these exogenous variables as shown in Figure 7.7.

Equations (7.5) and (7.6) can be substituted into this new level-1 equation to form a single multilevel equation:

$$y_{ti} = \alpha_0 + \lambda_{t1}\alpha_1 + \beta_{tti}x_{ti} + \lambda_{t1}\zeta_{1i} + \zeta_{0i} + \varepsilon_{ti}$$

The single equation helps emphasize that scores on the dependent variable at each time point, y_{ti}, are a function of the time-varying covariates as well as the intercept and slope factors (the first three terms on the right) plus multilevel error (the last three terms on the right). For values of β_{tti} greater than 0, the values of other estimates in this equation may be altered compared with the unconditional growth curve model that includes no covariates.

Predictor Variable Scaling

Because the average slope in the unconditional latent growth curve model can be shown to be a function of observed means at each time point (refer to Equation [7.8]), the inclusion of the time-varying covariate implies that the average slope is a function of the adjusted means. This modification of the interpretation of the slope suggests a need to be mindful of the scaling of the covariate. Recall that the intercept in any regression is a function of the mean of the predictor, $\beta_0 = \bar{y} - \beta_1\bar{x}$, and is therefore an adjusted mean. In many instances, it may not be sensible to consider the growth when the covariate is equal to 0. For example, a covariate measured on a Likert-type scale with the lowest value of 1 would lead to an expected mean on y_t when the covariate is equal to a value that is outside its possible range. Even if the covariate has a meaningful zero value, this may not be a desirable referent point for the adjusted means of y. If the covariate is self-rated health, with possible responses that range from 0 for "poor" to 4 for "excellent", poor health may not be a desirable referent point for evaluating growth parameters.

The solution may appear to be to center the time-varying covariate, similarly to the solution to the scaling issue with the time-invariant covariate. With this rescaling, the score for each case at each time point would be subtracted from the mean of all cases for each time point, $x^*_{ti} = x_{ti} - \bar{x}_t$. The mean of the centered scores, $\bar{x}^*_t$, would then be equal to 0 for all time points, and any change in the level of the covariate over time would not be taken into account in the estimation of the growth parameters. If we are interested in modeling change in marital conflict over time controlling for perceived stress, centering stress around the mean of stress at each time point would produce estimates of average marital conflict level of stress at that time point. If the two are positively related and both increase over time, the adjusted average of marital conflict would rise as stress rises, which is the pattern we would expect if we did not control for stress. Even in the case when the covariate is binary, it may not be desirable to interpret the adjusted mean of the dependent variable in terms of the group coded 0 on the covariate. The growth model parameters would represent changes in the dependent variable for cases that are in the 0 group at each wave.

Another possible solution is to center all values around the mean of the covariate at the first time point (or other desired referent point), thereby evaluating the adjusted mean

of the dependent variable in terms of an average score on the covariate while retaining information about the relative changes in the covariate over time. This would entail subtracting the covariate score at baseline from each individual's score for the covariate at each time point, so that the centered covariate is computed as $x^*_{ti} = x_{ti} - \bar{x}_1$. The average distance from the score at the first time point as well as the covariance among the scores is retained with this method. If stress is centered around the level at the first time point, then the average of marital conflict at each time point would be adjusted for the level of stress at the beginning of the study. If both variables increase over time, marital conflict at later time points would be adjusted downward toward the level of conflict experienced when stress was lower at the beginning of the study. This type of deviation score is a variant on *grand mean centering* commonly employed in multilevel regression analysis, in which a score at each time point is subtracted from the grand mean taken from all scores and all time points, $x^*_{ti} = x_{ti} - \bar{x}$. In the usual grand mean centering, however, the grand mean, $\bar{x}$, is the mean taken over all time points and all cases, which may not be a desirable point of reference in the longitudinal context (for time-varying covariates).

Yet another alternative is to center the covariate scores around the mean of scores across all time points for an individual, $x^*_{ti} = x_{ti} - \bar{x}_i$, which is referred to as *group-mean centering* or centering within context. If the mean for each case (i.e., averaging across the time points within each case), is used, the interpretation of the intercept in reference to the first time point is altered, because the centering of the covariate is no longer in reference to the first time point. It is possible to modify the group-mean centering approach, where $x^*_{ti} = x_{ti} - x_{1i} - \bar{x}_1$, which would center within context while retaining reference to the covariate score at the first time point. For example, level marital conflict at each time point would be adjusted for the individual's level of stress at the first time point (as opposed to the sample's average level of stress at the first time point in the grand-mean centering approach). Either of the centering within context approaches may have more substantial differences from the uncentered or the grand-mean centered approaches, because they represent a more complex transformation of the data and alter the covariance structure of the uncentered analysis. Each of the centering approaches will produce different interpretations of the adjusted means, and none are necessarily incorrect. The interpretation can be incorrect, however, given the method of centering that used, so great care is needed to ensure the proper interpretation (see Enders & Tofighi, 2007 and Algina & Swaminathan, 2011, for in-depth discussions of centering).

Although the means are conditional on the covariate at each time point, taking into account the covariate in estimation of growth parameters, the change over time in the covariate that is accounted for is the occasion-specific variation that is not solely due to increasing or decreasing trends in the covariate. We can contrast the interpretation of a growth curve model with a time-varying covariate to a model with parallel growth processes, in which intercept and slope factors are specified for the covariate and these factors are used as predictors of the intercept and slope factors for the dependent variable. The parallel growth model accounts for average change in the covariate over time but not occasion-specific fluctuations that lead to deviations from the linear trend in the covariate. Time-varying covariates consequently function more as a control for all occasion-specific variance that is confounded with the dependent variable rather than removing tends in the covariate in particular.

Interpretation of the Synchronous Path

The synchronous path between the time-varying covariate and the dependent variable, β_{tt}, represents the synchronous relationship between the two variables and not a prediction of change in the dependent variable. The paths may be set equal across time for theoretical

reasons or if there are problems with empirical underidentification, and their equivalence can be tested with a nested model comparison. In some instances, such a constraint may be necessary for the model to be empirically identified. The assumption that the covariate has the same relationship to the dependent variable at each time point may not be valid in every case, however. Some authors also have suggested a lagged effect of the covariate on the dependent variable, in which y_t is regressed on x_{t-1} for all time points except the first time point (e.g., Curran, Muthén, & Harford, 1998; Shaw & Liang, 2012). This variant on the model may include a synchronous effect as well or it may omit it. The inclusion of both the lagged and the synchronous effects may cause estimation difficulties if some equality constraints on the paths are not also employed. Omission of the synchronous path, on the other hand, involves an assumption that the covariate only has a delayed causal effect on the dependent variable and not a simultaneous effect (the simultaneous effect nearly always being the stronger relationship). In neither case is the autoregressive effect of the outcome variable taken into account.

Example 7.4: Time-Varying Covariate

It may be of interest to examine trajectories of weight gain independent of one's overall health at a given moment. Although weight surely influences health, health problems can sometimes have increasing or decreasing effects on weight. I therefore illustrated the use of a time-varying covariate by estimating a latent growth curve model of BMI controlling for the synchronous effects of self-rated health at each time point. The self-rated health variable at each time point was centered around the sample mean of self-rated health at baseline, so that $srh_t^* = srh_{ti} - \overline{srh}_1$. Each of the regression paths was allowed to be freely estimated in the initial model. The fit of the model was acceptable only according to one of the alternative fit indices, $\chi^2(46) = 1085.064$, $p < .001$, CFI = .983, SRMR = .057, RMSEA = .065. The lack of fit may be due to residual autocorrelation (i.e., not taking correlated measurement residuals into account), omitted variables, or incorrect functional form. The effect of self-rated health on BMI was only significant at the first time point, –.111, $p < .001$, with standardized values that ranged from .002 to –.025, suggesting a small effect. The mean of the intercept factor was 27.209, which was close to the value obtained with the unconditional model in Example 7.1. The average slope for BMI after taking self-rated health into account, .151, was also comparable to the value obtained in unconditional model. The intercept factor and slope factor variances were also comparable to the unconditional model, 23.232, $p < .001$ and .197, $p < .001$, respectively.

A model with a longitudinal equality constraint on the self-rated health-BMI path had a significantly poorer fit, indicating that the effect of self-rated health differed over time, $\chi^2(51) = 1{,}104.379$, $p < .001$, CFI = .982, SRMR = .129, RMSEA = .062, $\Delta\chi^2(5) = 19.315$, $p < .001$.

Reconsidering Time

The most common form of the latent growth curve model is applied to panel study that has observations taken at regular intervals. The picture I have presented thus far has been idealized, because I have implicitly assumed that (a) data are complete, (b) each case has equally spaced intervals, and (c) each case starts and ends at the same time relative to the desired time metric. The real world of research is usually not so perfect. Let us take departures from each of the three implicit assumptions in turn to explore some of the complexities of working with data that may have some level of individually varying time points.

Missing Data

Nearly all applied research studies will have some missing data, so that not all individuals may have data at every time point (i.e., the data matrix is not fully balanced). In rough terms, there are two broad patterns of missing data in longitudinal studies, intermittent missing data and attrition. With intermittent missing data, a case may have an omitted observation for one or more occasions, but the pattern is not one in which later waves are missing completely. With an attrition pattern, data for a case are completely missing for one or more waves at the end of the study. The distinction is not completely clear-cut. It may not be easy to decide whether a particular case is missing data due to attrition-related factors. Consider a case with the final time point missing, which would be ambiguous using my distinction. An intermittent pattern of missing data may be more likely to meet the missing at random assumption (MAR: Rubin, 1976) than an attrition pattern, but there is no guarantee of this. Given that data can be assumed to be at least MAR, full information maximum likelihood (FIML) estimation is simple to implement in most SEM software programs and will provide quality estimates. To the extent that data are not MAR, other considerations are necessary. A more extensive treatment of missing data issues related to longitudinal models can be found Chapter 13.

Unequally Spaced Observations

Another deviation from the simple, ideal time metric is when observations are not equally spaced. In practice, this is likely to be true in many data sets, in which interviews, for instance, are not conducted precisely one year apart. Small deviations from time structure of this nature are unlikely to create serious biases, and data are usually treated as time structured.[11] Provided that data are time structured (i.e., consistent across cases) and there is true unequal spacing of observations, there also is no special problem with estimating a latent growth curve. Loadings may be altered to appropriately reflect the difference in spacing. For example, observations from a survey conducted annually for the first three waves yet biennially for three subsequent waves could be assigned time codes 0, 1, 2, 4, 6, and 8. The slope factor mean would then represent change in y_t for each yearly increase, with the annual rate of increase estimated over the last three biennial waves for the shorter interval. In other circumstances, data may be considerably less orderly than this and the length of interval may vary for different cases. Depending on the degree of variation across cases, a combination of missing data estimation and appropriate coding of the slope loadings may be sufficient to handle the problem. If there is considerable variation across cases in the regularity of the intervals, special estimation of these individually varying time points can be used, a method described in greater detail shortly.

Cohorts

The third notable deviation from the ideal time structure concerns different starting and ending times for different individuals. The concept of varying start and end times, in this sense, is generally viewed in terms of an alternative underlying time metric. The simplest case is the *cohort sequential design* (also called accelerated designs). With cohort sequential designs, there are several distinct groups of cases observed over parallel intervals.[12] A common study design follows several age cohorts (e.g., students entering a four-year college course at ages 18, 19, 20, and 21). In such a design, there is variability of age (or other time metric) at Time 1 corresponding to the different cohorts, and therefore, relative to the age of the student the start and end times differ. Depending on the research questions and the data structure, there are several possible approaches to different cohorts.

Time-invariant covariate. One approach to dealing with different cohorts, such as age differences at the beginning of the college study, is simply to control for the cohort in the model. Such a solution would use age as a time-invariant covariate in a model with wave of the study as the time code (e.g., $t = 0$, 1, 2, and 3). In this approach, the interpretation of the intercept and slope are conditioned on the covariate (e.g., age), interpreted as the expected baseline value and rate of change for those with a particular value on the covariate (e.g., average age if centered). Or, in other words, common estimates for the intercept and for the slope are used for the full sample, ignoring variation by age. In the college student example, the researcher would not be concerned with the fact that freshmen start college at different ages, choosing to estimate common values for intercept and slopes across ages. Controlling for age may not be problematic if the desired interpretation is to derive a single estimate to represent all ages.

In this context as in others, it is important to carefully consider the coding of the covariate, bearing in mind that it will often be desirable to use some type of centering. If centering around average age is used, then the intercept will be interpreted as the value of the dependent variable at the first wave (assuming loadings are $t = 0$, 1, 2, … for the slope factor). If no centering is used, then the intercept would have the interpretation as the expected value of the dependent variable at birth. Centering around the earliest possible age, with each observed age at the beginning of the study subtracted from a theoretically chosen age, would give a different interpretation to the intercept. If, in the college student example, age 18 was subtracted from all student ages, the intercept would represent the expected value of the dependent variable at a time three years prior to when they actually entered the study for those students who entered college at age 21. If this centering approach has the goal of treating variation in age as an important factor, then one of the alternative approaches below would likely be preferable.

Multigroup Approach. When cohort differences are of interest, a time-invariant covariate would be problematic if the desired interpretation was to consider cohort differences as meaningful. A multigroup approach specifying a latent growth curve model in each cohort group would allow for comparisons of intercept or slopes across cohorts. Depending on how the slope factor loadings are specified in each of the groups, the researcher can use the waves of the study or the cohort metric as the referent point for the interpretation of the intercept. In the college student example, all cohorts would have the common referent point defined as the first year of college if the same loadings of the slope factor, such as $\lambda_{t1} = 0$, 1, 2, and 3, were used in each group. Alternatively, different loadings in each group could be used to provide a common intercept as the expected value at the earliest age ($\lambda_{t1} = 0$, 1, 2, and 3 for the group beginning college at age 18; $\lambda_{t1} = 1$, 2, 3, and 4, for the group beginning college at age 19; $\lambda_{t1} = 2$, 3, 4, and 5 for the group beginning college at age 20; $\lambda_{t1} = 3$, 4, 5, and 6 for the group beginning college at age 21).[13]

Using the Time Metric. Another approach to modeling growth when there are several cohorts is to use the time metric, such as age, as a basis for the slope factor loadings instead of time codes representing waves of the study. In this approach, the full range of ages is used for slope factor loadings, using missing estimation to accommodate values where cohorts do not overlap. In the college age cohort study, we would use the students age for growth curve loadings, with loadings spanning ages 18 to 25.[14] Freshmen starting at age 19 would all have missing data for age 18, and students entering the study at age 18 would have missing data for ages 23, 24, and 25. The strategy involves some extrapolation beyond the specific ages observed and assumes that missing data can be considered MAR.

Individually Varying Time Points

When there is considerable variability across cases in the start and end times, the data can be said to have individually varying time points. In this circumstance, a special estimation process, called the *variable definition method*, can be used (Neale, Boker, Xie, & Maes., 2002). The method can be viewed as a variant on the approach that sets factor loadings according to the full range of a time metric, such as setting loadings to student ages in the college example. With the variable definition approach, however, special software features allow the user to link the desired time metric to factor loadings for the slope factor by designating artificial time codes on a case-by-case basis. Figure 7.8 provides a conceptual representation of a latent growth curve model with individually varying time points linked to the loading matrix in the variable definition method. The time metric is represented by t_0–t_3 within dashed boxes to represent individually varying values of a time metric such as age for the four occasions. In the college student example, the observations for students entering the study at age 18 would have observations at ages 18 through 21 linked to the factor loadings $\lambda_{t1} = 0, 1, 2, 3$. Students entering the study at age 19 would have observations at ages 19 through 22 linked to factor loadings $\lambda_{t1} = 0, 1, 2, 3$, and so on.

Extrapolating to time points beyond the available data may be of greater consequence under some conditions than others. Consider a slightly more extreme example than the college cohort study. Depicted in Figure 7.9 are hypothetical results for a potential study of business income generated in cities in three states implementing corporate tax breaks. For each state, the average slopes are shown for revenues over a four-year period but spanning 12 years in total. The trajectory that is likely to best fit the full time range (from 2002 through 2011) is of curvilinear form with revenues increasing to a climax around 2006 and then decreasing (dashed line). The advantage of using tax year as the time metric instead of wave is that change over a larger range of years is modeled. The disadvantage is that a theoretical extrapolation must be made across the cohorts and assumes that each state would follow the same course if tracked over the full 12-year period. In other words, cross-sectional differences among cohorts are used to infer about a longitudinal trajectory over the full 12 years. Such a theoretical extrapolation is not necessarily invalid, but it does partially confound cohort and longitudinal or developmental effects. Results could potentially differ from a study in which all of the states were tracked over the full date range.

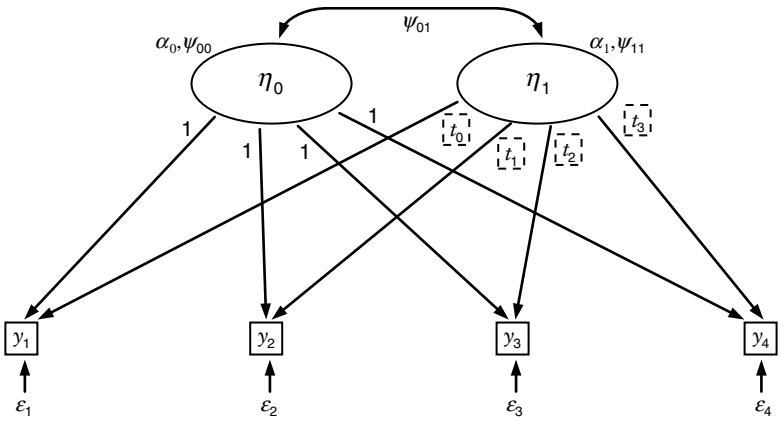

Figure 7.8 Latent Growth Curve Model with Individually Varying Time Scores.

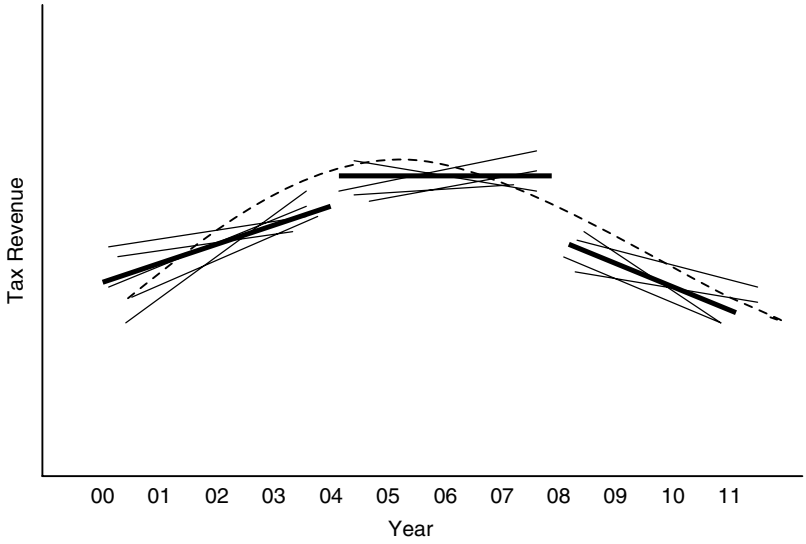

Figure 7.9 Example of Trajectories with Varying Start and End Times: Revenue for Cities Implementing Corporate Tax Breaks During Different Years.

An additional point could be made in connection with this example. For the hypothetical corporate tax study, Figure 7.9 shows no overlap between the periods of observation in the three studies and is therefore a purely hypothetical example. In practice, however, a model with little or no overlap between cohorts may be difficult to estimate because data are extremely sparse. The tax study data would have missing observations for two-thirds of the cities at any given occasion. If more states were involved in the study and observation periods began at different years, there would be fewer missing observations at any occasion and estimation would be more practical. Some aspect of the interpretational complications arising from cross-sectional differences would, of course, still apply to a lesser degree even if there was some overlap between trajectories.

Example 7.5: Some Explorations of Varying Time-Points

Several models were tested to explore some of the options for latent growth models with alternative time metrics. The health and aging data set has participants that range from 50 to 70 at baseline, and models from previous examples have examined growth over six waves (12 years). Instead of modeling growth over the six waves using time scores $\lambda_t = 0, 1, 2, 3, 4, 5$ and ignoring or controlling for age, age could be used as a time metric. Given that the youngest participants start the study at age 50 and the oldest participants start the study at age 70, the potential age range over the study is 50 to 82, given the 12-year span of the study. Accordingly, instead of using waves of the study as time codes, age of each participant could be used. Setting loadings to $\lambda_t = 50, 52, 54, \ldots, 82$, would create an undesirable interpretation for the intercept defined at age 0, so loadings could be set to $\lambda_t = 0, 2, 4, \ldots, 32$ to define the intercept at age 50. Such an analysis requires transformation of the dependent variable using conditional statements, computing a different variable for the baseline BMI for those aged 50 than for those aged 70 (e.g., if age at baseline is equal to 50, then BMI50 = BMI1; and if age at baseline is equal to 70, then BMI70 = BMI1). The result is 33 new variables, in which all participants would have

missing data for approximately two-thirds of the variables. Because this strategy results in too much missing data and in order to simplify the example, I created three age categories as a more convenient way of illustrating alternative time coding.

Three age categories were created, <65, 65 to 70, and >70. New variables were constructed for participants who started the study in one of these three age groups, so that the variables were defined as BMI50, BMI52, BMI54, …. BMI70 to take into account the three different starting age groups and the 12-year span of the study. An unconditional latent growth curve model was tested using this time coding scheme. Specification of the slope loadings corresponding to these variables was $\lambda_t = 0, 2, 4, \ldots, 20$ for the model. Thus, the estimated linear growth covered a 22-year period with two-year increments. The intercept mean was defined as the expected value of BMI for a participant who began the study in the 50 to < 65 age group. Note that only approximately a third of participants had scores for the first time point using this metric, so FIML estimation for missing data was necessary. The alternative fit indices provided somewhat ambiguous information about the adequacy of the fit of the model, $\chi^2(36) = 485.674$, CFI = .986, SRMR = .075, RMSEA = .065. The average intercept was 27.019, indicating the expected weight for individuals if they were to begin the study in the age 50 to < 55 age range. The slope estimate was .083, $p < .05$, telling us that BMI was expected to increase by this amount per year. Note that this value is roughly half of the estimate obtained for the conditional model controlling for age and using biennial waves as the time metric (Example 7.3). The standardized estimate of the average slope was .387, a value quite close to the standardized value obtained when using the per wave time coding approach for the model controlling for age. The variances for the age-metric model differ more substantially from the conditional model, 25.113 for the intercept factor variance and .046 for the slope factor variance.

A variable definition approach was also used. Even though individual ages could be used in this approach, barring estimation difficulties with data that are too sparse, the same time coding scheme was used to facilitate comparison to the previous model. Time score variables were created for each of the three age groups, with starting points staggered by five years and incrementing at two-year intervals. The time scores were as follows for the three groups: $t = 0, 2, 4, 6, 8, 10$ for ages 50–54, $t = 5, 7, 9, 11, 13, 15$ for ages 55–59, and $t = 10, 12, 14, 16, 18, 20$ for ages 60 and over. The full range of time scores was 0, 2, 4, …, 20. Ages could be used directly with this method, as the researcher need only link a time variable with the slope factor loading matrix, but, in many applications, rescaling ages to give the first age a value equal to 0 is likely to be preferred.

Traditional model fit information is not available for this estimation approach. The average intercept estimate was 26.948 and the average slope estimate was .073, $p < .001$. The slope value resembles that obtained with the age scoring method used in the prior model. The intercept and slope variances were significant and also highly comparable to the age coding model, 24.773, $p < .001$, and .048, $p < .001$, respectively.

Comments

Several complicating issues related to the time metric discussed above may exist together in any one study. Missing data will nearly always be present in longitudinal studies, and when missing data varies according to age or another time metric, confounding relationships with missingness may be particularly difficult to disentangle. Although there are clearly some contexts in which age, grade, or another time metric is a preferable basis for interpretation of growth parameters, there are many instances in which it may not be. Bias only exists to the extent that interpretation of the results does not match the data analytic strategy, so no method of coding time is truly incorrect. Choosing the appropriate time

coding for a particular problem and arriving at the appropriate interpretation can be challenging, so neither of these stages should be taken lightly or be concluded without careful consideration. Because the latent growth curve model must utilize the measurement model and factor loading to represent time, it is necessarily more cumbersome than the multilevel regression model approach to growth curves for handling individually varying time scores. Available software features in some programs that allow for individually-varying time scores makes the two general growth model approaches more comparable in this regard, however.

Second-Order Growth Curve Models: Multiple Indicators at each Occasion

The latent growth curve model can be augmented to include multiple indicators at each occasion, models that are variably referred to as *second-order growth curve models* (Sayer & Cumsille, 2001), curve-of-factors models (e.g., McArdle, 1988), multivariate latent curve models (MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997), or multiple-indicator growth curve models (Chan, 1998). Figure 7.10 illustrates the model specification for a second-order factor model of three time points. The loading specifications for the higher-order intercept and slope factors parallel the first-order growth curve specifications exactly. The loadings for the second-order growth factors, η_0 and η_1, are defined as in the first-order latent growth curve model, with the loadings for the intercept factor all set equal to 1 and the loadings for the slope factor set according to a chosen time metric (e.g., $\beta_{t1} = 0, 1, 2, \ldots T-1$). As with the first-order growth curve models, the intercept and slope factor means and variances will be of primary interest in most applications. The intercepts for the regression of the first-order factors on the second-order factors (α_{t1}, α_{t1}, and α_{t1} in Figure 7.10) are generally set equal to 0, so that the means of the first-order factors are defined by the identification constraints on the first-order factors. An assumption that occasion-specific variances are homogeneous over time can be imposed on first-order disturbances, ψ_{tt}, and can be tested using chi-square differences. The scaling choice for

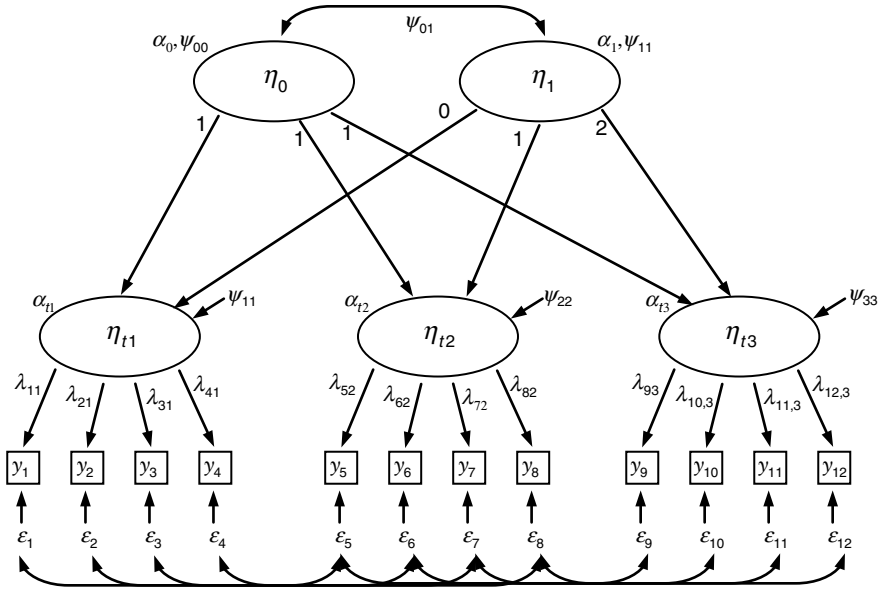

Figure 7.10 Second-Order Latent Growth Curve Model.

each first-order factor then determines how the values or the intercept and slope factors are interpreted (see below). Initial steps to establish measurement invariance are recommended, and appropriate constraints on loadings and measurement intercepts for repeated indicators, at minimum, typically would be employed (Chen, Sousa & West, 2005; Ferrer, Balluerka, & Widaman, 2008).

Consistent with other longitudinal structural models with multiple indicators, correlated measurement residuals would generally be included routinely in the second-order growth curve model. The measurement residual correlation structure might take different forms, depending on the number of available time points, either on an a priori theoretical basis or an empirical basis. Common structures would include correlations among repeated indicators across all time points, a lag 1 structure, or an alternative structure (e.g., first-order simplex). With few time points, there are fewer options for error structures. Nested model tests can be used to compare the fit of these structures provided there are sufficient degrees of freedom available. An alternative is to specify method factors following the strategies outlined for state-trait error models in Chapter 6.

Identification of First-Order Factors

Several issues related to identification of first-order factors require mention. Because the growth parameter estimates are dependent on the scaling of the first-order factors, the constraints used to identify the first-order variances and means are integral to how the intercept and the slope are interpreted (see Chapter 1 for a more detailed discussion of the impact of identification constraints in factor means and variances). The simplest (and often software default) method of identifying the factor mean and factor variance is referent identification (Sayer & Cumsille, 2001). With no longitudinal equality constraints on loading and measurement intercepts, the second-order growth parameters will reflect the values of the observed variable used as the referent so that, with the standard time codes, the intercept mean will give the expected value of the observed mean of the referent indicator and the slope of the referent indicator over time. Consequently, with referent indicator identification, the second-order-growth curve model will closely resemble a first-order growth curve model using only the referent indicator as the dependent variable. The equivalence is not exact, however, as the first-order factor variances will be partially determined by other indicators, and the use of equality constraints on loadings and measurement intercepts complicates the comparison (see discussion of the impact of equality constraints on means and variances in Chapter 2). The upshot is that the second-order factor model does not fully take into account all of the indicators for each first-order factor. In this regard, there is some disadvantage to the second-order growth curve model compared with a first-order-growth curve model that uses a composite of the indicators as a single-indicator at each occasion.

Ferrer and colleagues (2008) point out that the use of referent identification for loading and intercepts may lead to different results, depending on the choice of referent. The referent identification is problematic to the extent that results for the growth parameters could vary in magnitude or direction, as empirical examples demonstrate. Bollen and Curran (2006, p. 255) suggest using the "most reliable and valid indicator," which could be determined through a longitudinal confirmatory factor model. Presumably, the indicator with the largest loading (most reliable) would provide the best single representation of the factor. Once the best indicator is chosen, it would then serve as the referent indicator for the loading and the measurement intercept for the first-order factors, with the same indicator employed at each time point. The indicator with the largest loading at one time point may not be the indicator with the largest

loading at every time point, so the selection of the "best" indicator may be tricky in practice.

Ferrer and colleagues suggest an alternative method based on single occasion identification. The single occasion identification approach is not possible in the second-order growth curve model, because the first-order factors at each occasion are endogenous and would entail setting the residual variance and intercept. Their suggested approach involves two steps. The first step obtains first-order factor variance and mean estimates from a separate longitudinal confirmatory factor analysis, and the second step sets one factor loading as a referent in the growth model to the value obtained from the first step. The means are identified in a different manner by setting the intercept factor mean equal to 0 and allowing the first-order measurement intercepts to be estimated but constrained equal over time. With this approach, the slope estimate uses information from all first-order indicators but has the disadvantage of an intercept that is arbitrarily scaled rather than an intercept that provides information tied to the original scaling of the dependent variable.

Although not possible with available features in all SEM software packages, the effects coding identification approach (Little, Slegers, & Card, 2006) supplies the most ideal scaling solution for first-order factors, because it incorporates information from all indicators. When software does not allow complex equality constraints, one simple alternative would be to set all loadings equal to 1 and all measurement intercepts equal to 0. Requiring all loadings from each factor to be equal to 1, and therefore equal to one another, is likely to have detrimental effects on fit, however. Still, even this convenient solution might be preferable to deriving an estimate of change from only a single indicator, because all indicators contribute to the mean and the variance of the first-order factors and measurement error is taken into account by estimating the measurement residuals for each indicator. Meredith and Horn (2001) propose an alternative that creates factor means that are weighted averages of the observed indicators. Factors means are identified using factor scores obtained from an initial longitudinal confirmatory factor analysis step. Though more complicated to implement, their method suggests an alternative to equal weighting or referent indicator identification methods.

Advantages of the Second-Order Growth Curve Model

Although there are advantages to including multiple indicators at each occasion, the use of multiple indicators in the second-order factor latent growth curve model provides no special advantage over the first-order latent growth curve model in estimating occasion-specific factor means. Recall from Chapter 1 that measurement error does not affect the expected mean of a measure, so removal of measurement error from the factor variance has no impact on the factor mean. And, because intercept and slope factor means are a function of the observed means, as shown in Equation (7.8), the latent growth curve intercept and slope mean estimates also would not be impacted by measurement error.

There are several potential advantages of using multiple indicators in latent growth curve models, however. Consider first that second-order growth curve models should have higher reliability than first-order growth curve models, leading to greater statistical power (von Oertzen, Hertzog, Lindenberger, & Ghisletta, 2010; Wänström, 2009). Because observed variance at each occasion is partitioned into measurement residual variance (error and specific variance) and factor variance (true score variance), $\mathrm{Var}\left(y_j\right) = \lambda_{ij}^2 \psi_{jj} + \mathrm{Var}\left(\varepsilon_{(j)}\right)$ (see Equation [1.2]), the occasion-specific variance (or within-person variance) will be reduced.

In the second-order growth curve model, measurement residual variance, $\mathrm{Var}(\varepsilon_{(j)}) = \theta_{(tt)}$, no longer represents occasion-specific variance. Instead, the disturbance variance associated with the first-order factors, $\psi_{(tt)}$, represents occasion-specific variance. The reduced estimate of occasion-specific variance will increase reliability estimates. For example, consider the ρ_{RB} formula for reliability (e.g., Raudenbush & Bryk, 2002) as a representative example. Replacing the measurement residual variance with the factor disturbance variance, gives

$$\rho_{RB} = \frac{\psi_{11}}{\psi_{11} + \psi_{(tt)} / T}$$

The value used could be a single value derived from a model with equality constraints for the parameter or the average of disturbance variance estimates.

To the extent that measurement residual variance is greater than 0, the estimate of occasion-specific variability using factor disturbances in the second-order model will be smaller than the estimate obtained from the measurement residuals in the first-order model. The end result then is that there will be an improved estimate of the slope reliability, which will have beneficial effects for standard error estimation and power.[15] Conceptually, we can see this by considering Figure 7.1 of the individual growth curve again. For any observation with measurement error, random fluctuation of observed scores from occasion to occasion and, therefore, the deviation of the observed score y_{ti} from predicted value at each occasion, will be larger than for an observation without measurement error. There will, of course, be a similar advantage of the precision of the intercept estimate for the second-order growth curve model compared with the first-order growth curve model, with a corresponding modification to the reliability estimate for the intercept by using the occasion-specific disturbance, $\psi_{(tt)}$.

An additional advantage of the second-order latent growth curve model is the potential for inclusion of autocorrelated measurement residuals to account for stable specific variance. The first-order latent growth model cannot estimate the autocorrelation of specific variance, and would, therefore, result in biased estimates when true autocorrelations exist. Because autocorrelations among measurement residuals concern the covariance structure and not the mean structure, the effect on growth parameter estimates from including autocorrelations among measurement residuals will mostly be seen in the covariance between intercept and slope and the heterogeneity of occasion-specific variance (i.e., factor disturbances). Equation (7.13) can be adapted to illustrate how the covariance among two repeated measures, j and j° over time, $\mathrm{Cov}\left(y_j, y_{j^{\circ}}\right)$, is partly accounted for by the covariance among the associated measurement residuals, $\mathrm{Cov}\left(\varepsilon_j, \varepsilon_{j^{\circ}}\right) = \theta_{jj^{\circ}}$.

$$\mathrm{Cov}\left(y_j, y_{j^{\circ}}\right) = \psi_{00} + \lambda_{j1}\lambda_{j^{\circ}1}\psi_{11} + \lambda_{j1}\psi_{01} + \lambda_{j^{\circ}1}\psi_{01} + \mathrm{Cov}\left(\varepsilon_j, \varepsilon_{j^{\circ}}\right)$$

In the partitioning of the observed covariance into these components, it can be seen that when values of the measurement residual autocovariance differ from 0, it will necessarily lead to adjustments in the factor variance and covariance parameters. When true autocorrelations are greater than 0, the fit of the model will also be improved by including these elements in the model specification.

Inclusion of time-invariant or time-varying covariates is certainly possible with second-order growth curve models. There are few special considerations beyond those discussed in connection with first-order growth curve models. With second-order growth curve models, time-varying covariates will predict the occasion-specific factors rather than

observed variables.[16] Similar extensions to understanding group differences in growth curve models can be applied to the second-order case using either a MIMIC or multigroup modeling approach.

Example 7.6: Second-Order Growth Curve Model

I illustrate the second-order latent growth curve model (Figure 7.10) with the health and aging data set. Three subscale scores (negative affect, positive affect, and somatic symptoms) from the Center for Epidemiologic Studies-Depression scale were used as indicators of an overall depression factor at each time point. The initial model used the referent indicator approach to identifying the first-order factors with the first loading and the first measurement intercept at each time point set equal to 1 and 0, respectively. For this particular model, the remaining loadings and measurement intercepts were not set equal over time in order to illustrate the relationship to the first-order model. Correlated measurement residuals were also omitted from the model for purposes of illustration. Model results from the second-order growth curve model had an intercept factor mean, α_0, equal to .276 and the slope factor mean, α_1, of .009, $p < .001$. The standardized estimate for the slope average was .172, suggesting a small to moderate increase in depression over the six waves. The intercept factor mean closely resembled the observed mean for the referent indicator (negative affect), which was equal to .270. The variance of the intercept was .160, $p < .001$, and the variance of the slope was .003, $p < .001$.

For comparison, a first-order latent growth curve model was tested using only the variable that served as the referent indicator (negative affect) in the second-order model. For this model, the mean of the intercept factor was .276 and the mean of the slope factor was .009, $p < .001$, values that exactly matched the second-order model using all three indicators at each time point. The variance of the intercept, .156, $p < .001$, and slope factor, .003, $p < .001$, were equal or approximately equal to their respective values in the second-order model. Standard errors and significance values from the first- and second-order growth models also corresponded closely. For example, the Wald ratios for the significance test of the slope factor mean was 5.203 for the second-order model and 5.171 for the first-order model.

Finally, to demonstrate a more optimal model specification, a second-order latent growth curve model with effects coding identification of the first-order factors was tested. The effects coding identification approach will produce growth model estimates based on weighted means of the indicators at each occasion, which will be more representative of the first-order factor indicators. The model included equality constraints on factor loadings and measurement intercepts over time as well as correlations among measurement residuals for repeated indicators. Standard time values for the intercept and slope factor loadings were used, with $\lambda_{t1} = 0, 1, 2, 3, 4, 5$. According to the alternative fit indices, this model fit the data well, $\chi^2(111) = 815.215$, CFI = .981, SRMR = .047, RMSEA = .034. The intercept factor mean was .349, and the slope factor mean was .006, $p < .001$, indicating an increase in depression across the six waves of the study. The standardized estimate of the slope factor mean was .119, which suggested a small increase in depression over time. The variances for the intercept and slope factors were .130 and .002, both significant at $p < .001$. The reliability for the slope, which was computed with the ρ_{RB} formula using the average disturbance variance for convenience, was $\psi_{11} / \left(\psi_{11} + \psi_{tt} / T\right) = .002 / \left(.002 + .089 / 6\right) = .119$.

Comments

Applications of the first-order latent variable growth curve model are ubiquitous, but applications of the second-order latent growth curve model are scarce. Given the several potential advantages of the second-order growth curve model and the frequent availability of multiple indicators for a construct, there are few reasons why the second-order growth curve model should not be applied more often. Attention to identification of the first-order factors is critical, however, as the perfunctory use of the referent indicator approach eliminates some of the potential advantages for improved validity in representing the construct of interest. The fact that results and conclusions could potentially differ depending on the choice of referent (Ferrer et al., 2008) is troublesome. Results reported in Example 7.6 were based on a large sample size, and the advantages of the second-order factor model for statistical precision may be more evident with smaller samples sizes or fewer time points. I turn next to more general issues of statistical precision and power.

Power, Number of Time Points, and Sample size Issues

It will be useful to provide a few brief comments about power, the number of time points, and sample size with latent growth curve models, though there is not sufficient space for a complete treatment of this topic. Minimum sample size guidelines for latent growth curve models should follow general recommendations for SEM, often stated in terms of a minimum of 100 cases for normally distributed continuous variables or 5–10 cases per measured variable (e.g., Anderson & Gerbing, 1988; Finney & DiStefano, 2013; Hu & Bentler, 1999; Tanaka, 1987). These are oversimplifications, however, and convergence and power depend on model complexity, multivariate distributions, missing data, and estimation method (e.g., ML for continuous variables, WLSMV, bootstrapping). At minimum, three time points are required for a standard linear latent growth curve model with no special constraints, but simulation work by Fan and Fan (2005) found convergence problems in up to a quarter of the samples for a latent growth curve with only three time points. For five or more time points, there were no convergence problems even for as few as 50 or 100 cases.

Beyond these minimal requirement recommendations, there are some additional guidelines that may be relevant to growth curves in particular. When considering statistical power for growth curve models, it is imperative to distinguish between tests of significance of fixed effects (average intercepts and slopes) and random effects (variances of intercepts and slopes). Fewer cases and time points are required for tests of fixed effects than for random effects. Based on simulation work on multilevel regression models (Maas & Hox, 2004, 2005), there will typically be unbiased standard errors and sufficient power to test fixed effects (average intercept and slope estimates) with a minimum of 50 groups with 5 cases per group. For growth curves, this would suggest a minimum of 50 cases with five time points. Fewer time points would generally require more cases. Results from the study by Fan and Fan (2005) on latent growth curve models were fairly consistent with this, suggesting a minimum of 50–75 cases were needed with a medium effect size and 100 cases for a smaller effect size, even for as few as three time points.

Sufficient power for tests of variance components appears to require more cases. Intercept and slope variances are not only important for describing growth but also for predicting initial status or changes with covariates. Because predicting initial status is usually of less interest and requires fewer cases for sufficient power, most of the focus of simulation work has been on slope variance. Muthén and Curran (1997) estimated that approximately 500 cases with five time points would be required for sufficient power to

detect a small effect size for slope variance. Although Hertzog and colleagues (Hertzog, von Oertzen, Ghisletta, & Lindenberger, 2008) reported far fewer cases would be needed for sufficient power (only approximately 100 cases with six occasions or 500 cases with fewer occasions), their results are based on a simultaneous test of slope variance and intercept variances, which makes it difficult to gauge the relevance for slopes specifically.

Rast and Hofer (2014) point out that there is considerable variability in power as a function of the ratio of slope variance to error variance, based on an analysis of 35 longitudinal studies. Their results are for single-parameter variance tests that provide information specific to the slope test. Using a small effect size from the range of those observed in the studies and single-unit increments, their simulation results suggest that over 3,000 cases would be needed for three occasions, approximately, 1,800 cases would be needed for four occasions, and approximately 750 cases would be needed for five occasions. These results are fairly consistent with those of Muthén and Curran for small effect sizes. With larger intervals between occasions, the minimum number of cases required decreases, and, in general, more waves can have important effects on power. None of the simulations report the sample size required for sufficient power with moderate effect sizes for the slope variance. Moderate effect sizes, which are realistically attainable goals for applied studies, should require considerably fewer cases, but more extensive simulation work is needed to better understand the required sample sizes for various effect sizes.

Several simulation studies examined power for detecting the effects of a covariate (cross-level interactions) or the correlation between slopes for two simultaneously estimated growth curves (Fan, 2003; Hertzog, Lindenberger, Ghisletta, & Oertzen, 2006; Mathieu, Aguinis, Culpepper, & Chen, 2012; Muthén & Curran, 1997; Wänström, 2009). All of these studies suggest that more cases are required than tests of slope variance to achieve sufficient power. The nature of the effect of the covariate may be important, as Fan (2003) illustrates that disordinal interactions (with crossed slopes) for tests of the effect of a binary covariate on slopes are particularly low in power, requiring as many as 300 cases to detect a medium effect size as opposed to 150 cases for detecting a cross-level interaction of a different form.

Wänström (2009) and von Oertzen and colleagues (2010) investigated power with second-order latent growth curve models. Both studies suggest advantages of multiple indicators. Wänström's findings suggest that more indicators and higher loadings (i.e., higher measurement reliability) increase power to detect the effects of covariates on growth factors. The sample size required to detect the effect of a covariate decreases by nearly 50% for factors with high loadings compared to factors with low loadings. Results suggest that the second-order factor model has advantages in precision of estimates of variance effects of both intercepts and slopes, reducing the minimum required sample size, and increasing sensitivity for detecting smaller effect sizes.

In summary, simulation work provides some useful general guidance for the number of cases and time points required for linear latent growth curve analysis. Fixed effects require far fewer cases and time points, with three or four time points and 100 cases a potential minimum for detecting moderate effects. For random effects and cross-level interactions, the minimum number of cases is likely to be larger, with perhaps several hundred cases needed. The use of multiple indicators at each occasion reduces the sample size necessary and increases sensitivity for detecting smaller effect sizes. These studies also suggest, however, that sufficient power depends heavily on effect size and can vary widely across studies. As a consequence, study planning should not rely solely on this general simulation work but should be based instead on power analysis and careful consideration of expected effect size for the study.

Extensions

This chapter presents a basic introduction to growth curve models, but there are a variety of extensions possible. The most obvious extension is to nonlinear growth, a topic that will be covered in detail in the next chapter. The latent growth curve model can be modified to take into account autoregressive effects of occasion-specific observations, what Bollen and Curran (2004, 2006) term the *autoregressive latent trajectory model*. The goal is to remove autoregressive effects from the estimation of trend over time, potentially eliminating a source of confounding in growth estimates. Such models may have the effect of obscuring nonlinearity, however (Bianconcini, 2012; Voelkle, 2008). This issue is revisited in Chapter 8 and its relation to autoregressive processes and differences in time series models is considered again in Chapter 11.

Because the latent growth curve model is specified within the general structural modeling framework, any expansion of the growth curve model using other elements of structural models is possible. Whereas the multilevel regression approach to growth curves allows for prediction of intercept or slope parameters, the SEM framework allows for these parameters to serve as predictors as well. Theories concerning the effects of decreases or increases over time are ripe for exploration with this approach, and, to date, there have been few studies that have taken advantage of this level of modeling flexibility. For example, declines in physical health in later life may precede onset of dementia symptoms. Investigation of simultaneous growth curves for multiple variables is also not possible with the multilevel regression approach (Willett & Sayer, 1996). These models, referred to as cross-domain or parallel process growth curves examine the correspondence between trajectories of two or more variables. Growth parameters can also be incorporated into mediation models as predictors, mediators, or outcomes (Cheong, MacKinnon, & Khoo, 2003; von Soest & Hagtvet, 2011).

Although not emphasized here, multigroup analysis of growth curves provides a powerful tool for analyzing a classic longitudinal design, the cohort sequential design (Meredith & Tisak, 1990). Multigroup SEM is a highly flexible approach for analyzing these designs, allowing for constraints on means or variances of intercepts or slopes as well as variable error structures.

Recommended Readings

Book-length treatments of latent growth curve models provide more background on this analysis approach than I can provide in the larger context of longitudinal SEM (Bollen & Curran, 2006; Duncan, Duncan, & Stryker, 2013; Preacher, Wichman, MacCallum, & Briggs, 2008; Singer & Willett, 2003). The text by Bollen and Curran is somewhat more technical than the other sources but a thorough consideration of many of the relevant issues. Bollen (2007) provides an excellent discussion of the history of growth curves and latent growth curve models. Chou, Bentler, and Pentz (1998) review the similarities and differences of the latent growth curve and multilevel regression growth curve approaches. Masyn, Petras, and Liu (2014) have an excellent discussion of growth curve models with non-continuous variables. Discussion of cross-level interactions, including plotting and testing simple trajectories can be found in Preacher, Curran, and Bauer (2006), Carrig, Wirth, and Curran (2004), and Curran, Bauer, and Willoughby (2004). Several sources delve into the considerable complexities of the impact of time coding on growth parameter estimates and more detailed discussion of individually varying time points (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004; Hancock & Choi, 2006; Metha & West, 2000; Stoel & van den Wittenboer, 2003). Sayer and Cumsille (2001) give a good general

introduction to specification of second-order growth curves. Finally, Rast and Hofer (2014) extensively consider reliability estimation and power with latent growth curves.

Notes

1 I deviate from traditional LISREL notation by using a 0 subscript for the intercept factor to maintain a parallel to growth curve analysis in multilevel regression and to emphasize the special nature of these growth factors. The usual numbering of paths is also modified to be consistent with these alternative subscripts.

2 Parameter estimates, standard errors, and significance tests obtained with the multilevel regression approach and the SEM approach will be very similar (Chou et al., 1998), given comparable specifications (more on these details below) and continuous variables. Although both derive estimates from a ML estimation process, the algorithms used in the two types of software are related but not identical. The multilevel regression analysis typically uses an expectation maximization (EM) algorithm (Raudenbush & Bryk, 2002), a two-stage "restricted" maximum likelihood estimation process, whereas SEM programs typically employ algorithms based on Fletcher-Powell or Gauss-Newton minimization by default (Bollen, 1989). Although multilevel regression programs usually have a Fisher scoring "full" maximum likelihood estimator available, it is not identical to the ML estimation algorithm used in SEM software.

3 A very serious concern, which is often overlooked, is whether there is sufficient power to detect variance of slopes. Generally, there is greater power to detect variance of intercepts than variances of slopes (Maas & Hox, 2005). Without sufficient power, it is possible to erroneously conclude that the slopes do not differ across cases. This topic is addressed further below in the section Power, Number of Time Points, and Sample Size Issues.

4 This formula is clearly the most commonly used definition of ICC, but there are others that also have been proposed (Bliese, 2000; Kuljanin, Braun, & DeShon, 2011).

5 ICC and reliability are more commonly reported when multilevel regression software is used, because some software programs generate ICC or reliability coefficients in the output (e.g., HLM; Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). For other multilevel software programs and all SEM software programs, these quantities must be manually computed and this usually results in less frequent reporting.

6 The degrees of freedom for growth curve models estimated with multilevel regression differ from the degrees of freedom for latent growth curve models. If the intercept and each level-1 predictor has a random effect and all possible covariances among them are estimated, then, with V as the number of level-1 variables, $\left[(V+1)(V+2)\right]/2+1$ parameters are estimated in a multilevel regression model if homogeneity of variance is assumed (Goldstein, 2011; Snijders & Bosker, 2012). The number of available parameters to be estimated is equal to the number of variance–covariance elements, $V(V+1)/2$, with V equal to the number observed variables. For an unconditional growth curve model, there is only one predictor, the time variable, so the minimum number of covariance elements needed for identification of the model is $\left[(1+1)(1+2)\right]/2+1=4$. Allowing heterogeneity of variance requires additional parameters to be estimated.

7 Plots of predicted growth curve lines are not available in most SEM software programs (Mplus is one exception) and these plots are typically generated using other software (e.g., Carrig, Wirth, & Curran, 2004; Preacher, Curran, & Bauer, 2006).

8 Setting the first threshold equal to 0 is not the only approach to identifying binary or ordinal models. It is also possible to set the intercept mean to 0, where information about the baseline proportions would be contained in the estimated thresholds for the baseline measure (Muthén & Asparouhov, 2002). Yet another alternative to these approaches is to set thresholds to particular theoretical values using prior knowledge of the underlying continuous distribution. Koran and Hancock (2010) argue that this approach will result in more accurate estimates of growth with observed ordinal variables for some measures. Partial threshold invariance has also been suggested as a possible alternative assumption if full threshold invariance cannot be met (e.g., Muthén & Asparouhov, 2002).

9 Another simple effect analysis for the growth curve model is the effect of the predictor on the outcome (e.g., effect of x_1 on y_{t1}) at particular values of time, λ_{t1} (Curran, Bauer, Willoughby, & 2004), but this type of simple slope seems to be rarely reported.

10 For this model, I constrained the intercept and slope factor variances to be equal in the two groups. The constraints were used to obtain a more stable estimate of the variances in the over 65 group, which had a small sample size, and to provide more comparable results to the MIMIC modeling approach. The variance constraints did not significantly degrade the fit of the model, so these appeared to be empirically justifiable assumptions.

11 Where the deviation of the actual time from the assumed time point is normally distributed and centered on the time point assumed (e.g., actual interview dates that vary around the assumed anniversary date), there is likely some imprecision but there should be minimal bias. It may be wise, however, to check whether any given data set at least approximately conforms to this pattern. There are practices, such as rounding to a specific date, which could potentially have important biases in some situations. For example, rounding to month of age in an infant study could lead to results that are based on an older development age than the month rounded to week of age. This would have little effect for older ages, but could be important where small units of time matter, as with a few weeks or even days of development with young infants.

12 My terminology is imprecise. Demographers and epidemiologists distinguish between age, periods, and cohorts, although these factors are often confounded and difficult to separate (e.g., Kupper, Janis, Karmous, & Greenberg, 1985). I use the term "cohort" here simply to refer to a time-related grouping.

13 Mehta and West (2000) discuss the implications of using different intercept referent points and recommend considering defining growth factor loadings in a way that derives a common referent intercept with respect to the desired time metric (e.g., age). The approach taken should depend on whether the desired interpretation is according to the cohort time metric or the wave of the study.

14 If the multigroup approach is used and means and variances of the intercept and slope are set equal across groups, the implied time series also spans from age 18 to age 25. Under restricted conditions, the estimates of slope and intercept should be equivalent to the approach that modifies slope loadings according the range of ages.

15 The benefit of multiple-indicator models for precision of mean estimates in the context of ANOVA has been discussed by Hancock (2003). Considering that mean differences are also the basis of the latent growth curve model, we should expect a parallel improvement in precision of estimates of latent growth parameters. Simulation results from Wänström (2009) and von Oertzen and colleagues (von Oertzen, Hertzog, Lindenberger, & Ghisletta, 2010) would appear to confirm this.

16 There may be special measurement-related hypotheses, however, in which paths between time-varying covariates and observed variables at each time point might be included in a second-order growth curve model. For example, differential item functioning or biases associated with population groups or other factors are often modeled in measurement-focused applications, and it is foreseeable that investigation of similar biases could be investigated with a second-order growth curve model if there is particular interest in understanding how these biases affect change over time.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, NJ: Sage.

Algina, J., & Swaminathan, H. (2011). Centering in two-level nested designs. In. J. Hox, & K. Roberts (Eds.), *The handbook of advanced multilevel data analysis* (pp. 285–312). New York: Routledge.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Baker, G. A. (1954). Factor analysis of relative growth. *Growth*, *18*, 137–143.

Bartlett, M. S., & Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, *8*, 128–138.

Bliese, P (2000). Within-group agreement, non-independence, and reliability. In K. Klein & S. Kozlowski (Eds.), *Multi-level theory, research, and methods in organizations* (pp. 349–381). San Francisco: CA: Jossey-Bass.

Bianconcini, S. (2012). Nonlinear and quasi-simplex patterns in latent growth models. *Multivariate Behavioral Research*, *47*(1), 88–114.

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*, 30–52.

Bock, R. D. (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds.), *Modern advances in psychometric research* (pp. 103–115). Hillsdale, NJ: Erlbaum.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A. (2007). On the origins of latent curve models. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100* (pp. 79–98). Mahwah, NJ: Lawrence Erlbaum Associates.

Bollen, K. A., & Curran, P. J. (2004). Autoregressive latent trajectory (ALT) models a synthesis of two traditions. *Sociological Methods & Research*, *32*, 336–383.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158.

Carrig, M. M., Wirth, R. J., & Curran, P. J. (2004). A SAS macro for estimating and visualizing individual growth curves. *Structural Equation Modeling 11*, 132–149.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, *1*, 421–483.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471–492.

Cheong, J., MacKinnon, D. P., & Khoo, S. (2003). Investigation of mediational processes using parallel process latent growth curve modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 238–262.

Chou, C-P, Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*, 247–266.

Curran, P. J., Bauer, D. J., & Willoughby, M. T. (2004). Testing main effects and interactions in latent curve analysis. *Psychological Methods*, *9*, 220.

Curran, P. J., Muthén, B. O., & Harford, T. C. (1998). The influence of changes in marital status on developmental trajectories of alcohol use in young adults. *Journal of Studies on Alcohol and Drugs*, *59*, 647–658.

Dumenci, L., & Windle, M. (2001). Cluster analysis as a method of recovering types of intraindividual growth trajectories: A Monte Carlo study. *Multivariate Behavioral Research*, *36*, 501–522.

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 Academic.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*, 121–138.

Kupper, L. L., Janis, J. M., Karmous, A., & Greenberg, B. G. (1985). Statistical age-period cohort analysis: a review and critique. *Journal of Chronic Diseases*, *38*, 811–830.

Fan, X. (2003). Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. *Structural Equation Modeling*, *10*, 380–400.

Fan, X., & Fan, X. (2005). Power of latent growth modeling for detecting linear growth: Number of measurements and comparison with other analytic approaches. *The Journal of Experimental Education*, *73*, 121–139.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*, 22–36.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller, (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age Publishing.

Goldstein, H. (2011). *Multilevel statistical models*. New York: John Wiley & Sons.

Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods*, *2*, 293–305.

Hancock, G. R., & Choi, J. (2006). A vernacular for linear latent growth models. *Structural Equation Modeling*: *A Multidisciplinary Journal*, *13*, 352–377.

Hertzog, C., Lindenberger, U., Ghisletta, P., & Oertzen, T. V. (2006). On the power of multivariate latent growth curve models to detect correlated change. *Psychological Methods*, *11*, 244–252.

Hertzog, C., von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling*, *15*, 541–563.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Koran, J. & Hancock, G. R. (2010). Using fixed thresholds with grouped data in structural equation modeling, *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 590–604.

Kreft, I. G. G. (1997). The interactive effect of alcohol prevention programs in high school classes: An illustration of item homogeneity scaling and multilevel analysis techniques. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *Science of prevention: Methodological advances from alcohol and substance abuse research*. Washington, DC: American Psychological Association.

Kuljanin, G., Braun, M. T., & DeShon, R. P. (2011). A cautionary note on modeling growth trends in longitudinal data. *Psychological Methods*, *16*, 249–264.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics, 38*, 963–974.

Leite, W. L., & Stapleton, L. M. (2011). Detecting growth shape misspecifications in latent growth models: An evaluation of fit indexes. *The Journal of Experimental Education*, *79*, 361–381.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Maas, C. J., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 86–92.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*, 215–253.

Masyn, K. E., Petras, H., & Liu, W. (2014). Growth curve models with categorical outcomes. In *Encyclopedia of Criminology and Criminal Justice* (pp. 2013–2025). New York: Springer.

Mathieu, J., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Improving the accuracy of inferences about cross-level interaction tests in random coefficient modeling. *Journal of Applied Psychology*, *97*, 951–966.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. Nesselroade & R. B. Cattel (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., p. 561–614). New York: Plenum.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*, 110–133.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*, 23–43.

Mendes, M., Özcaya Turhan, N., & Gürbüz, F. (2006). A new alternative in testing for homogeneity of variances. *Journal of Statistical Research*, *40*, 65–83.

Meredith, W., & Horn, J. L. (2001). The role of factorial invariance in modeling growth and change. In. L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association.

Meredith, W., & Tisak, J. (1984). *Statistical considerations in Tuckerizing curves with emphasis on growth curves and cohort sequential analysis*. Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Muthén, B. & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Retrieved from https://www.statmodel.com/download/webnotes/CatMGLong.pdf.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371–402.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2002). *Mx: Statistical modeling* (6th ed.). Richmond, VA: Author.

Newsom, J. T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 431–447.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrika*, *51*, 83–90.

Rast, P., & Hofer, S. M. (2014). Longitudinal design considerations to optimize power to detect variances and covariances among rates of change: Simulation results based on actual longitudinal studies. *Psychological Methods*, *19*, 133–154.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics*, *13*, 85–116.

Raudenbush, S. W. (1993). Hierarchical linear models and experimental design. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science*. New York: Marcel Dekker.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*. 1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, NJ: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Chicago, IL: Scientific Software International.

Raykov, T. (1998). Satisfying a simplex structure is simpler than it should be: A latent curve analysis revisit. *Multivariate Behavioral Research*, *33*, 343–363.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726–748.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In. L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 179–200). Washington DC: American Psychological Association.

Shaw, B. A., & Liang, J. (2012). Growth models with multilevel regression. In J. T. Newsom, R. N. Jones, & S. M. Hofer, *Longitudinal data analysis: A practical guide for researchers in aging, health and social sciences* (pp. 217–242). New York: Routledge.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Snijders, Tom A. B., and Bosker, Roel J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: Sage Publications.

Stoel, R. D., & van den Wittenboer, G. (2003). Time dependence of growth parameters in latent growth curve models with time invariant covariates. *Methods of Psychological Research*, *8*, 21–41.

Tanaka, J. S. (1987). "How big is big enough?" Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*, 134–146.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, *23*, 19–23.

Voelkle, M. C. (2008). Reconsidering the use of autoregressive latent trajectory (ALT) models. *Multivariate Behavioral Research*, *43*, 564–591.

Von Oertzen, T., Hertzog, C., Lindenberger, U., & Ghisletta, P. (2010). The effect of multiple indicators on the power to detect inter-individual differences in change. *British Journal of Mathematical and Statistical Psychology*, *63*, 627–646.

von Soest, T. & Hagtvet, K.A. (2011) Mediation analysis in a latent growth curve modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 289–314.

Wänström, L. (2009). Sample sizes for two-group second-order latent growth curve models. *Multivariate Behavioral Research*, *44*, 588–619.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies on individual growth. *Educational and Psychological Measurement*, *49*, 587–602.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363–381.

Willett, J. B., & Sayer, A. G. (1996). Cross-domain analyses of change over time: Combining growth modeling and covariance structure analysis. In G. A. Marcoulides & R. E. Schumaker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 125–157). Mahwah, NJ: Erlbaum.

Wu, W., & West, S. G. (2010). Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behavioral Research*, *45*, 420–452.

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*, 183–201.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles.

Keywords
latent growth curve model, multilevel, growth curve, latent curve model

8 Nonlinear Latent Growth Curve Models

The latent growth curve model is highly flexible and can be extended to a variety of complex mathematical patterns of change that go beyond the simple linear slope. Building upon the fundamental principles of linear growth curves detailed in the previous chapter, this chapter will discuss latent growth curve models of discontinuous change, quadratic change, polynomials, and flexible forms. These more complex functions can be studied by making some fairly minor modifications to the common linear latent growth curve model.

Nonlinear change has been of interest to many fields for nearly a century or more. The first efforts to model individual nonlinear growth were in biological sciences. Early interest was in growth of plants, livestock, and children (Fisher, 1921; Jenss & Bayley, 1937; Wishart, 1938). The first uses of factor analysis were devoted to understanding growth and gave rise to the general SEM approach to growth curve models. Early works on latent growth curve models also discussed some methods of modeling nonlinear change (McArdle & Epstein, 1987; Meredith & Tisak, 1990), but many developments have occurred since. Despite their long history, nonlinear models have been vastly underutilized in the social sciences. There may be many reasons for their scarcity, including lack of theoretical specificity, data limitations, imprecise measurement, or insufficient technical knowledge. Advances in theory and design are likely imminent and the following pages seek to provide additional technical tools for investigating nonlinear change.

Nonlinear change can be conceptualized as a non-constant change, an increase or decrease that accelerates or decelerates over time. This generally must imply a more complicated mathematical relationship between the time variable and the predicted outcome. I will try to keep the mathematics to a level as accessible as possible, but we will necessarily delve into some simple calculus to discuss one way of conceptualizing nonlinear rates of change. The chapter starts with a discussion of linear piecewise models that allow for estimation of different slopes for two or more segments of the time series.

Piecewise Models

Discontinuous change is one way in which growth may depart from a simple linear trajectory. When the rate of change across a portion of the time points is different from the rate of change across another portion, the process can be modeled with a *latent piecewise growth curve model*.[1] The primary application is to assess whether the rate of change prior to an event, which may correspond to the introduction of an intervention or other disruption, differs from the rate of change after the event. For example, an economist might be interested in whether the rate of corporate profits changes in a sample of oil and gas companies after the introduction of new environmental regulations. Piecewise models are an alternative approach to the analysis of the regression continuity design (Cook & Campbell, 1979), which has traditionally been analyzed with ANOVA or regression

models. Latent piecewise growth curve models have advantages over traditional analytic approaches because, as random coefficient models, they can be used to compare individual differences in intercepts or slopes before and after an event. In the latent growth curve approach, the intercept and slope factors also can be incorporated into more complex structural equation models in which they are predicted by covariates or serve as predictors of outcome variables.

Single-Intercept Model

The latent piecewise growth curve model extends the linear growth curve model in a fairly simple fashion. In the basic model, two slope factors are specified using loadings that correspond to segments of the study before and after the event. Figure 8.1 shows the most common version of the latent piecewise model with an intercept factor, η_0, and two slope factors, η_1 and η_2. A piecewise model requires five time points at minimum for identification (Bollen & Curran, 2006). In this specification, illustrated with six time points, the loadings for each of the slope factors uses a different set of time codes: $\lambda_{t1} = 0, 1, 2, 3, 3, 3$, and $\lambda_{t2} = 0, 0, 0, 0, 1, 2$. The placement of the last 0 for the second factor's loadings determines the point at which the second slope rate begins. In this case, the second slope rate begins at Wave 4. Notice that adding the elements of the two loading vectors together gives the usual one-unit increments of the linear growth curve model, 0, 1, 2, 3, 4, 5, so the two slope factors can be viewed as a decomposition of the single growth trajectory. The intercept factor sets all loadings equal to 1 as in the basic latent growth curve model.

Figure 8.2a shows the general form of the relationship between time and the predicted value of *y* represented by the model where there is a single intercept defined as the baseline value but different rates of change, one rate for the first three time points and one rate for the last three time points. The pattern of results represented in the figure suggests two different positive slope values, but either slope could be flat or negative instead. The case-level equation for the latent growth curve is extended by adding the second factor and loading vector.

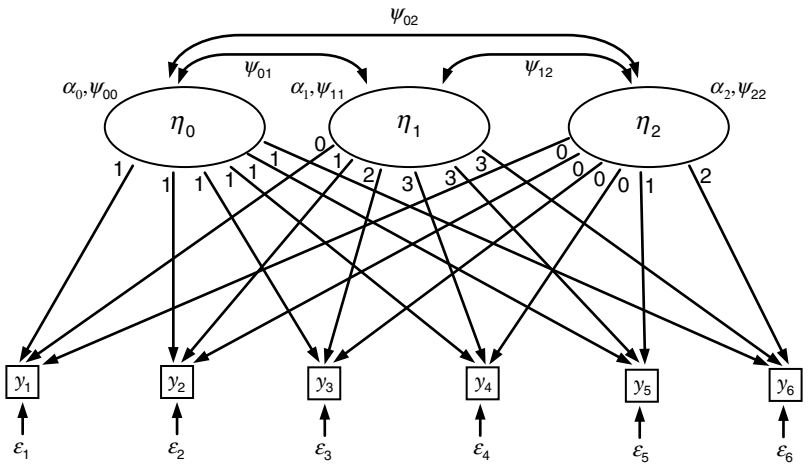

Figure 8.1 Linear Piecewise Model.

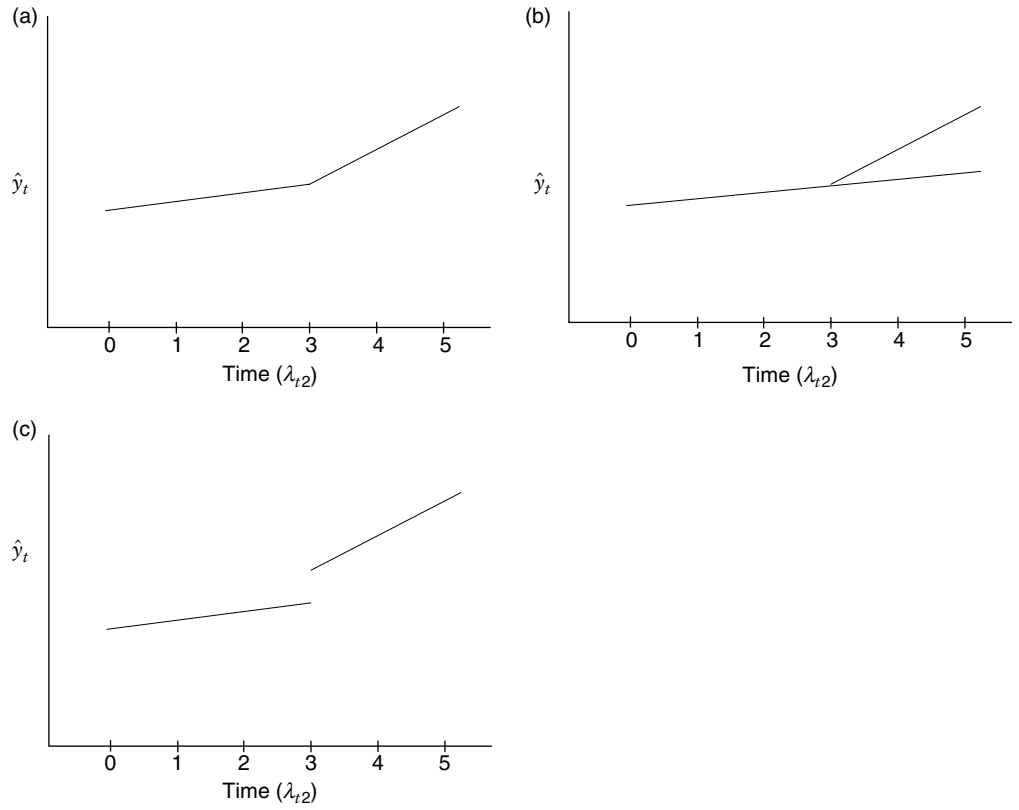

Figure 8.2 Plots Illustrating Piecewise Slope Patterns: (a) differing slopes; (b) increment (decrement) model; (c) differing slopes and intercepts.

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\eta_{1i} + \lambda_{t2}\eta_{2i} + \varepsilon_{ti}$$

The time codes used in Figure 8.1 represent just one example of a variety of coding schemes that could be used. The researcher may wish to define the intercept as the final time point or at the point of intervention, for instance. The slope means and/or the slope variances can be compared for the two segments using a likelihood ratio test by imposing equality constraints on these parameters. The piecewise model is nested within the single trajectory linear growth curve model, so the comparison of the two overall chi-square values also provides an omnibus test of the utility of the piecewise model. Note that the omnibus test implies a test of the covariances among the growth factors in addition to the comparisons of means and variances of each.[2]

Increment (Decrement) Model

An alternative coding scheme also estimates two slopes. The first slope represents a base rate across the full length of the study, whereas the second slope estimates the extent to which there is an increase or decrease relative to the base rate after the event (see Figure 8.2b). This variation has been referred to as the *increment (decrement) model* (Raudenbush & Bryk, 2002) or the added growth model (Flora, 2008). The loadings for the base rate

slope factor follow the codes from the common linear growth curve model, $\lambda_{t1} = 0, 1, 2, 3, 4, 5$, and the loadings for the increment (decrement) factor follow the post-event codes from the piecewise model, $\lambda_{t2} = 0, 0, 0, 1, 2, 3$.

Two-Intercept Model

Piecewise models do not have to be restricted to a single intercept. Theory may predict an immediate level increase after the event, and this may occur with or without a change in the slope (Singer & Willett, 2003). The piecewise model with two slopes can be extended by specifying two intercept factors to obtain separate intercept estimates for the periods before and after the event (Figure 8.2c). Consider again a design with six time points. This model retains the loadings for the slope factors in the original piecewise model, $\lambda_{t1} = 0, 1, 2, 2, 2, 2$ and $\lambda_{t2} = 0, 0, 0, 1, 2, 3$, but two sets of loadings consisting of 1s and 0s are used to specify two intercept factors, $\lambda_{t0\text{pre}} = 1, 1, 1, 0, 0, 0$, and $\lambda_{t0\text{post}} = 0, 0, 0, 1, 1, 1$. The two intercept means, $\alpha_{0\text{pre}}$ and $\alpha_{0\text{post}}$, represent the expected means for the outcome at the first time point and the fourth time point, respectively. This specification allows the researcher to compare the intercept or slope factor means or their variances using equality constraints and likelihood ratio testing. Theory may also predict different intercepts but the same rate of change in the pre- and post-event segments, which could be modeled by allowing the two intercepts to differ but constraining the two slopes to be equal.

Example 8.1: Piecewise Models

A latent piecewise growth model was tested using body mass index (BMI) from the health and aging data set. Syntax and data sets used in the examples are available at the website for the book. The model was specified precisely as the model depicted in Figure 8.1, using a single intercept for this initial model. Just for the purposes of illustration, the transition point was chosen arbitrarily in between the third and fourth waves of data, so that the first slope modeled the rate of change over the first three waves and the second slope modeled the rate of change over the last three time points. Loadings were set equal to $\lambda_{t1} = 0, 1, 2, 3, 3, 3$ and $\lambda_{t2} = 0, 0, 0, 0, 1, 2$, for the slope factors for first three time points and the last three time points, respectively. This model fit the data well according to the alternative fit indices, $\chi^2(12) = 158.173$, $p < .001$, CFI = .998, RMSEA = .048. The mean of the intercept factor was 27.166, which corresponds closely with the observed mean of 27.176 at baseline. The mean of the first slope factor was .192, $p < .001$, indicating an increment of about one-fifth of a point on the BMI every two years over the first three waves of the study. The standardized value for this slope was .343, suggesting a moderate effect. The mean of the second slope factor was .064, $p < .001$, which was about a third of the rate of change seen in the first three waves of the study. The standardized estimate for this slope was .076. The difference in slopes suggests that rate of increase in weight was not constant throughout the study. This hypothesis was tested by constraining the two slope means to be equal, with the results showing the slopes differed significantly, $\chi^2(13) = 195.765$, $\Delta\chi^2\ (1) = 37.592$, $p < .001$. The difference in fit was of small to moderate magnitude, $w = .256$ ΔMc = .028. The small difference in intercepts before and after the change point is not surprising, given that the change point was arbitrarily assigned in this case. Figure 8.3a plots the average slopes.

A subsequent model specified separate intercepts for the two slopes, allowing for a step increase in the expected value at the fourth wave. The mean of the first intercept factor was 27.180, which corresponded closely with the observed mean at baseline. $\chi^2(7) = 20.586$, $p = .004$, CFI = 1.000, RMSEA = .019. The mean of the second intercept factor was 27.254, which resembled the observed mean of 27.757 at the fourth wave. Although the second

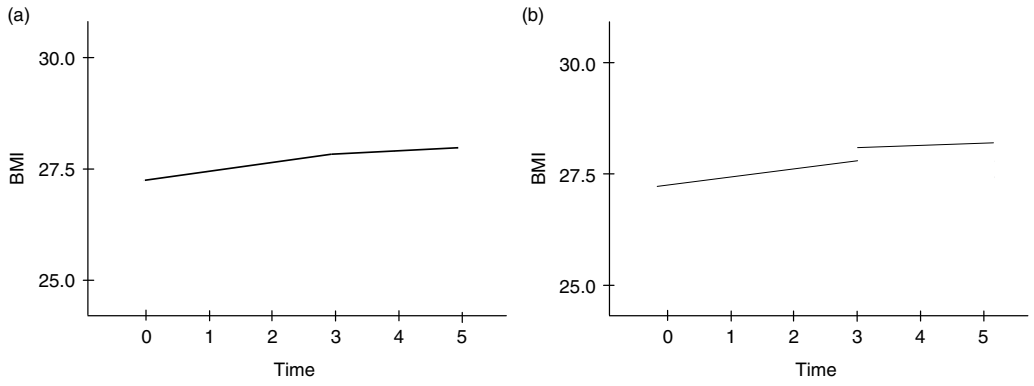

Figure 8.3 Results from Piecewise Models of BMI: (a) single intercept at baseline; (b) two intercepts, before and after change point.

intercept appeared to be only slightly higher than the first intercept, constraining the intercepts to be equal resulted in a significantly poorer fit, $\Delta\chi^2$ (1) = 6.045, $p < .001$, indicating the two intercepts differed statistically. The magnitude of this difference appeared to be small, $w = .103$ $\Delta Mc = .004$, however. Figure 8.3b plots the average slopes.

An increment (decrement) model was tested with loadings set equal to 0, 1, 2, 3, 4, 5 for the first slope factor and 0, 0, 0, 0, 1, 2 for the second slope factor. The mean of the first slope factor was .192, $p < .001$, with a standardized value of .343. The mean of the second slope factor was –.127, $p < .001$, with a standardized value of –.124. The negative coefficient represents a slower rate of increase in body weight by .127 points. Observe that if the two unstandardized slopes are added, their sum is close to the mean of the post-transition point slope in the initial piecewise model, .192 + (–.127) = .065. The significance test of the second slope in the increment (decrement) model, therefore, assesses the same hypothesis as the likelihood ratio test comparing the two slopes in the initial model. In fact, the square of the *z*-test from the increment (decrement) factor mean is nearly identical to the chi-square difference from the first piecewise model, $z^2 = (6.142)^2 = 37.724$ vs. $\Delta\chi^2 = 37.592$.

Extensions

These simple piecewise models can be extended in several ways. Slopes can be specified for three or more segments of a time series. Flora (2008) provides an example of such a model. Methods have also been suggested for determining the location of the change point when one is not predicted a priori (Harring, Cudeck, & du Toit, 2006). One adaptation of piecewise models includes multiple indicators at each occasion and is a fairly natural extension of the model with observed variables (Kohli & Harring, 2013). Nonlinear curves also can be modeled for one or more of the segments, a topic I will return to later in the chapter.

Comments

The latent piecewise model has been underused by researchers (cf. Cohen, 2008), despite a number of advantages for examining change prior to and after an event or intervention. This analysis strategy is well-suited for comparing groups as well, where application

to quasi-experimental studies could yield considerable benefits. Demonstrating treatment and comparison groups have similar rates of change prior to an intervention but significantly different rates of change after the intervention has been introduced would provide strong evidence in favor of the intervention's effectiveness even without randomization (Cook & Campbell, 1979).

Quadratic Growth Curves

Basic Concepts

The simplest continuous nonlinear change is a quadratic function, which can be easily investigated with a few alterations of the linear latent growth curve model. The quadratic latent growth curve model typically estimates two latent variables representing linear and quadratic change. The quadratic factor is specified with loadings equal to the square of the values used for the linear factor loadings. Figure 8.4 depicts the model with linear factor loadings $\lambda_{t1} = 0, 1, 2, 3, 4, 5$ and quadratic factor loadings $\lambda_{t2} = \lambda_{t1}^2 = 0, 1, 4, 9, 16, 25$. The single-case quadratic can be written as

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\eta_{1i} + \lambda_{t1}^2\eta_{2i} + \varepsilon_{ti} \tag{8.1}$$

The mean of the second factor, $E(\eta_{2i}) = \alpha_2$, represents the average quadratic effect over and above the linear effect. In other words, whereas the linear effect represented the constant or average rate of change over the study, the quadratic effect represents the amount of acceleration or deceleration in the rate of change. Another way to state this is that the quadratic effect gives the "rate of change of the rate of change."[3] Equation (8.1) emphasizes this point:

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\left(\eta_{1i} + \lambda_{t1}\eta_{2i}\right) + \varepsilon_{ti}$$

Observe that the term inside the parentheses resembles its own linear model. A positive coefficient for the quadratic effect represents a tendency toward a concave shape (*u*-shaped), whereas a negative coefficient for the quadratic effect represents a tendency toward a convex shape (*n*-shaped). The degree of "concaveness" or "convexness" is determined by the quadratic coefficient, where the average quadratic effect is given by the mean of its factor, α_2.

Figure 8.5 illustrates quadratic trajectories for several values of quadratic effect, ranging from –1.5 to 1.5, and assuming no linear effect. In practice, the shape of the predicted curve will be a combined function of the linear and quadratic effects, so the appearance will deviate from *u* and *n* shapes in this plot. An example combining both linear and quadratic effects is reported for Example 8.2.

Identification

At least three time points are needed to fit a quadratic curve. With means, variances, and covariances estimated for the intercept, linear, and quadratic factors, the model will be underidentified for three time points without constraints, however. Recall that the number of parameters estimated in the typical linear latent growth curve model is $T + 5$. For the quadratic latent curve model that estimates three factor means (α_0, α_1, α_2), their variances (ψ_{00}, $\psi_{11,}$ ψ_{22}), the covariances among them (ψ_{01}, $\psi_{02,}$ ψ_{12}), and heterogeneous errors at each time point (θ_{11} through θ_{tt}), the number of parameters estimated is $T + 9$. With three time points, there are only $[3(3 + 1)]\ 2 = 6$ observed variance–covariance elements

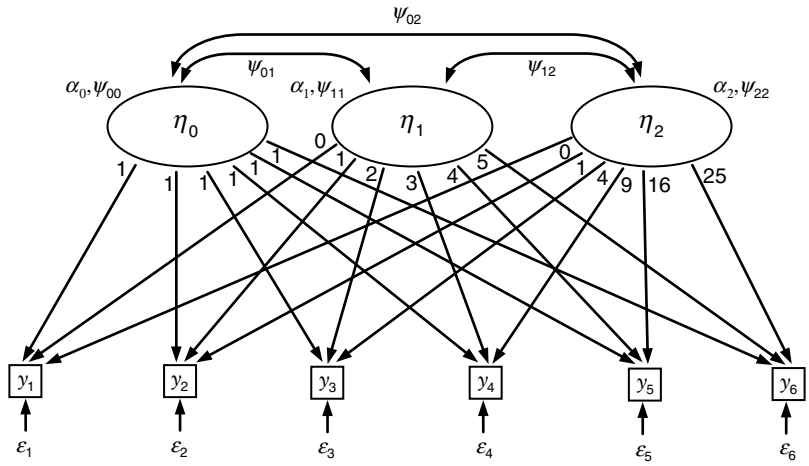

Figure 8.4 Quadratic Latent Growth Curve Model.

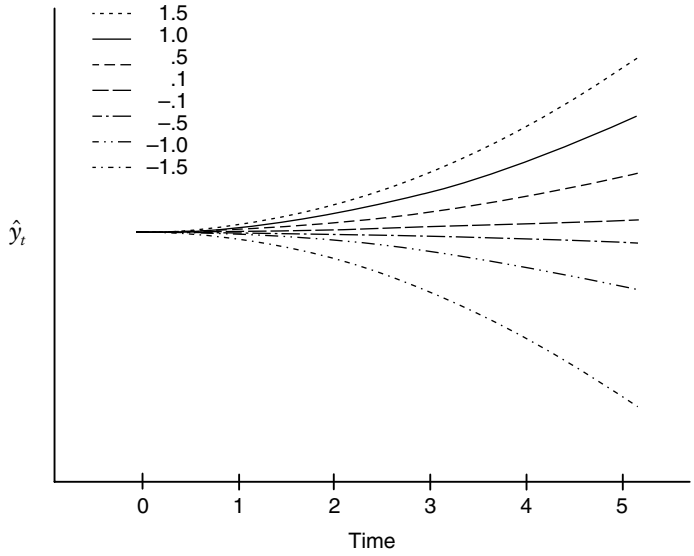

Figure 8.5 Quadratic Effects with Varying Coefficient Values.

and 3 observed means, a total of 9 observed values. This leaves the quadratic model with three more parameters estimated than available elements. Two parameters can be saved by requiring equal errors over time, but an additional constraint is needed. In the multilevel regression framework, a common strategy is to omit one of the random effects, which, in the context of the latent growth curve model, entails setting variance of one of the factors (intercept, linear, or quadratic) equal to 0. With the variance fixed, a constant value is used for the effect, eliminating the ability to estimate associations with other factors or covariates.

Although a model can be specified using only a single quadratic factor, omitting the linear factor, modeling both linear and quadratic effects together is often recommended

as at least an initial step to demonstrate that there is a significant quadratic change that exists independently of the linear change. Without the linear factor in the model, the mean of the quadratic factor may differ from zero simply because there is a linear effect. To determine whether a quadratic function is appropriate, specific parameter tests of the quadratic mean or variance can be examined with the Wald test or the models with and without the quadratic effect can be compared with a likelihood ratio test. The two tests are not equivalent, because the likelihood ratio test necessarily must include the variance of the quadratic factor and the covariance between the quadratic factor and the other two factors. Willett and Sayer (1994) recommend the more global likelihood ratio test, based on the reasoning that the variance of the quadratic effect also may be of interest. If the mean of the quadratic factor is not significant, it indicates that there is no departure from the linear effect on average; but there may still be meaningful variability in the quadratic effect that could be accounted for by other factors. The likelihood ratio test, however, is not specific to variance of the quadratic effect. It necessarily includes a test of the covariances among the factors as well.

Coding Time

Centered Time Codes. The simultaneous estimation of the linear and quadratic effect may cause multicollinearity between the two factors, because the quadratic time values are a multiplicative function of the linear time values. This non-essential multicollinearity results in inflated standard errors of the linear effect (Aiken & West, 1991; Stimson, Carmines, & Zeller, 1978). The test of the quadratic effect or highest-order effect remains unbiased. It is only the tests of the lower-order coefficients that are impacted. Fit of the model also is unchanged by centering. A solution to the multicollinearity problem is to rescale the time values and center them around the midpoint. With three time points, the linear time scores become –1, 0, 1, and with five time points, the linear time scores become –2, –1, 0, 1, 2. With an even number of waves, the centering point falls in between the two middle scores. For example, with six waves, the centered linear time scores would be –2.5, –1.5, –.5, .5, 1.5, 2.5. Rescaling the time scores generally leads to a dramatic reduction in the magnitude of the correlation between the linear and quadratic time scores. Although a plot of the centered time scores ignoring the linear effect, –2, –1, 0, 1, 2, would suggest a centered parabola, the plot of the predicted scores will look much the same from the centered and uncentered models when both linear and quadratic effects are used to compute the predicted scores; this is because the estimates of the average linear and quadratic effects are altered in a compensatory manner when the time scores are centered.

Instantaneous Rate of Change. Rescaling changes the interpretation of the coefficients, and this can be disconcerting. With centered values, the intercept is no longer associated with a baseline value but with the midpoint of the study. For models with quadratic or higher order effects, the intercept estimates the average across the time points when centered time scores are used. The linear effect is the constant or average rate of change taken over the full time series in the linear growth curve model, but its meaning differs when there is a quadratic effect in the model. With a quadratic effect included, the linear effect represents the estimate of the instantaneous rate of change when $t = 0$. We can compute the *instantaneous rate of change*, which is the expected linear increase as the time increment approaches 0, at any other point along the curve. The instantaneous rate of change, which I will denote $\alpha_{\Delta t \to 0}$, is given by $\alpha_1 + 2\alpha_2(t-1)$, and is the first derivative of the quadratic model. Some further detail on derivatives of growth curves can be found in Appendix B.[4]

The instantaneous rate of change can then be computed for any time value $(t-1)$ desired. With a little manipulation, this equation can then be used to find the time point at which the instantaneous rate is equal to 0, $(t-1)_{\alpha_{\Delta t \to 0}} = -\alpha_1 / (2\alpha_2)$. This is a way to find the minimum or maximum of the curve – the point at which the curve begins to change directions. Stimson and colleagues (1978) recommend finding this point and centering the time codes at $(t-1)_{\alpha_{\Delta t \to 0}}$ to obtain a more meaningful estimate of the intercept, the linear effect, and the correlation between the linear and quadratic trajectories that correspond to the minimum or maximum of the curve.

Orthogonal Polynomials. An additional approach to using raw or centered time coding schemes is to use *orthogonal polynomials*. Orthogonal polynomial codes can be used to model quadratic (as well as higher-order) effects and are constructed to minimize the correlation among the factors. Any two sets of codes are orthogonal if the sum of their products equals zero. For example, the centered codes for six time points mentioned earlier are orthogonal were $\lambda_{t1} = -2.5, -1.5, -.5, .5, 1.5, 2.5$ and $\lambda_{t1}^2 = 6.25, 2.25, .25, .25, 2.25, 6.25$. The sum of their products is equal to 0: $(-2.5)(6.25) + (-1.5)(2.25) + (-.5)(.25) + (.5)(.25) + (1.5)(2.25) + (2.5)(6.25) = 0$. The raw time scores, however, are not orthogonal, because they do not sum to 0: $(0)(0) + (1)(1) + (2)(4) + (3)(9) + (4)(16) + (5)(25) = 225$. A common set of orthogonal set of linear and quadratic coefficients used for ANOVA is given by Hays (1994, appendix F, table VII) and are available from many other sources: $\lambda_{t1} = -5, -3, -1, 1, 3, 5$ and $\lambda_{t2} = 5, -1, -4, -4, -1, 5$.

Quadratic models with centered time codes and commonly used orthogonal polynomial codes will not produce the same coefficient estimates, though their significance tests will be the same. Coefficients obtained with orthogonal polynomials will tend to be larger as a result of smaller intervals used for the time codes (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004). A related issue is that higher-order effects always involve larger time code values than lower order effects, which puts the raw coefficients on a different scaling metric and makes it difficult to compare the relative magnitudes of the raw coefficients associated with different order effects.

When the smaller raw coefficient values are undesirable, the orthogonal polynomial coefficients can be rescaled to put the linear and quadratic effects on a similar metric (Hedeker & Gibbons, 2006). One method (Bock, 1975) rescales the coefficients by dividing by the square root of the sum of the squares of the codes. For example, the sum of the squares of the linear codes from Hays is 70, so dividing all of the linear codes by $\sqrt{70}$ gives –.598, –.359, –.120, .120, .359, .598. The sum of the squares of the quadratic codes from Hays is 84, so dividing all of the quadratic codes by $\sqrt{84}$ gives .546, –.109, –.436, –.436, –.109, .546. Table 8.1 gives orthogonal and rescaled orthogonal codes for designs with three through seven waves. Although rescaling the intercept codes is indicated by Hedeker and Gibbons, this produces intercept mean and variance estimates that are very different from the corresponding observed values. A standard set of loadings for the intercept factor, in which all loadings are set equal to 1, will produce results for the intercept that are more comparable to the observed values. Results from a model with rescaled orthogonal polynomial codes can be translated back to results from the original uncentered model (see Hedeker & Gibbons, 2006, for details).

Example 8.2: Quadratic Growth

The health and aging data set was used to model quadratic growth of BMI scores over the six periods of the study. The first model used uncentered loadings, $\lambda_{t1} = 0, 1, 2, 3, 4,$

Table 8.1 Orthogonal Polynomials for $T = 3$ through $T = 7$

	Orthogonal codes									
	$T=3$		$T=4$		$T=5$		$T=6$		$T=7$	
t	*Linear*	*Quadratic*	*Linear*	*Quadratic*	*Linear*	*Quadratic*	*Linear*	*Quadratic*	*Linear*	*Quadratic*
1	−1	1	−3	1	−2	2	−5	5	−3	5
2	0	−2	−1	−1	−1	−1	−3	−1	−2	0
3	1	1	1	−1	0	−2	−1	−4	−1	−3
4			3	1	1	1	1	−4	0	−4
5					2	2	3	−1	1	−3
6							5	5	2	0
7									3	5
	Rescaled orthogonal codes									
	$T=3$		$T=4$		$T=5$		$T=6$		$T=7$	
t	*Linear*	*Quadratic*	*Linear*	*Quadratic*	*Linear*	*Quadratic*	*Linear*	*Quadratic*	*Linear*	*Quadratic*
1	−.707	.408	−.671	.500	−.632	.535	−.598	.546	−.567	.546
2	.000	−.816	−.224	−.500	−.316	−.267	−.359	−.109	−.378	.000
3	.707	.408	.224	−.500	.000	−.535	−.120	−.436	−.189	−.327
4			.671	.500	.316	.267	.120	−.436	.000	−.436
5				.000	.632	.535	.359	−.109	.189	−.327
6							.598	.546	.378	.000
7									.567	.546
Sum of squares (SS)	2	6	2	4	1	14	7	84	28	84
Divisor ($\sqrt{SS}$)	1.414	2.449	4.472	2	3.162	3.742	8.367	9.165	5.292	9.165

5, for the linear factor and the squares of these values, $\lambda_{t2} = 0, 1, 2, 9, 16, 25$, as loadings for the quadratic factor. Measurement intercepts were set equal to 0, errors at each time point were freely estimated, and all factor means, variances, and covariances were freely estimated. The model fit the data well, $\chi^2(12) = 148.416$, $p < .001$, CFI = .998, RMSEA = .046. The average of the intercept factor was 27.164, which closely matched the observed mean at baseline, 27.176. The average of the linear factor, .230, was significant, $p < .001$. The standardized value was .247. The linear effect can be intercepted as the instantaneous rate of .230 points increase per wave (two-year period) on the BMI estimated from baseline. The quadratic effect was also significant, –.017, $p < .001$, suggesting a curvilinear slowing of the rate of increase over time. The standardized value was –.101, suggesting that the magnitude of the quadratic effect was small. The correlation between the linear and quadratic factors was –.865. A plot of the average predicted values was obtained by entering the time score values into the equation $\hat{y}_t = \alpha_0 + \alpha_1\lambda_{t1} + \alpha_2\lambda_{t2}$, and the predicted scores were used to generate the plot shown in Figure 8.6. The obtained parameters can also be used to estimate the point $(t-1)$ at which the instantaneous rate of change is equal to 0, which is the point along the curve where the growth levels off or begins to decline. Inserting the average linear and quadratic into the equation gives $t_{\alpha_{\Delta t \to 0}} = -\alpha_1 / (2\alpha_2) = -.230 / [(2)(-.017)] = 6.765$, which suggests that increases in BMI will continue past the end of the study period to between the seventh and eighth wave (between $t-1$ equal to 6 or 7) when it will begin to level off.

To illustrate the effect of time coding choice on the model parameters, three additional models were tested using centered ($\lambda_{t1} = -2.5, -1.5, -.5, .5, 1.5, 2.5$; $\lambda_{t2} = 6.25, 2.25, .25, .25, 2.25, 6.25$), common orthogonal polynomial ($\lambda_{t1} = -5, -3, -1, 1, 3, 5$); $\lambda_{t2} = 5, -1, -4, -4, -1, 5$), and rescaled polynomial loadings ($\lambda_{t1} = -.598, -.359, -.120, .120, .359, .598$; $\lambda_{t2} = .546, -.109, -.436, -.436, -.109, .546$). The results from these three models are presented in Table 8.2 along with the results from the uncentered model. Each model had identical fit to the uncentered model and, importantly, the tests of the significance of the quadratic effect were all identical to the test obtained in the uncentered model.

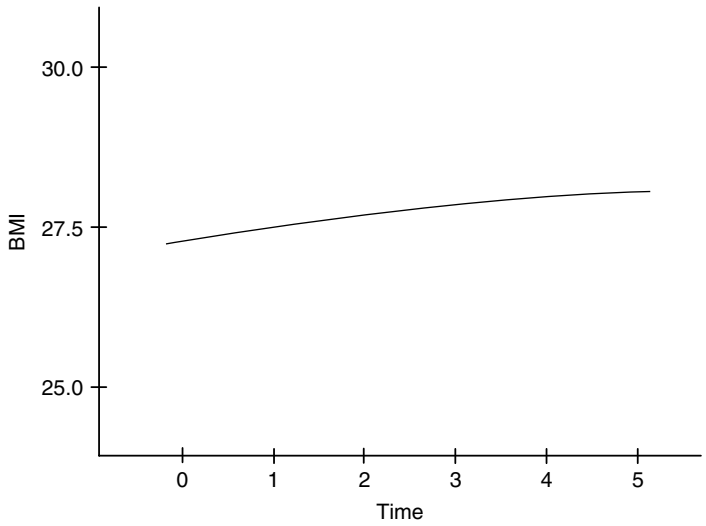

Figure 8.6 Plot of Predicted Scores from the Quadratic Model of BMI.

Table 8.2 Latent Quadratic Growth Curve Results for BMI Using Four Coding Methods

		Uncentered			*Centered*			*Common orthogonal polynomial*			*Rescaled orthogonal polynomial*		
		Estimate	*SE*	*z*	*Estimate*	*SE*	*z*	*Estimate*	*SE*	*z*	*Estimate*	*SE*	*z*
α	Intercept	27.164	.031	404.806	27.631	.069	398.652	27.581	.068	406.527	27.581	.068	406.527
	Linear	.230	.018	12.674	.144	.007	19.288	.072	.004	19.289	.601	.031	19.290
	Quadratic	−.017	.003	−4.958	−.017	.003	−4.958	−.012	.002	−4.958	−.106	.021	−4.959
ψ	Intercept	23.19	.467	49.640	25.046	.496	50.472	24.319	.475	51.149	24.319	.475	51.149
	Linear	.868	.045	19.278	.220	.006	35.927	.055	.002	35.927	3.848	.107	35.923
	Quadratic	.029	.002	18.642	.029	.002	18.643	.013	.001	18.642	1.083	.058	18.629
χ^2		148.416			148.416			148.416			148.278		
df		12			12			12			12		
CFI		.998			.998			.998			.998		
RMSEA		.046			.046			.046			.046		

Note: Parameter estimates are unstandardized. Common and rescaled orthogonal polynomial coefficients are given in Table 8.1 for $T = 6$ occasions.

On close inspection of the results from the three models, it is clear that the intercept factor means all differed somewhat from the uncentered model. The difference is due to the altered interpretation of the intercept given the various time coding schemes. The intercept in the centered model represents the expected value of BMI at the midpoint of the study (between waves 3 and 4), whereas the intercept in the two orthogonal polynomial models represents the expected value for the average across time points. Both values are relatively close to the observed values of 27.544 and 27.551, respectively. Even though the means of the linear and quadratic factors differed across each of the time coding approaches, it is important that the significance tests are identical.[5] Similarly, the standardized estimates for the linear and quadratic means are identical among the three scaling approaches, .306 for the linear effect and –.102 for the quadratic effect. Compared with the uncentered model, the correlation between the linear and the quadratic factors has been greatly reduced, .098, although it remained significant, < .005. Thus, even though the unstandardized values differ across the three models, the three versions of centered coding all represent models that are equivalent to one another. Of course, none of the three centered models have the same values as the uncentered model, because the location of the intercept and the interpretations of the linear effects differ in the uncentered and centered approaches. A plot of the predicted values from each of the three models (not shown) gives curves identical to the curve obtained for the uncentered model (Figure 8.6), indicating that the uncentered and centered models equally account for the observed data but simply partition the intercept, linear, and quadratic parameters in different ways.

Comments

As illustrated by Figures 8.5 and 8.6, the quadratic growth model is a form that can fit a variety of trajectories that may be predicted by theory or are suggested by the data. The quadratic factor in the latent growth curve model represents acceleration or deceleration in change over and above a linear increase or decrease. A wide range of curvilinear shapes can be modeled when the linear and quadratic effects are considered together. Stimson and colleagues (1978) argue that it is artificial to partition the curve into linear and quadratic components and may not represent the true form of the data pattern. One can argue the other side, however, that a parabolic shape alone may not be a fair representation of the data pattern either. Regardless, in initial tests, at least, both the linear and the quadratic factors are needed to verify that a significant quadratic function is not simply due to a linear increase or decrease. Another interesting point, made by Singer and Willett (2003), is that a nonsignificant average quadratic effect may not imply necessarily that it is fruitless to model a quadratic effect. When there exists significant variability in the quadratic effect, it may be valuable to attempt to explain such variation with covariates. Thus, attention should be paid to the variance of the factor as well as its average.

Even when a quadratic effect is not predicted by theory, it may be wise to investigate nonlinear effects for several reasons. And the analysis of growth data should be conducted in conjunction with alternative tools, such as diagnostic tests and graphical explorations. Plots of growth curves, individual and averaged are essential. Lack of fit of the model in the linear model may be due to the presence of a nonlinear trend and researchers should keep in mind that this is one potential cause of misfit in a linear model. Moreover, variance of the linear effect may be spurious and due to the presence of a nonlinear effect (Bauer & Cai, 2009). Another complicating concern is that quadratic and other curvilinear effects may be indistinguishable from other models. Raykov (1998) and Bianconcini (2012) show that a linear model with an autoregressive component is equivalent to a quadratic model under some circumstances. Researchers should therefore fully explore their data with as

many approaches as possible and should carefully consult theory when choosing the particular model and functional form.

Choosing an appropriate coding scheme for the problem is not an easy matter. One source of solace is the fact that the test of the quadratic effect will be the same regardless of the coding scheme one chooses. Multicollinearity should not be a concern in terms of assessing the quadratic effect (Aiken & West, 1991), given equivalent tests of the quadratic factors mean and variance whether the times scores are centered or not. Each coding scheme implies a different interpretation of the intercept, the average linear effect, the variance of the linear effect, and the covariance between the factors. Biesanz and colleagues (2006) and Hedeker and Gibbons (2006) differ on whether orthogonal polynomials are desirable. One states "the coefficients obtained via orthogonal polynomial contrast codes can be difficult to properly interpret" (Biesanz et al., 2006, p. 40), whereas the other states "An additional advantage of using orthogonal polynomials, over simply centering time, is that the polynomials are put on the same scale" (Hedeker & Gibbons, 2006, p. 86). But the apparent differences in the estimates are a result of different partitionings of the linear and quadratic effects and none represents an "incorrect" method.[6] Statistically there is no difference. One consideration not mentioned thus far is that the orthogonal polynomial codes presented above are designed for equally spaced interval designs. When spacing is unequal, appropriate modifications can be made to the codes. The general method of Cholesky factorization, which serves as the basis of the orthogonal polynomials presented in Table 8.1, can be used to derive codes for designs with unequally spaced intervals (see Bock, 1975; Hedeker & Gibbons, 2006 for detailed descriptions).[7]

Other Polynomial Growth Curves

The general concepts and specification of the quadratic growth curve model can be applied to a near limitless variety of other functions that modify time codes using exponential transformations. The first of these to be considered is the cubic growth curve model (Figure 8.7), in which the time scores are raised to the third power, λ_{t1}^{3}. In general, a cubic trend, which has an S-shape or two bends in the curve, requires a minimum of four time points to model. A model with linear, quadratic, and cubic factors that freely estimate measurement residuals at each time point, all factor means, variances, and covariances, requires $T+14$ parameters to be estimated. This suggests a minimum of five time points with this parameterization, $df = \left[J(J+1)/2+J\right]-(T+12) = \left[5(5+1)/2+5\right]-(5+12) = 1$. With four time points, four additional constraints would be needed that could be accomplished by setting measurement residuals equal over time and constraining one other parameter (e.g., a factor variance).

The cubic polynomial model can fit a variety of patterns when considered with lower-order effects. Figure 8.8 displays just a few of the curves that might result from a combination of coefficient values for the linear, quadratic, and cubic factor means. Unlike the quadratic model, centering of time codes has an important impact on the shape of the function with the cubic model. Centering moves the location of the first bend in the curve (i.e., stationary point or point of inflection), and when uncentered time codes are used, the first bend in the curve is moved to the first time point. This results in a shape to the curve that resembles the quadratic curve, because the first bend is eliminated. Centered time scores will produce two bends in the curve where the change in the direction is located at the center time point. The plot in Figure 8.8 was generated using centered time codes.

Beyond the quadratic and cubic forms, other transformations of the time codes can be used to model other forms. Higher-order power transformations create curves with additional bends. Root transformations (e.g., square root or $x^{1/2}$, cube root or $x^{1/3}$) are

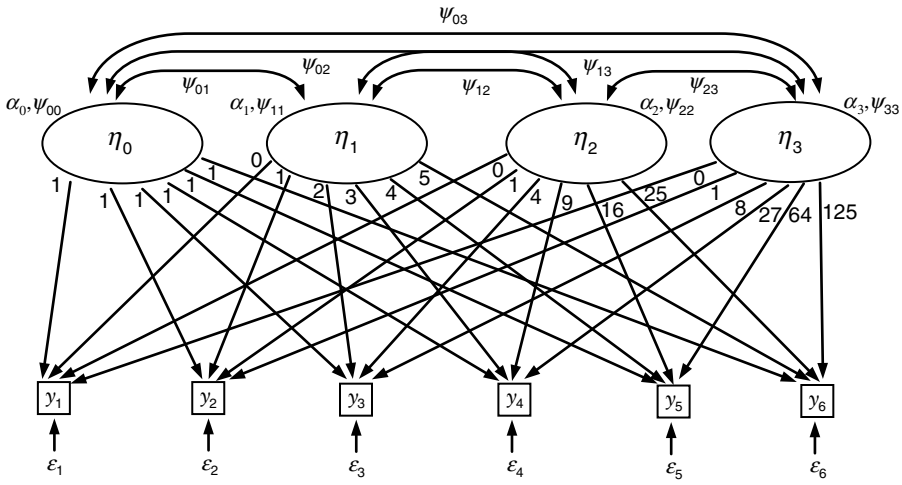

Figure 8.7 Cubic Growth Curve Model.

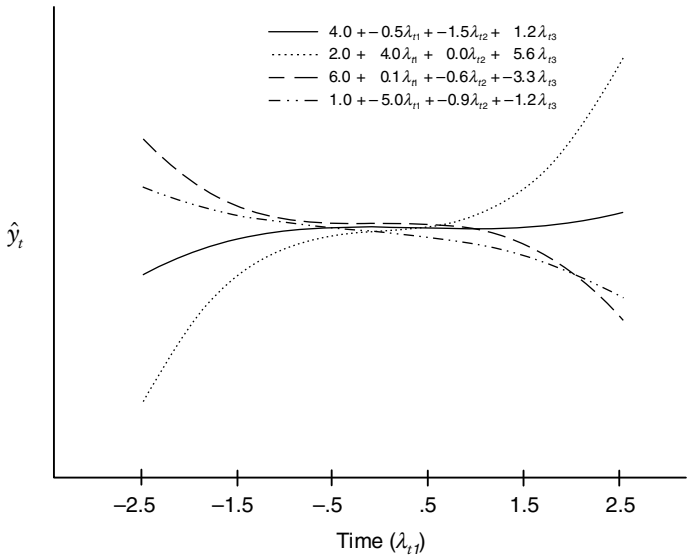

Figure 8.8 Several Possible Cubic Trends.

also possible. Root transformations produce a convex-upward sloping curve, similar to Figure 8.6, but they cannot be used with centered time codes, since they are undefined with negative numbers unless imaginary numbers are involved. Exponential transformations, using the mathematical constant *e* as the base (e.g., e^x), are popular and can be used with uncentered or centered time codes. When the coefficient is positive (i.e., positive factor mean), the exponential trend represents a concave upwardly sloping curve, and when the coefficient is negative, the exponential trend is a downwardly sloping convex curve. Logarithmic transformations, which are popular for transformations in regression analysis, are undefined for zeros or negative numbers, so are not useful for transforming

time codes. Similarly, inverse functions (e.g., x^{-1}, x^{-2}) are problematic, because they are undefined whenever the time value is zero.

Growth Models with Nonlinear Parameters

All of the above growth curve models are said to be "linear in their parameters," because the equations that describe them involve a set of additive terms or can be stated as a set of additive terms. For example, Equation (8.1) has an intercept term, a linear growth term, and a quadratic growth term. The parameters in this model all have first-order exponents, even though the variables (vis-à-vis the factor loadings for the quadratic factor) do not. Models with nonlinear parameters, on the other hand, involve products of the parameters (e.g., $\beta_{i1} \times \beta_{i2}$) or transformations of the parameters (e.g., $e^{\beta_{i2}}$). Models that are linear in their parameters are said to be *dynamically consistent*, whereas models that are inherently nonlinear are not dynamically consistent. Dynamically consistent models have average parameter estimates that are equal to the average of the individual case-level parameter estimates. The average of the curve is equal to the curve of the averages. For example, the quadratic growth curve model is dynamically consistent, because the curve fitted to the means taken at each time point is equal to the average of the curves fitted to each individual case. The determination of whether a model is dynamically consistent is not always obvious, because sometimes a non-additive equation can be re-expressed as an equivalent or approximate model that is additive (Preacher & Hancock, 2012).

Exponential curves, logistic curves, negative binomial curves, monomolecular curves, and Gompertz curves are just a few of the possible "truly nonlinear" curves available. Exponential and logistic curves are widely used and take on several different forms. The Richards curve (Richards, 1959), originally used to describe plant growth, is a variant on the logistic curve also referred to as the generalized logistic curve. The Jenss–Bayley curve (Jenss & Bayley, 1937), developed to describe infant growth, is a variant of the exponential curve. These curves are often sigmoidal or S-shaped and have an inflection point where the curve changes direction and an asymptote where the curve approaches some maximum or minimum value. These points can be specified a priori or derived from the data, depending on the application. Although originally rooted in biological applications, these nonlinear curve families have wide applicability and have been used in research on learning, population growth, and economics.

Beginning with Browne and du Toit (1991), several authors have illustrated how the latent growth curve framework can be adapted to estimate models that are nonlinear in the parameters (Choi, Harring, & Hancock, 2009; Grimm & Ram, 2009; Grimm, Zhang, Hamagami, & Mazzocco, 2013; Preacher & Hancock, 2012; Ram & Grimm, 2007). Rather than recount all of these particular forms, I will describe the latent growth curve modeling specification of the exponential and logistic curves to illustrate some of the basic principles. Modifications to this strategy can then be used to estimate other nonlinear forms as needed. The exponential model closely resembles the curve modeled by the quadratic model when a linear effect is included (see Figures 8.5 and 8.6), and the logistic model closely resembles the cubic model with lower-order effects included (see Figure 8.8; Burchinal & Appelbaum, 1991).

One exponential growth curve is described by the following equation:

$$\hat{y}_{ti} = b_{0i} + b_{1i}\left(1 - e^{-b_{2i}t}\right) \tag{8.2}$$

Its form resembles the case-level growth model except that the mathematical constant e (approximately 2.718) is raised to $-1 \times$ the parameter b_{2i}, making it a nonlinear equation.

Equation (8.2) is the theoretical equation for the case level, but the average values, b_0, b_1, and b_2 will be interpreted in practice. If we assume that t represents time scores of 0, 1, 2, etc., then b_0, gives the predicted baseline score (i.e., the usual interpretation of the intercept). The parameter b_1 represents the amount of increase or decrease to the highest or lowest point on the curve. Then $b_0 + b_1$ gives the highest or lowest point on the curve. With centered time scores, b_0 is the predicted value at the middle time point, b_1 is the amount of increase from the middle point to the highest point, and $b_0 + b_1$ is the highest point. Centered and uncentered scores will produce identical curves, with an adjustment to the *y*-axis for the rescaling. The rate parameter, b_2, gives the exponential rate increase, if positive, and exponential rate decrease, if negative, per wave. The parameters b_1 and b_2, however, do not have simple interpretations, because they act to compound one another. Although b_1 determines the minimum or maximum point on the curve, the sign of b_2 also can operate to determine whether the curve is concave or convex. Generally, a positive value of b_1 will tend to produce a concave curve that slopes upward, but, if the point implied by b_2 is higher than the intercept b_0, it will act to change the curve to a convex shape or an S-shape.

The exponential in Equation (8.2) can be reparameterized so that it is more easily testable using our latent growth curve framework. The reparameterization is based on Taylor series approximation, a method of approximating a function with the sum of all its derivatives. I will spare you the mathematical details of the reparameterization here, but more explanation can be found elsewhere (Blozis, 2004; Browne, 1993). The reparameterized model fits our general structure as the model in Figure 8.4 (Grimm, et al., 2013; Ram & Grimm, 2007),

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\eta_{1i} + \lambda_{t2}\eta_{2i} + \varepsilon_{ti} \tag{8.3}$$

but the exponential model replaces the loadings for the two factors with $\lambda_{t1} = 1 - e^{-b_2 t}$ and $\lambda_{t2} = \alpha_1 t\left(e^{-b_2 t}\right)$. The loadings, λ_{t1} and λ_{t2} are computed using the exponential function and time codes (e.g., 0, 2, 3,…$T-1$). The parameter α_0 is the mean of the intercept factor η_0, which has loadings all equal to 1, and its variance estimated, ψ_{00}. The factor η_1 has mean α_1, representing the rate parameter (same as b_1 in Equation [8.2]) and its variance ψ_{11} is estimated as in the latent growth curve models previously discussed. The parameter b_2 is new and must be estimated using complex modeling constraints.[8]

The logistic function can be estimated with the same general model as the linear model,

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\eta_{1i} + \varepsilon_{ti}$$

where the loadings are transformed values of time scores, such as $t = 0, 1, 2, \ldots T-1$. The transformation has an additional level of complexity relative to the exponential model, because two parameters are estimated that provide additional information about the shape and location. Also based on a reparameterization of the logistic function, the transformation of the time scores are used as the loadings for factor η_1.

$$\lambda_{t1} = \frac{1}{1 + e^{-(t - b_2)b_3}} \tag{8.4}$$

There are two parameters needed to determine the loadings: b_2, a location parameter, which gives information about the inflection point, and b_3, which gives information about the overall increase or decrease over the time points of the study. As with the exponential model, the parameters are very much interdependent and not simple to interpret individually. The mean of the rate factor, α_1, gives information about the change in *y*. Generally,

when the overall rate is positive, there is an increase over time; and when the overall rate is negative, there is a decrease over time. The mean of the intercept factor α_0 is the predicted baseline value if $t = 0$. When the overall function is increasing (α_1 is positive), α_0 is the lowest value, and when the overall function is decreasing (α_1 is negative), α_0 is the highest value. We can determine the high or low point on the curve with these two values, because $\alpha_0 + \alpha_1$ is the upper asymptote for an increasing function and $\alpha_0 + \alpha_1$ is the lower asymptote for a decreasing function.

Depending on the rate of increase, the upper/lower asymptote may not occur in the time frame of the study. The parameter b_2 is the inflection point, which is the value on the *x*-axis (time point) in which the curve changes direction. If time scores are centered, the value of α_0 is still associated with the baseline value, but it changes the interpretation of the inflection point, which is interpreted as the distance on the *x*-axis from the middle time point. The sign of the inflection point can change the direction of the curve (e.g., from concave to convex), although the general direction of the curvature can also be determined by the other parameter, b_3. The sign of b_3 can reverse the direction indicated by α_1. Larger absolute values for b_3 indicate more curvature (i.e., more of an S-shape), and smaller absolute values indicate less curvature (i.e., flatter).

Example 8.3: Exponential and Logistic Curve Models

Although the quadratic model fit the BMI data well, an exponential curve may be a more precise representation of the weight trajectory over time. The general model structure is described by Equation (8.3) and is the same as that for the quadratic model (Figure 8.4) with modified loadings. The exponential transformations of uncentered time scores (with $t = 0, 1, 2, 3, 4, 5$), $\lambda_{t1} = 1 - e^{-b_2 t}$ and $\lambda_{t2} = \alpha_1 t\left(e^{-b_2 t}\right)$, were used to obtain the loadings for the two growth factors. The parameter b_2 and loadings for all λ_{t1} and λ_{t2} were estimated as part of the model using complex constraints following these formulas. The chi-square was significant, $\chi^2(12) = 149.797$, $p < .001$, but the model fit the data well according to the alternative fit indices, RMSEA = .046, CFI = .998. This model fit the data nearly as well as the quadratic model in Example 8.2, which had a chi-square value of 148.416 with 12 *df*. The mean of the intercept factor in the exponential model was 27.169. The mean of the rate factor, α_1, was 1.158, $p < .001$, suggesting an increase of a little over a point on the BMI scale to the highest value reached during the study period. The sum of these two parameters, $\alpha_0 + \alpha_1 = 28.327$, gives the predicted baseline value for the curve and indicates that the initial average BMI was in the overweight range. I will omit the plot of the curve, as it is indistinguishable from the plot in Figure 8.6.

A logistic curve was also fit to the BMI data with an intercept factor, η_0, and a rate factor, η_1, and loadings computed according to Equation (8.4). Complex constraints were needed to estimate these loadings as well as the b_2 and b_3 parameters. Although the alternative fit indices indicated that the fit of this model was good overall, $\chi^2(14) = 442.147$, $p < .001$, RMSEA = .076, CFI = .993, the fit was substantially worse than the quadratic or the exponential models. The intercept estimate, α_0, was 27.232, a value similar to the baseline values predicted in the previous models. The average of the rate factor, α_1, was .707, suggesting a general increase over time. The sum of these two values, 27.232 + .707 = 27.939 gives the estimate of the upper asymptote on the curve. The estimated shape parameter b_3 was 3.488 and suggested a departure from a flat curve toward an S-shape. The estimated inflection point b_2 was 1.233, indicating that the bend in the curve was between the first and second time points, and this would be the point at which weight levels off and increases less rapidly over the remainder of the study. Figure 8.9

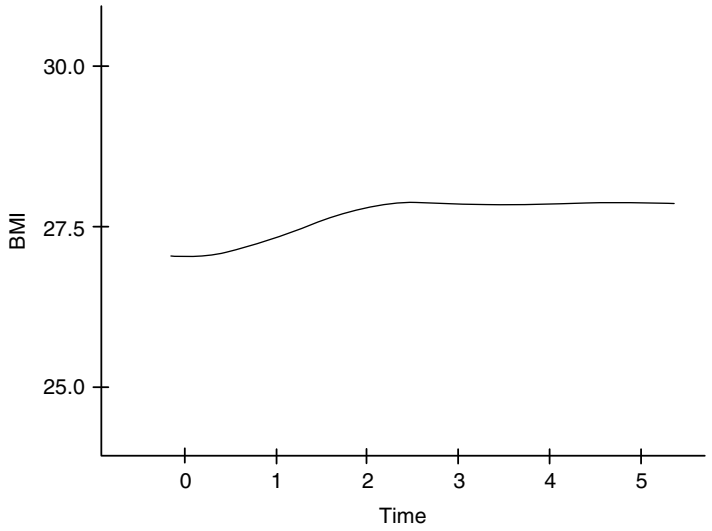

Figure 8.9 Plot of the Logistic Latent Growth Curve Results.

plots the average logistic curve, although the poorer fit of the model relative to the fit of the quadratic or the exponential models would suggest that this function is not optimal for the data.

Comments

The nonlinear parameter models I have described are only two of the possible functional forms that can be modeled. The models are based on the general strategy suggested by Browne and du Toit (1991; Browne, 1993) and elaborated on by others (e.g., Blozis, Harring, & Mels, 2008; Grimm & Ram, 2009; Grimm et al., 2013; Preacher & Hancock, 2012; Ram & Grimm, 2007). This approach has proven to be highly flexible and suitable for testing a variety of functions. Choi and colleagues (2009) have suggested an alternative formulation based on transformation of the dependent variable that also may be promising for nonlinear latent growth curve models. Estimation of the models in Example 8.3 for six time points needed starting values to achieve convergence, and convergence may be a bigger issue with fewer time points or more complex forms that require additional shape and location parameters. The model reported under Example 8.3 allowed freely estimated measurement residual variances with no distributional parameter constraints. Homogeneity of variance might also be assumed by constraining residual variances to be equal over time, potentially improving the likelihood of convergence. Grimm and colleagues (Grimm et al., 2013) also placed constraints on the pattern of measurement residual variance over time based on what would be expected from the exponential function. More work also may be needed to understand the extent to which these models conform to assumptions of error distributions of occasion-specific residual variance under ML.

The original formulation proposed by Browne and du Toit (1991) is referred to as the structured latent curve model, in which the loadings are constrained based on a nonlinear form. The latent growth curve approach presented here has been referred to as conditionally linear (Blozis & Cudeck, 1999; Blozis et al., 2008), because the loading matrix is estimated within nonlinear constraints. The conditionally linear model is a

combination of linear and nonlinear components but is a nonlinear function approximated to be linear in the parameters. The factors for these models can be random, allowing their values to vary across cases in the data set. Shape and location parameters, b_1, b_2, and b_3, are estimated through complex equality constraints and do not have estimated variances, however, so they are not random effects. Multiplicative random effects can be estimated in multilevel regression programs, and this is one potential advantage over the latent growth curve approach to nonlinear models (Grimm & Ram, 2009).

Differential Structural Equation Models

Following physical science and engineering concepts of oscillating systems, Boker (1996, 2001; Boker & Graham, 1998) has formulated a flexible approach to modeling nonlinear longitudinal change rooted in differential calculus.[9] The approach, which is less complicated than it sounds, uses a structure similar to the general latent growth model to realize several additional concepts to describe complex functions over time. The essential concepts come from the first and second derivatives of curves that we have already discussed in relation to other nonlinear models. The first derivative of a function gives information about the instantaneous rate of change and the second derivative gives information about the change in the rate, or acceleration, two concepts related to nonlinear change that can be exploited using the general SEM framework.

Loadings for the Differential Model

The loadings for the differential structural equation model are obtained from a simple method of approximating the first and second derivative, called *local linear approximation*. The essential idea can be stated in terms of differences. A linear slope can be described in terms of the first derivative, where the slope is a ratio of the change in y between any two points relative to the change in some time metric, t, $\beta_1 = (y_2 - y_1)/(t_2 - t_1) = \Delta y / \Delta t$. The concept of the derivative, however, is to estimate this ratio for an infinitesimally small increment, so theoretically Δ represents a difference between two points on a curve that approaches, but does not reach, a limit of 0. As we know, the linear slope gives information about the rate of change. A simplified description of the second derivative can be stated in terms of two sets of differences. For example, if the first two points mentioned are said to be Δy_{2-1} and Δt_{2-1}, then, using the same type of difference between the subsequent adjacent points, the local linear approximation of the second derivative is $(\Delta y_{3-2} / \Delta t_{3-2}) - (\Delta y_{2-1} / \Delta t_{2-1})$. This "difference between differences" captures the change in the rate over time or acceleration. (Appendix B provides a brief introduction to derivatives.)

The process for obtaining the loadings to estimate these derivatives can be stated in more than one way. One method involves constructing an "embedded" matrix, offsetting the series of time codes in each row (e.g., 0, 1, 2, 3 and 1, 2, 3, 4), and then subtracting the average value for each row. This process produces centered time scores. The centered time scores are then transformed according to the appropriate derivative order (for details, see Deboeck, Montpetit, Bergeman, & Boker, 2009). This process can be summarized with the following formula, which is general for any order derivative model:

$$\lambda_{ta} = \frac{\left[\Delta_t (t - \bar{t})\right]^a}{a!}$$

The derivative order is represented by a, with 1 used for the first derivative, and 2 used for the second derivative factor. Δ_t is the spacing of intervals, and when time points are single-unit increments, this value is equal to 1. The "!" symbol in the denominator is for the factorial operation. For the first derivative, the denominator is 1, because $a! = 1$. For the second derivative, the denominator is 2, because $a! = 2 \times 1 = 2$.

The resulting matrix of loadings for single-unit spacing is the following, with columns for the intercept, first derivative, and second derivative factor loadings, respectively (see Appendix A for more information about the lambda loading matrix):

$$\mathbf{L} = \begin{bmatrix} 1 & -2.5 & (-2.5)^2/2 \\ 1 & -1.5 & (-1.5)^2/2 \\ 1 & -.5 & (-.5)^2/2 \\ 1 & .5 & (.5)^2/2 \\ 1 & 1.5 & (1.5)^2/2 \\ 1 & 2.5 & (2.5)^2/2 \end{bmatrix}$$

Alternatively, the same values can be obtained by taking the indefinite integral (i.e., the antiderivative) of first and second order (Boker, Neale, & Rausch, 2004). It should be clear that these loadings are the same as the loadings for centered quadratic model, with the quadratic loadings rescaled by one half. Cudeck and du Toit (2002) show how the standard quadratic model can be reparameterized using differentiation. Higher-order derivatives, such as the third or fourth derivatives, similarly correspond to rescaled, centered polynomials for the cubic and quartic coefficients (Boker, 2007). The denominators used for rescaling are $3! = 6$ and $4! = 24$, respectively. The method of approximating derivatives can be extended through automated procedures with specialized software, known as generalized local linear approximation, allowing the user to choose any number of embedded dimensions, estimate derivatives for higher orders, use different time increments, and apply to more occasions (Boker, Deboeck, Edler, & Keel, 2010).

Damped Oscillator Model

Conceptualizing the factors as first and second derivatives in this way makes it possible to obtain useful longitudinal interpretations with some minor modifications of the growth curve model. Consider three factors like those we have defined thus far with factor loadings specified such that η_0 represents an intercept, or baseline value, η_1 represents a linear factor, and η_2 represents a quadratic factor. Until now, we have only considered models in which these three factors were allowed to correlate. We did not impose a causal structure on their relationship. If we allow η_2 to be predicted by η_1 then the path β_{21} represents the extent to which the acceleration is a function of the linear rate of change (see Figure 8.10). The model can be expressed as the following structural equation:[10]

$$\eta_2 = \beta_{20}\eta_0 + \beta_{21}\eta_1 + \zeta_2$$

Assume for a moment positive values for the mean of these two factors (both $\alpha_1 > 0$ and $\alpha_2 > 0$). The form of the curve would be an increasing function that curves upward, so that the increase over time becomes more rapid. Human population growth is one trend that resembles this pattern where there is an increasingly rapid rise over time. In

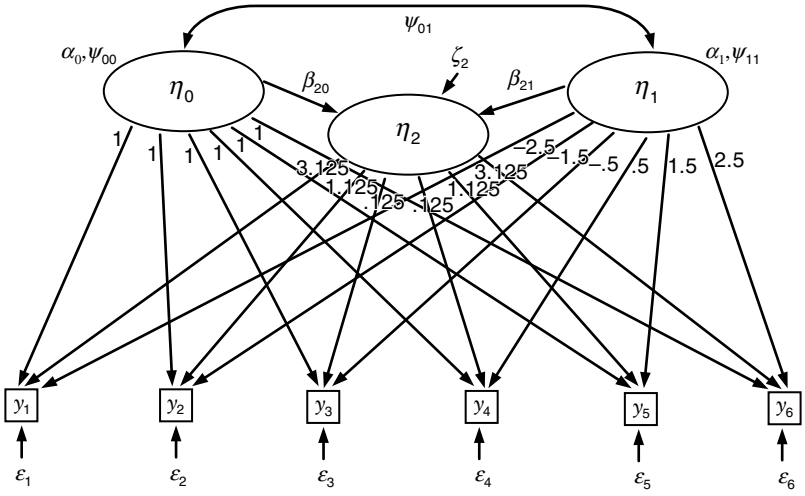

Figure 8.10 Damped Oscillator Model.

this case, a positive path between the linear and nonlinear factors would suggest that greater linear increase is associated with greater acceleration of the curve or a more dramatic rise in the predicted value over time. The linear effect has an *amplifying effect* on the nonlinear component. Countries or regions with steep increases would tend to have more rapid acceleration over time. A negative path between the linear and the nonlinear factors would suggest that greater linear increase is associated with a deceleration over time. The linear component has a *damping effect* on the nonlinear component. Some plant and animal populations may follow this pattern – when the general population rate is too steep, environmental limitations (e.g., food scarcity) force a slowing of the rate of increase. Multiplication of beer yeast accelerates to a certain point, but when its production of alcohol reaches a level of concentration that is beyond its tolerance level, reproduction is slowed.

More complex functions can model oscillation of the trend over time (Boker, 2007). As shown in Figure 8.11, data may increase and decrease within a wave form over time. When the oscillations decrease over time, the trend is said to converge toward a state of *equilibrium*. If negative, the coefficient β_{21} is a damping coefficient indicating the speed at which oscillations decline over time. Note that the damping process is related to the regression toward the mean and entropy concepts discussed in Chapters 4 and 5 (see also Voelkle & Oud, 2013). If the coefficient β_{21} is positive, it is an amplification coefficient indicating the speed at which the oscillations increase over time. Though the curve modeled is nonlinear, the differential structural model has been referred to (possibly confusingly) as the "linear oscillating model," because the model is linear in the parameters with additive terms.

The parameter β_{20} represents the relation between the starting point and the acceleration. The starting point may be the center time point if loadings are defined as in the above example or it could be defined at baseline. For oscillating systems, the parameter β_{20} is interpreted as the squared frequency of the oscillation rate. Smaller values of the coefficient represent a longer rate oscillation and larger values of the coefficient represent more rapid oscillations.

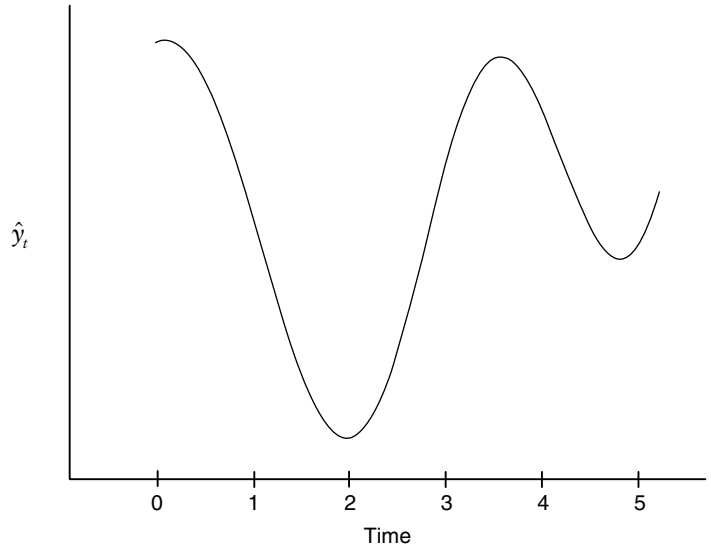

Figure 8.11 Example of an Oscillating Curve.

Example 8.4: A Simple Differential Equation Model

A damping oscillator model was tested for the BMI data using the differential approximation for the loadings (see Figure 8.10). The intercept factor, η_0, had loadings set equal to 1, the first derivative factor, η_1, had loadings set equal to –2.5, –1.5, –.5, .5. 1.5, 2.5, and the second derivative factor, η_2, had loadings set equal to 3.125, 1.125, .125, .125, 1.125, 3.125. The second derivative factor was predicted by the intercept factor and the first derivative factor, and these factors were allowed to correlate with one another. The fit of the model was the same as the quadratic model fit in Example 8.4, with $\chi^2(12) = 148.416$. The effect of the first derivative factor on the second derivative factor, β_{20}, had an unstandardized value of .099, $p < .001$, suggesting that there was an amplifying effect. Greater linear increase in weight over time was associated with an increasing rate of weight increase. The standardized estimate was .136, suggesting this was a relatively small effect. The unstandardized effect of the intercept factor on the second derivative factor was –.015, $p < .001$. The standardized effect was –.219, a small to medium effect. In the context of this quadratic model, the frequency suggests those who weighed more at the middle point in the study had slower acceleration of weight gain over time. Had this been an oscillating system, we could estimate the frequency of oscillation by squaring the unstandardized coefficient.

Latent Basis Model

The nonlinear models discussed thus far have tested an a priori hypothesis about the shape of the growth curve over time. A more data-driven approach is also possible if the latent growth curve is fit to the observed points by allowing the loadings to be freely estimated. This *latent basis model* has been referred to by a number of terms, including unspecified growth model, a linear spline model, and estimated loadings model among others.[11] This approach to latent growth modeling has been around since its inception, discussed both by Meredith and Tisak (1984, 1990) and McArdle and Epstein (1987). The idea for the flexible form approach flows naturally out of the exploratory factor analysis approach to

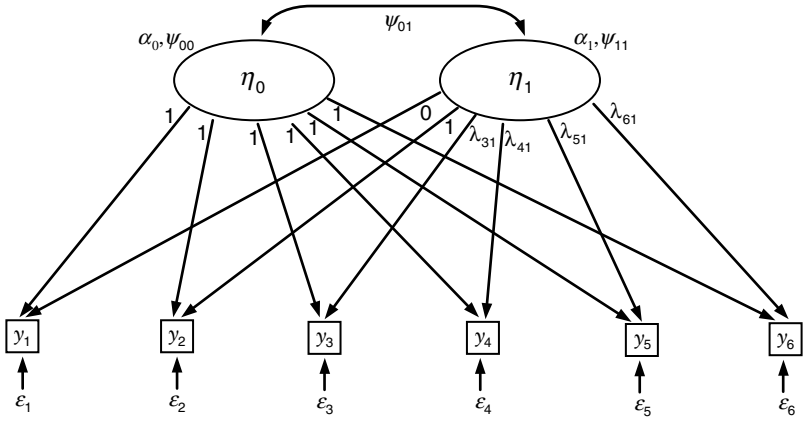

Figure 8.12 Latent Basis Model.

growth curves that were the precursors of the latent growth curve model approach (e.g., Tucker, 1958; Rao, 1958). Because of the form of this type of model, it will generally fit as well or better than a priori specified models.

Figure 8.12 shows one form of the model specification with an intercept factor, η_0, having loadings set equal to 1, and a nonlinear factor, η_1, having estimated factor loadings. To identify the model, two loadings for the latent basis (or shape) factor need to be constrained. The most common approach is to constrain the first and second loadings to 0 and 1. With this specification, as long as the first loading is set equal to 0, then the intercept factor is interpreted as the baseline value.

When the two loadings are constrained to 0 and 1, they set a scaling referent by which the remaining loadings can be interpreted. The loading for an indicator at any time point can be shown to be a function of the ratio of differences between the observed mean at that time point and the mean at the second time point in comparison to the mean at the first time point (Bollen & Curran, 2006; Stoolmiller, 1995).

$$\lambda_{t1} = \frac{\bar{y}_t - \bar{y}_1}{\bar{y}_2 - \bar{y}_1} \tag{8.5}$$

If $t = 3$, for example, then the loading for the third time point can be obtained by dividing the difference of the mean at the third time point by the difference between the means for the first and second time points. (Note that the difference between the first two time points cannot be exactly zero.) This series of linear ratios is essentially the same process as that used to estimate b-spline regression models. A loading of 2 is what would be expected if the degree of change is equal between $\bar{y}_3 - \bar{y}_1$ and $\bar{y}_2 - \bar{y}_1$, indicating a linear rate of increase. A perfect linear increase then produces loadings of 0, 1, 2, 3, …, $T - 1$. If the ratio obtained is larger than $t - 1$, the mean change is greater than what would be expected by a linear increment. This should sound familiar, as it is related to the concept of acceleration. When loadings are larger or smaller than a one-unit increment, they indicate that the rate of change is not constant.

The mean of the latent basis factor, α_1, represents an average change over all of the time points and can be approximated by $(\bar{y}_{T-1} - \bar{y}_1) / \lambda_{T-1,1}$, where the last time point is $T - 1$ and the final estimated loading is $\lambda_{T-1,1}$. A linear slope is approximated by dividing the difference between the first and last mean by $T - 1$, whereas the mean of the latent

basis model divides the difference by the estimated loading, which adjusts for the nonlinear increment over time. Theoretically, the model can be estimated with only three time points, but convergence problems may arise if there are few data points. When estimation difficulties arise, constraints on residual variances to be equal may increase the chances of estimation.

Alternative scores for the model are possible as well. McArdle (1988) proposed setting the first and last loadings to 0 and 1 to obtain loadings that represent the mean change relative to the total difference between the first and last time points. Thus, the loading represents a proportional increment. The two models have identical fit values, and the estimates from the two models can be obtained from one another. To find the conversion formula, we can use the ratio in Equation (8.5). The interpretation of the new loadings, λ'_{t1}, will be in relation to 0. So, to obtain the original loading estimates, λ_{t1}, from the model with the first two time points constrained, we can simply divide each new loading, λ'_{t1}, by the second new loading $\lambda_{t1} = \lambda'_{t1} / \lambda'_{21}$. The process is simple to reverse to get the new loadings from old loadings, where the last loading from the model that constrains the first two loadings, $\lambda_{T-1,1}$, must be used as the denominator to obtain the loadings for the model that constrains the first and last loadings, $\lambda'_{t1} = \lambda_{t1} / \lambda_{T-1,1}$.

For the model depicted in Figure 8.12, the latent basis factor represents both linear and nonlinear change, and it is impossible to distinguish the two with just a single factor (Stoomiller, 1995). We cannot know if the standard linear growth curve would have been sufficient to account for the data. A modified version of the model addresses this by specifying a latent basis factor and linear factor together. In such a model we have η_0, the intercept factor; η_1, the linear effect; and η_2, the nonlinear effect. The loadings for the linear factor are specified as usual, with $\lambda_{t1} = 0, 1, 2, 3, \ldots T-1$ for an uncentered model. Two of the loadings for η_2 are constrained for identification, with the remaining freely estimated. The model bears some resemblance to the piecewise acceleration (deceleration) model, because the nonlinear factor provides information about the degree to which the trajectory departs from linear growth.

Example 8.5. A Latent Basis Model

Following the specification of the model depicted in Figure 8.12, a latent basis model with a single nonlinear factor was estimated to fit the BMI data over six waves. The initial model did not converge, so a second model was tested constraining residual variances equal across time. The resulting model had a considerably higher chi-square than previous models, $\chi^2(17) = 897.835$, $p < .001$. Still the model did not fit poorly, CFI = .985, RMSEA = .099. With the first two loadings constrained to 0 and 1, the remaining loadings for waves 3 through 6 were estimated to be 2.680, 4.470, 6.606, and 8.202. Because these loadings are larger than what would be expected with a linear increase, it suggests that change in BMI had an increasing rate over time. The mean of the intercept factor was 27.262, which is similar to the baseline observed value for BMI. The mean of the latent basis factor was .083, which indicates that the average nonlinear increase over the full period of the study was less than a tenth of a point per wave. The difference between the observed means at the first and last time point divided by the last loading is very close to this value, (27.864 – 27.176)/8.202 = .084. The loadings, $\lambda_{t1} = 0$, 1, 2.680, 4.470, 6.606, 8.202, can be entered into the following equation to plot an average trajectory over time: $\hat{y}_t = \alpha_0 + \alpha_1\lambda_{t1} = 27.262 + .083\lambda_{t1}$. Figure 8.13 shows the estimated line. For the plot, the *x*-axis uses *t* rather than estimated loadings to allow the line to be nonlinear.

A second model was tested constraining the first and last loadings to 0 and 1. This model also required equal variances over time. The chi-square and *df* were identical to the previous model. The estimated loadings were 0, .122, .327, .545, .805, and 1. Because the models are equivalent, the new loadings can be used to find any of the loadings estimated in the first model. For example, the fourth loading in the original model was equal to $\lambda_{41} = \lambda'_{41} / \lambda'_{21} = .545 / .122 = 4.407$. The mean of the basis factor was .683, $p < .001$. Although statistically equivalent, the interpretation of the basis factor mean differs because of the different interpretation of the loadings. The mean for this version of the model represents the average cumulative proportional change over the course of the study.

A third model was tested with a linear factor and a latent basis factor. The loadings for the linear factor were $\lambda_{t1} = 0, 1, 2, 3, 4, 5$, and the loadings for the latent basis factor (λ_{t2}) set the first two loadings equal to 0 and 1 and freely estimated loadings for the remaining time points. As was the case with the prior latent basis models, homogeneity of variance had to be assumed to obtain a solution. Still, there were convergence problems with this model and starting values were needed to obtain a solution. Convergence was sensitive to the specific starting values and centering scores on the linear factor did not help. Although a solution was printed, a warning was issued that standard errors might be problematic. Indeed, standard errors appeared to be unusually large. The model fit was good $\chi^2(13) = 170.144$, $p < .001$, CFI = .997, RMSEA = .048 and parameter estimates appeared to be reasonable. The estimated loadings for the latent basis factor were 0, 1, 1.476, 1.813, –.661, –4.123, indicating an increasing trend with deceleration relative to the linear slope in the last two time points. Although the last loading is unusually large in magnitude and is likely problematic, the pattern generally fits with the expected trend based on prior analyses. The estimate of the average of the linear slope factor was .169, ns, and the estimate of the average of the basis factor was .034, ns. Both of these values are similar to estimates from prior models, but significance tests are likely incorrect because the standard errors are large relative to those obtained with other models.

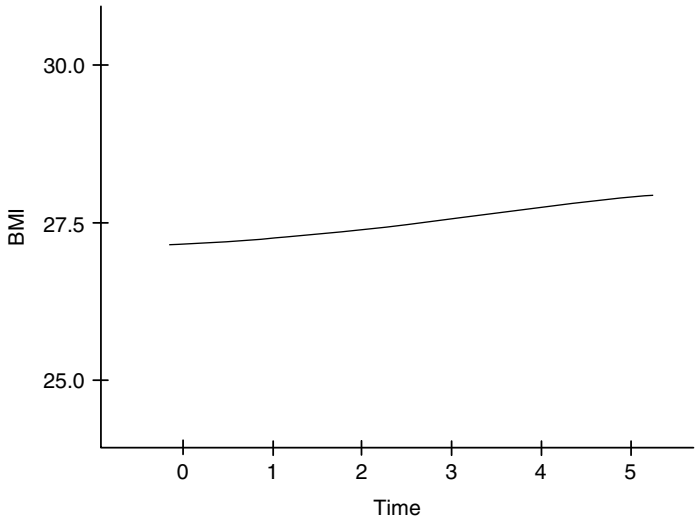

Figure 8.13. Plot of Average Line from the Latent Basis Model of BMI.

Comments

The latent basis models provide a potentially powerful method of exploring functional form of the data. As with most data-driven analyses, proper cognizance is needed that the method is fitting a model to the data rather than the data to a model. There is always the potential that the model will be "overfit" to the data, with stochastic fluctuations at specific occasions affecting the shape of the function, perhaps leading to incorrect conclusions. An important precaution is to ensure that the results of a model with only a latent basis factor are not simply driven by a linear trend. The mean of the basis factor will be significant even if only a linear trend is present, and, without a significance test to determine whether the basis factor adds information beyond a linear trend, interpretation of results will remain ambiguous (Stoolmiller, 1995). Such a test can be achieved either by examination of the Wald ratio for the basis factor mean or through a likelihood ratio test comparing a model that includes a linear slope with and without the latent basis factor. Either involves at least an initial test of a latent basis model that also includes a linear factor.

The latent basis model is essentially a simple spline regression model and could be extended in several ways. Spline regression models use smoothing and polynomial expansions to improve estimates, and both are possible extensions within the SEM framework. Such extensions have yet to be explored in this context, however. Differencing and autocorrelations can be used to smooth curves, a topic that will be addressed further in Chapter 11 on time series models. Complex constraints on loadings potentially could be used to produce polynomial spline models.

Power and the Number of Time Points

General issues regarding power with linear models were discussed in Chapter 7, but a few additional issues regarding power with nonlinear models deserve mention. Estimation of nonlinear model is more likely to result in convergence problems. Welch (2007) found that convergence was best for the latent basis model compared with other nonlinear models when fitting a logarithmic curve, although within the conditions of the study even this convergence rate was only 60%. Convergence is less likely when the model is incorrect and, given that correct models were not compared to incorrect models, it is unknown whether correct polynomial models would have higher convergence rates than latent basis models.

Increasing the number of time points is likely to increase power to detect polynomial change in growth (fixed effects) up to approximately seven time points (Raudenbush & Xiao-Feng, 2001), where the increases in power appear to reach an asymptote. Increases in sample size seem to have a more potent effect on power than increases in the number of time points, although the two factors interact somewhat. For a medium effect size and three time points, a minimum of approximately 300 cases was needed for sufficient power to detect fixed effects (average slopes). The minimum number of cases declined only slightly, however, if more time points were included. These authors did not examine power to detect random effects of polynomials (variance of the nonlinear slope), and much less is known about the required sample size and number of occasions for sufficient power to detect random effects for polynomials. But the requirements for the number of time points and sample size no doubt are likely to be higher than they are for fixed effects, given data requirements are more stringent for random effects for linear factors. Power to detect random effects will be a function of the reliability of the estimate, with higher between-case and lower within-case variability increasing power to detect significance.

Extensions

This chapter has focused on many of the fundamental concepts associated with testing nonlinear growth curve models, but these concepts can be applied to more elaborate models. Incorporation of multiple indicators at each occasion, individually-varying time points, and binary and ordinal variables are all possible. An introduction to these features was given in the previous chapter on linear growth curves, and the same features can be readily transferred to the nonlinear models discussed in the present chapter. Boker and colleagues (Boker, Neale, & Rausch, 2004), for example, show how to test latent differential models with multiple indicators. To cover the most essential concepts, I scarcely mentioned inclusion of covariates, but both time-invariant and time-varying covariates can be added to any of the nonlinear models discussed in this chapter. Steele and Ferrer (2011) apply the damped oscillator model to illustrate coupled effects involving two corresponding dynamic systems models. Linear piecewise models can be extended to include nonlinear components (Cudeck & Harring, 2007), greatly expanding their potential for testing a number of sophisticated hypotheses about change. Such models can be used as an alternative method of handling asymptotes, combining linear and nonlinear components, or specifying different nonlinear functions.

Comments

There are many nonlinear functions to choose from, but there is no universal guidance I can give on which function is best to use. That should depend on theory. The more that models can be guided by theory, the more confident we can be about results when they are supported by a particular model. That said, it is also important to consider the possibility that there may be trends other than a linear one present in the data, moving beyond the default assumption that change processes will be linear. The latent basis is one method of exploring the function form of the data, but there are many plotting approaches that can be helpful (see Singer & Willett, 2003, for several good suggestions). Individual plots are useful for visualizing individual variability, although generation of plots is often inconvenient because it usually must take place outside of the SEM software program. Grimm and colleagues (Grimm, Steele, Ram, & Nesselroade, 2013) discuss exploratory approaches to modeling nonlinear change, including Tuckerized curves, latent basis models, and exploratory SEM methods. When data have guided decisions about the functional form of the trajectory, it is doubly important that replication with other samples be obtained before concluding with any degree of certainty that the appropriate functional form has been chosen.

Theory is particularly important in light of the fact that many nonlinear models can also be accounted for by autoregressive processes. Bianconcini (2012) shows that quadratic latent growth curves and latent basis curves can be equivalently stated as linear model with autoregressive processes. Similarly, linear oscillators can also be seen as autoregressive processes (Deboeck, 2010). Although these models have different theoretical bases, they cannot be distinguished empirically. Consequently, other evidence and theory must be marshalled to justify one over the other.

Recommended Readings

Hedecker and Gibbons (2006, chapter 5) and Singer and Willet (2003, chapter 6) have very useful discussions of nonlinear trajectories in the context of multilevel regression models. Grimm and Ram (2009) illustrate and compare nonlinear growth curve modeling

with structural equations and multilevel regression. Highly informative discussions of how theories about nonlinear change can be modeled with growth curve approaches are given by Burchinal and Appelbaum (1991), Cudeck and Harring (2007), and Ram and Grimm (2007), all of which have multiple illustrations. For differential structural equation models, a good place to begin is with Boker and Nesselroade (2002), who have the gentlest introduction to the rationale of the approach. A second step would be to read Boker (2007) for additional detail. More on the rationale and mathematical detail of oscillating systems and dynamic models can be found in Beltrami (1998).

Notes

1 Piecewise growth curve models have been referred to by a dizzying array of terms, including discontinuous change models, change-point models, multi-phase models, and growth curve models with knot points, transition points, or turning points. There are really no practical differences in the models that fall under these labels, and the goal in each case is to obtain separate estimates of linear trajectories for different portions of the longitudinal study.

2 McArdle and Epstein (1987) in their early paper on growth curves proposed a model they referred to as a "stopping" model in which a single slope factor is defined by increments in loadings up to a particular change point, after which the loading values remain constant. Their slope factor is much like the first slope factor in the standard piecewise model. They also suggested extensions of such a model could be used to analyze regression discontinuity designs.

3 The observant reader may notice some resemblance of the quadratic effect to the increment (decrement) model. Although they both represent an effect that occurs over and above the linear effect in a sense, the increment (decrement) model represents an additive rate for a specific subset of occasions that occurs over and above the average linear rate across all occasions, whereas the quadratic effect represents a multiplicative effect that occurs over and above the linear rate across all occasions.

4 Briefly, the instantaneous rate of change is given by the first derivative, which provides information about the slope of any tangent line to the curve. The slope of the tangent line gives the change in y for an infinitesimally small increment in x (i.e., with limit of 0) for any chosen point along the curve (i.e., at any value of x).

5 The values obtained with the rescaled orthogonal polynomial model do differ very slightly from the other two centered models. This is due to rounding of the fractional time codes to only three decimal places. If the values used in the analysis were carried out to further decimal places, the estimates would match even more closely.

6 I should point out that the equivalences of the several centered coding schemes depicted in Figure 8.2 may be less of a clear case when missing data are present, individually varying time points are used, or there is an unequal spacing of intervals.

7 The factorization process, referred to as orthonormalization, is actually rather simple with the aid of a program that can perform matrix operations such as procedures available in many standard statistical packages. Hedeker and Gibbons (2006) provide PROC IML code for SAS, for example.

8 Not all SEM programs have features for complex modeling constraints, though an increasing number are adding this feature. Fewer (e.g., Mplus, Lisrel, Mx) allow a new parameter to be estimated based on these constraints. Grimm and Ram (2009) show how phantom variables can be used when new parameters cannot be estimated.

9 These more complex functions are sometimes referred to as dynamic or "dynamical" systems. The terms are not always specific, however, and can refer to a variety of mathematical models other than oscillators.

10 To be consistent with the rest of the book, my notation differs from that used by Boker and colleagues (e.g., Boker, 2001), who state the damped oscillation model in terms of the second derivative of the outcome $x(t)$ with respect to time, t, on the left, the regression coefficient representing damping as z, and the regression coefficient representing the squared frequency estimate, η,

$$\frac{d^2x(t)}{dt^2} = \zeta\frac{dx(t)}{dt} + \eta x(t) + e(t).$$

11 The term "latent basis" comes from the basis spline regression model or "b-spline" model, which estimates a flexible function to fit the data. The term "basis" is a linear algebra term for independent vectors of numbers that can be used to produce all other dependent vectors. Within the SEM literature, the term "basis" is sometimes used loosely to refer to any latent factors in a growth curve model.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, NJ: Sage.

Beltrami, E. J. (1998). *Mathematics for dynamic modeling*. New York: Academic Press.

Bianconcini, S. (2012). Nonlinear and quasi-simplex patterns in latent growth models. *Multivariate Behavioral Research*, *47*, 88–114.

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*, 30–52.

Blozis, S. A., & Cudeck, R. (1999). Conditionally linear mixed-effects models with latent variable covariates. *Journal of Educational and Behavioral Statistics*, *24*, 245–270.

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological Methods*, *9*, 334–353.

Blozis, S. A., Harring, J. R., & Mels, G. (2008). Using LISREL to fit nonlinear latent curve models. *Structural Equation Modeling*, *15*, 346–369.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGrawHill.

Boker, S. M. (1996). *Linear and nonlinear dynamical systems data analytic techniques and an application to developmental data* (Doctoral dissertation). University of Virginia.

Boker, S. M. (2001). Differential structural equation modeling of intraindividual variability. In L. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 3–28). Washington, DC: American Psychological Association.

Boker, S. M. (2007). Specifying latent differential equations models. In S. M. Boker & M. J. Wenger (Eds), *Data analytic techniques for dynamical systems in the social and behavioral sciences* (pp. 131–159). Mahwah, NJ: Lawrence Erlbaum Associates.

Boker, S. M., Deboeck, P. R., Edler, C., & Keel, P. K. (2010). Generalized local linear approximation of derivatives from time series. In S.M. Chow, E. Ferrer, & F. Hsieh (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue. The Notre Dame series on quantitative methodology* (pp. 161–178). New York: Routledge.

Boker, S. M., & Graham, J. W. (1998). A dynamical systems analysis of adolescent substance use. *Multivariate Behavioral Research*, *33*, 479–507.

Boker, S., Neale, M., & Rausch, J. (2004). Latent differential equation modeling with multivariate multi-occasion indicators. In K. van Montfort, H. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models* (pp. 151–174). Dordrecht, The Netherlands: Springer.

Boker, S. M., & Nesselroade, J. R. (2002). A method for modeling the intrinsic dynamics of intraindividual variability: Recovering the parameters of simulated oscillators in multi–wave panel data. *Multivariate Behavioral Research*, *37*, 127–160.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: John Wiley & Sons.

Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions 2* (pp. 171–198). Amsterdam: North-Holland.

Browne, M. W., & Du Toit, S. H. (1991). Models for learning data. In L. M. Collins, & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 47–68). Washington, DC: American Psychological Association.

Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development*, *62*, 23–43.

Choi, J., Harring, J. R., & Hancock, G. R. (2009). Latent growth modeling for logistic response functions. *Multivariate Behavioral Research*, *44*, 620–645.

Cohen, P. (Ed.). (2008). *Applied data analytic techniques for turning points research*. New York: Routledge.

Cook, T. D., & Campbell, D. T. (1979). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, *58*, 615–637.

Cudeck, R., & du Toit, S. H. C. (2002). A version of quadratic regression with interpretable parameters. *Multivariate Behavioral Research*, *37*, 501–519

Deboeck, P. R. (2010). Estimating dynamical systems, derivative estimation hints from Sir Ronald A. Fisher. *Multivariate Behavioral Research*, *43*, 725–745.

Deboeck, P. R., Montpetit, M. A., Bergeman, C. S. & Boker, S. M. (2009). Describing intraindividual variability at multiple time scales using derivative estimates. *Psychological Methods*, *14*, 367–386.

Fisher, R. A. (1921). Studies in crop variation i: An examination of the yield of dressed grain from broadbalk. *Journal of Agricultural Science*, *11*, 107–135.

Flora, D. B. (2008). Specifying piecewise latent trajectory models for longitudinal data. *Structural Equation Modeling*, *15*, 513–533.

Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in Mplus and SAS. *Structural Equation Modeling*, *16*, 676–701.

Grimm, K. J., Steele, J. S., Ram, N., & Nesselroade, J. R. (2013). Exploratory forms of latent growth modeling in structural equation modeling. *Structural Equation Modeling*, *20*, 568–591.

Grimm, K. J., Zhang, Z., Hamagami, F., & Mazzocco, M. (2013). Modeling nonlinear change via latent change and latent acceleration frameworks: Examining velocity and acceleration of growth trajectories. *Multivariate Behavioral Research*, *48*, 117–143.

Harring, J. R., Cudeck, R., & du Toit, S. H. C. (2006). Fitting partially nonlinear random coefficient models as SEMs. *Multivariate Behavioral Research*, *41*, 579–596.

Hays, W. L. (1994). *Statistics*. Orlando, FL: Harcourt Brace.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York: Wiley.

Jenss, R. M., & Bayley, N. (1937). A mathematical method for studying the growth of a child. *Human Biology*, *9*, 556–563.

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, *48*, 370–397.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. Nesselroade & R. B. Cattel (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). New York: Plenum.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*, 110–133.

Meredith, W., & Tisak, J. (1984). Statistical considerations in Tuckerizing curves with emphasis on growth curves and cohort sequential analysis. Paper presented at the annual meeting of the Psychometric Society, Santa Barbara, CA.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122.

Preacher, K. J., & Hancock, G. R. (2012). On interpretable reparameterizations of linear and nonlinearlatent growth curve models. In J. R. Harring & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 25–58). Charlotte, NC: Information Age Publishing.

Ram, N., & Grimm, K. J. (2007). Using simple and complex growth models to articulate developmental change: Matching method to theory. *International Journal of Behavioral Development*, *31*, 303–316.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrika*, *51*, 83–90.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, NJ: Sage.

Raudenbush, S. W., & Xiao-Feng, L (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387–401.

Raykov, T. (1998). Satisfying a simplex structure is simpler than it should be: A latent curve analysis revisit. *Multivariate Behavioral Research*, *33*, 343–363.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, *10*, 290–300.

Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York: Oxford University Press.
Steele, J. S., & Ferrer, E. (2011). Latent differential equation modeling of self-regulatory and coregulatory affective processes. *Multivariate Behavioral Research*, *46*, 956–984.
Stimson, J. A., Carmines, E. G., & Zeller, R. A. (1978). Interpreting polynomial regression. *Sociological Methods & Research*, *6*, 515–524.
Stoolmiller, M. (1995). Using latent growth curve models to study developmental processes. In J. M. Gottman (Ed.), *The analysis of change* (pp. 103–138). New York: Erlbaum.
Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, *23*, 19–23.
Voelkle, M. C., & Oud, J. H. L. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *British Journal of Mathematical and Statistical Psychology*, *66*, 103–126.
Welch, G. W. (2007). *Model fit and interpretation of non-linear latent growth curve models* (Unpublished doctoral dissertation). University of Pittsburgh.
Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*, 363–381.
Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, *30*, 16–28.

Keywords
nonlinear, piecewise, change-point, latent growth curve model, multilevel, growth curve, latent curve model

9 Latent Difference Score Models

The latent difference score model (McArdle & Hamagami, 2001; McArdle & Nesselroade, 1994) is an extension of the simple difference (or gain) score approach to the analysis of change. Difference scores are computed by subtracting an earlier observed value from a later observed value. The latent difference score model extends this simple bivariate difference to consecutive intervals over many waves of a study. The result is a flexible modeling approach to the investigation of change. In this chapter, I will introduce the basic concepts and specifications for the model, discuss some of the basic elaborations, consider how the model relates to other longitudinal models, and explore issues related to the inclusion of covariates. I begin with a brief review of some of the fundamental concepts of differences discussed at greater length in Chapters 3 and 4.

Review of Fundamental Concepts

The central notion of the latent difference score model is that we can represent the difference between two consecutive measurements within the autoregressive structure. The residual from a simple autoregression is equal to the difference score if y_t is regressed on y_{t-1} with a regression slope equal to 1 (i.e., a perfect relationship). For a simple autoregression written in terms of a simple structural model, the equation is

$$y_{ti} = \beta_{t,t-1} y_{t-1,i} + \zeta_{ti} \tag{9.1}$$

The subscript i is added to emphasize individual scores – a point I will return to shortly. When $\beta_{t,t-1}$, the autoregression coefficient, is replaced by 1, the equation can be rearranged to show that the residual can be expressed as the difference between y_{ti} and $y_{t-1,i}$.

$$\begin{aligned} y_{ti} &= \beta_{t,t-1} y_{t-1,i} + \zeta_{ti} \\ y_{ti} &= (1) y_{t-1,i} + \zeta_{ti} \\ \zeta_{ti} &= y_{ti} - y_{t-1,i} \end{aligned} \tag{9.2}$$

With the equation stated in this manner, the average residual is equal to the average of the difference scores across all of the cases, $E(\zeta_{ti}) = E(y_{ti} - y_{t-1,i})$. We know also that the average of the differences equals the difference of the averages, $E(y_{ti} - y_{t-1,i}) = E(y_{ti}) - E(y_{t-1,i})$. Thus, a test of whether this difference is significant is equivalent to a repeated measures *t*-test or ANOVA (see Chapter 3).

What this restatement tells us is that by placing particular constraints on autoregression coefficients in a structural model, we can derive difference scores. If we relabel the residual ζ_{ti} from Equation (9.2) as the difference, $\Delta y_{t,t-1,}$ and insert the value back in to Equation

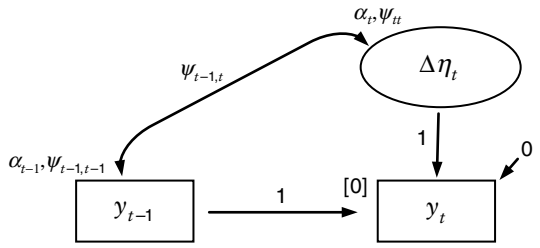

Figure 9.1 Latent Difference Score Model for Two Waves.

(9.1), then, assuming $\beta_{t,t-1}$ is equal to 1, the score at the subsequent time point is equal to the score from the prior time point plus the difference.

$$\begin{aligned} \Delta y_{t,t-1} &= y_{ti} - y_{t-1,i} \\ y_{ti} &= y_{t-1,i} + \Delta y_{t,t-1,i} \end{aligned} \tag{9.3}$$

Equation (9.3) is the foundation of the latent difference score model, because this formulation can be exploited to specify the structural model illustrated in Figure 9.1. In the figure, the difference score factor, $\Delta\eta_t$, is equal to the simple difference score $\Delta y_{t,t-1,i}$. The difference score factor is estimated by specifying a single indicator, y_{ti}, with its loading set equal to 1, that is predicted by the observed variable at the previous time point, $y_{t-1,i}$, with its path set equal to 1. After omitting the *i* subscript to be consistent with usage throughout the text, a restatement of the structural model for two waves as depicted in Figure 9.1 parallels Equation (9.3).

$$\begin{aligned} y_2 &= (\beta_{21})y_1 + (\lambda_{22})\Delta\eta_2 + \zeta_2 \\ y_2 &= (1)y_1 + (1)\Delta\eta_2 + 0 \\ y_2 &= y_1 + \Delta\eta_2 \end{aligned}$$

The final equation states that the observed variable at Time 2 is equal to the sum of the observed variable at Time 1 and the difference between Time 2 and Time 1.

In practice, because we are interested in the average difference at each time point, the mean structure will be estimated to obtain α_3, the mean of the difference factor. For identification purposes, however, the intercept for the autoregression, α_2, will be set equal to 0.[1] The model may also be stated more formally in the LISREL single-indicator form, which specifies a latent variable for each measured variable used in the structural component. For this specification, y_1 and y_2 would be single indicators of factors η_1 and η_2 with loadings set equal to 1 and latent difference score factor $\Delta\eta_2$ representing the difference between the two factors. Most software programs do not require this type of specification, however, and the observed variables can be used directly in specifying autoregressive paths.

Although this structural equation model may seem elaborate, the difference score factors are simply a method for deriving difference scores. They can be conveniently incorporated into a larger structural equation model, however, making it possible to test a variety of more sophisticated hypotheses. The model can be extended to multiple indicators to remove measurement error at each time point, but this does not affect the mean difference. Even though the latent difference score factor is sometimes described as embodying "true differences," the difference between two observed scores is equal to the

difference between two true scores on average, $E(y_t - y_{t-1}) = E(T_t - T_{t-1})$, where T used here is the true score. Estimates obtained from a latent difference factor, therefore, equal estimates obtained from a simple calculation of differences from observed scores. This equality holds because measurement error is equal to 0 on average (see Chapter 4 for additional detail on these points).

Latent Difference Score Model for Several Occasions

Model Specification

Additional time points can be added to the model following the general structure illustrated for two time points. Figure 9.2 illustrates a version of the model for six waves. At each occasion except for the first, a latent variable representing the difference score is specified with the loading set equal to 1. The autoregressive parameters between each successive observed variable are set equal to 1, and the disturbances, ζ_2 through ζ_6, are set equal to 0. The variances and means of each of the exogenous variables (y_1 and $\Delta\eta_2$ through $\Delta\eta_6$) are estimated as are their covariances, but constraints on the variances or covariances may be necessary if empirical underidentification problems are encountered.

The degrees of freedom for the latent difference score model follow the usual formula for a structural model with means.

$$df = \left[\frac{J(J+1)}{2} + J\right] - q$$

There are $[J(J+2)]/2$ variance and covariance elements and J means in the observed data. The number of free parameters is equal to q, which can be stated generally as

$$q = \left[2T + \left(\frac{J(J-1)}{2}\right)\right]$$

for the latent difference score model when there are T time points. Because autoregressive paths and structural disturbances are constrained, the free parameters consist of $2T$ means and variances for the exogenous variables (y_1 and $\Delta\eta_2$ through $\Delta\eta_6$ for the model in Figure 9.2) and $[J(J-1)]/2$ covariances among these variables. For a model with six time points, the degrees of freedom are

$$df = \left[\frac{J(J+1)}{2} + J\right] - \left[2T + \frac{J(J-1)}{2}\right] = \left[\frac{6(6+1)}{2}\right] - \left[(2)(6) + \frac{6(6-1)}{2}\right] = 0$$

The model is therefore just identified for six time points, and, in fact, given the same model specifications for any other number of time points will be just identified.

Several possible strategies can be used to increase likelihood of convergence if problems arise. One possible constraint is to require variances of the difference factors to be equal over time. Another approach, which can be used as an alternative or as an additional constraint if necessary, is to set all or some of the covariances among exogenous variables to be equal to 0. These constraints can be imposed without affecting mean difference estimates. The covariance between the baseline measure and latent difference score factors are likely to be of more theoretical interest, and covariances among difference score factors are likely to be of lesser theoretical interest.

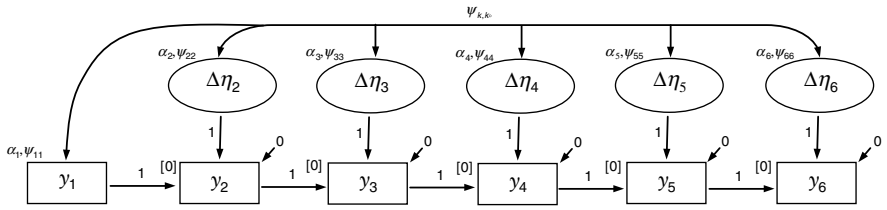

Figure 9.2 Latent Difference Score Model for Six Waves.

Interpretation

The mean for each latent difference variable, α_2 to α_T, is the average of the two consecutive difference scores or, alternatively, the difference between their respective means, $\bar{y}_t - \bar{y}_{t-1}$. The variance of each latent difference variable, $\psi_{22} - \psi_{TT}$, gives the variability of the difference scores across cases. In growth curve terms, factor variances can be conceptualized as random effects of each bivariate difference (Jajodia, 2012). Either the means or the variance of the difference factors can be constrained to be equal to obtain an omnibus test of whether these parameters differ over time. When the means are set equal, the single estimate of the mean of the difference factor gives the average change at each interval.[2] The average difference across all of the contiguous waves should be recognizable as the interpretation given to the slope in a linear growth curve analysis (Chapter 7 in this volume; Bollen & Curran, 2006). Indeed, an estimate of the average difference across waves in the latent difference model and average slope from the linear latent growth curve model will be nearly identical. A test of the equality of mean differences across all intervals is the same as the test of the equality of all means obtained with ANOVA (see Chapter 3 for elaboration of this point).

Yet another connection to familiar longitudinal models can be made. The latent difference score model closely relates to the perfect simplex model. Recall from Chapter 5 that the perfect simplex model is a series of autoregressions with the observed score at each time point predicted by the observed score at the prior time point. The latent difference score model is equivalent to the simplex model if its autoregressive paths have been set equal to 1. Inclusion of the intercept in the simplex model then obtains the mean difference for adjacent scores. This can be seen by inspecting the common formula for the intercept under the condition that the autoregression coefficient is equal to 1.

$$\beta_0 = \bar{y}_t - \beta_{t,t-1}\bar{y}_{t-1}$$
$$\beta_0 = \bar{y}_t - (1)\bar{y}_{t-1}$$
$$\beta_0 = \bar{y}_t - \bar{y}_{t-1}$$

Thus, the intercepts in a restricted simplex model with autoregressive paths are set equal to 1. The variances of the disturbances in this restricted simplex model give the variance of the difference scores, even though its expected value is equal to 0 when the intercept is included in the model. The latent difference score model is therefore a special case of the simplex model where each difference factor is a phantom variable that estimates the autoregressive intercept and variance of the disturbance.

The equivalence to the simplex model suggests that the latent difference score model can be modified further by adding a quasi-simplex structure. The quasi-simplex (or

quasi-Markov) model is a variant of the perfect simplex that adds a latent variable for each measured variable (see Chapter 5 for more description). The measurement residual for each single-indicator factor can be estimated if there are four time points and two residuals are constrained to be equal. The advantage of the quasi-simplex specification is the estimation of measurement error derived from repeated measurements (McArdle, 2001). The estimation of measurement error will impact the correlational structure but will not affect mean differences. To identify the quasi-simplex model, the first factor variance is set equal to 1 and loadings for observed scores at each time point are set equal to 1, freely estimating all measurement residuals except for two (usually the first and last time points). Four or more time points may be required to identify this version of the latent difference score model, as this is the minimum for the quasi-simplex model. Estimating measurement residuals will not affect difference factor means but may reduce difference factor variance. This increased precision may come at the cost of convergence difficulties due to empirical underidentification, a problem associated with quasi-simplex models (Marsh, 1993).

Example 9.1: Latent Difference Score Model

To illustrate the latent difference score model, biennial differences in body mass index (BMI) were examined in the health and aging study. Syntax and data sets used in the examples are available at the website for the book. The model parameters were specified consistent with Figure 9.2. The mean of the observed BMI score was estimated along with the means of all the latent difference factors, $\Delta\alpha_2$ through $\Delta\alpha_6$. Variances and covariances for the observed BMI at Time 1 and all latent difference factors were estimated. The autoregression coefficients, $\beta_{t,t-1}$, and loadings, λ_{tt}, for each difference factor were set equal to 1, with the autoregressive disturbances constrained to be equal to 0. All covariances among exogenous variables (y_1 with $\Delta\eta_2 - \Delta\eta_6$) were set equal to 0 in the initial model.

The model was just identified. The estimate of the mean of BMI at the first wave was 27.176, which, as expected, is the same as the observed mean value. The means for the difference factors for Waves 2 through 6, were .181, .155, .245, .060, and .047, respectively, and corresponded nearly perfectly with differences between the means calculated from the raw scores. The means of the difference factors were constrained to be equal in a subsequent model, $\chi^2(4) = 43.909$, $p < .001$, suggesting the mean differences were not equal across all of the waves. Several plotting strategies could be used to aid in the interpretation of latent difference score models but will not be illustrated here (for examples, see Boker & McArdle, 2005; Jajodia, 2012).

Binary Variables and Ordinal Variables

Latent difference score models can be applied to binary or ordinal observed variables using the appropriate estimator (e.g., full ML, WLSMV). We saw in Chapter 3 that a repeated measures test of two proportions can be conducted that is equivalent to the McNemar test using the latent difference framework. The structure of the model generally follows Figure 9.2, but, depending on the software implementation, it may be convenient to construct a latent variable for each occasion in order to restrict the autoregression intercepts to be equal to 0. Constraints on the measurement thresholds will be required for identification of the factors for each occasion. With binary variables, this constraint typically involves setting the single threshold equal to 0. With ordinal variables, the constraint typically involves setting the first threshold equal to 0 and the remaining thresholds equal across time points. Constructing a latent variable corresponding to the first observed

variable is not necessary, because it serves as an exogenous variable in the model. Overall the model requires more parameters to be estimated than the model for continuous measures when multiple thresholds are required, and additional constraints may be required for convergence.

With noncontinuous variables, latent difference factors represent change in the logit or probit function and require transformation using the appropriate cdf (see Chapter 1 for a review of this transformation). As a natural circumstance of contingency table analyses for binary and ordinal variables, the mean of each difference factor will not be precisely equal to the difference in proportions for binary or ordinal models (Agresti, 2013). Thus, researchers only should present the difference factor mean as an approximation of the mean difference. Constraint of the difference factor means can provide an omnibus test of all proportion differences. The latent difference score model with more than two binary or ordinal variables tests the same hypothesis as alternatives for matched pairs, such as Friedman's chi-square, Cochran–Mantel–Haenszel, or the loglinear model.

Example 9.2: Latent Difference Score Model with Binary Variables

I present just one example with binary variables to illustrate the basic specification, but the model is easy to extend to ordinal variables or different estimation approaches. A latent difference score model was specified for reported diagnosis of diabetes (yes or no) over six waves of the health and aging study. I estimated this model using WLSMV with delta parameterization, but other estimation approaches could be used. To specify the model, latent variables corresponding to each measured variable were constructed with each respective single loading set equal to 1. Other aspects of the model were specified as with the continuous model, with autoregressive paths set equal to 1, loadings for each difference factor set equal to 1, and factor means and variances estimated.

Covariances among latent variables are usually estimated, but this model would not be identified for binary variables. The covariances among exogenous factors do not impact on the mean estimates of the latent variables, so little critical information is lost if they are omitted. Because the covariances between the mean for the baseline measure and the difference factors were of most interest, I estimated the covariance between baseline diabetes and each of the difference factors but constrained the covariances among all exogenous difference factors to be equal to 0. The initial model did not converge, and a subsequent model was tested with the variances of the difference factors constrained to be equal. The fit of this model was excellent, $\chi^2(8) = 12.572$, $p = .128$, CFI = 1.000, RMSEA = .010, WRMR = .528, suggesting that the constraints did not detract substantially from the fit. The mean of the factor for Wave 1 was −1.336. The cumulative normal distribution (cdf) transformation of the mean gives an estimate of the proportion of those with diabetes equal to .091, which is the same as the observed proportion. The cdf transformations of the estimated means for the differences factors using .5 as the referent returned the proportions .037, .036, .040, .044, and .045. These values are slightly larger than the observed proportional differences of .016, .018, .022, .026, and .031, although both indicated increasingly larger differences in the proportion reporting diabetes diagnosis across consecutive waves. The discrepancy is a consequence of the fact that, for binary variables, marginal differences in proportions are not equivalent to the average of the individual differences. The results generally suggest an acceleration of the rate of diabetes diagnosis with increasing age, a trend observed with the growth curve model analyses of these data.

A subsequent model constrained the means of the difference factors to be equal for Waves 2 through 6 to investigate whether the changes in diabetes diagnosis was the same

across all of the two-year intervals. This model had a slightly higher chi-square value, $\chi^2(12) = 15.718$, which was not significantly higher than the chi-square in the previous model according to the likelihood ratio test appropriately weighted for WLSMV estimation, $\Delta\chi^2(4) = 2.636$, ns, indicating that the rate of increase in diabetes was constant across all time points.

Latent Difference Score Model with Added Intercept and Slope Factors

One elaboration of the basic latent difference score model is to specify an intercept and slope factor with the difference score factors serving as indicators of higher-order slope factor (e.g., McArdle, 2001). Depicted in Figure 9.3, the model resembles the linear latent growth curve model in which slope loadings are set according to a time metric (e.g., 0, 1, 2,…$T-1$).[3] For identification of the slope factor mean, the intercepts for each difference score factor is set equal to 0. Disturbances associated with each difference score factor are freely estimated but may be set equal over time. If there are sufficient degrees of freedom and the model is empirically identified, the covariances among disturbances [e.g., $\text{Cov}(\zeta_{t-1},\zeta_1)$] can be estimated as well (not shown in the figure).

Because the indicators for the slope factor represent difference scores between consecutive time points rather than single observed scores, the interpretation of the average slope is not the same as the interpretation of the slope factor in the linear growth curve model. If the difference scores are equal across intervals, the slope mean will be equal to 0. Such a result can be interpreted as a constant rate of change – the same as a linear increase or decrease over time (Grimm, An, McArdle, Zonderman, & Resnick, 2012). In fact, setting all of the loadings for the slope factor equal to 1 would be an alternative way to represent the linear growth curve model. The mean of the slope factor for such a model would be equal to the average of the $T-1$ difference factor means estimated by imposing longitudinal equality constraints on the means of the difference score factors. For a latent difference score model with added intercept and slope, a positive mean for a slope factor specified with linear increasing loadings (e.g., 0, 1, 2,…$T-1$) would indicate that the rate of change at each interval is increasing rather than constant. The latter interpretation should sound familiar as the concept of acceleration discussed in connection with nonlinear growth curve models (Chapter 8). Conversely, a negative slope would indicate a deceleration in the rate of change across occasions.

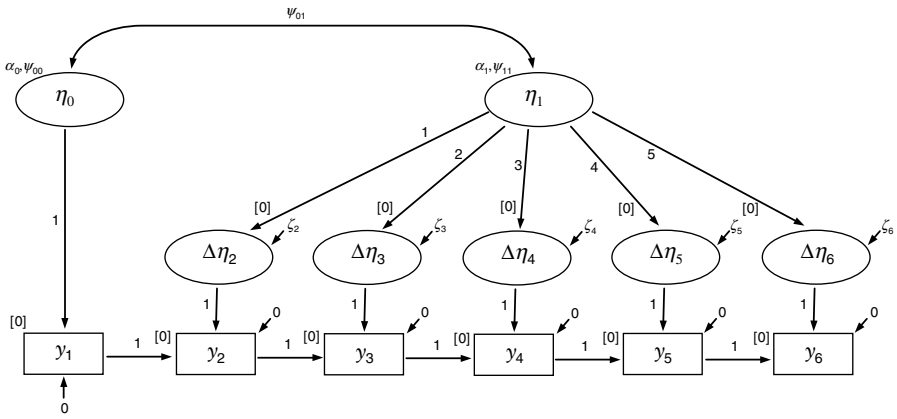

Figure 9.3 Latent Difference Score Model with Added Intercept and Slope Factors.

The latent difference score model with added intercept and slope factors can further be interpreted in light of the differential equation modeling perspective. If the true growth function is quadratic, for instance, the slope factor mean with linearly incremental loadings represents the average difference among the consecutive differences. Recall from Chapter 8 that this concept of the difference of the differences is the local approximation of the first derivative. The first derivative for a quadratic model gives the instantaneous rate of change that can be represented as the tangent line to the curve. For common loadings of 0, 1, 2, …, $T-1$, the slope factor would represent the instantaneous rate of change from baseline, and for centered loadings, the slope factor would represent the instantaneous rate of change from the midpoint of the study. Other loading schemes are possible, including freely estimating one or more loadings as in the latent basis growth curve model. The model can also be conceptualized in terms of time series models, because it can be viewed as a growth model with first-order differencing – the "integrating" component in the autoregressive, integrating, and moving-average (ARIMA) model (Box & Jenkins, 1970). First-order differencing, which is considered at greater length in Chapter 11, is used to detrend trajectories in time series models.

Example 9.3: Latent Difference Score with Added Intercept and Slope Factors

The previous example with continuous BMI was extended by adding an intercept and slope factor similar to that depicted in Figure 9.3. The intercept factor was specified by setting a single loading equal to 1, with the observed BMI score at baseline serving as its indicator. The slope factor was specified to have loadings equal to 0, 1, 2, 3, 4, and 5. The solution for the initial model with freely estimated means, variances, and covariances for both the intercept and slope factor was not admissible, with a negative variance estimate obtained for the slope factor. This is not unusual for this type of model. One method of addressing this problem is by standardizing the factor variance for the slope factor (McArdle, 2001). The constrained model did not fit well, $\chi^2(18) = 14{,}815.976$, $p < .001$, CFI = .749, SRMR = 1.845, RMSEA = .393. As expected, the intercept mean was equal to the baseline observed mean for BMI (27.176). The mean of the slope factor was significant, .030, $p < .05$, and this value suggested that the rate of change across successive intervals increased over time. Because the variance of the slope factor was set equal to 1, the standardized estimate was equal to the unstandardized estimate of the mean, a value that was of small magnitude.

Latent Dual Change Score Model

A variation on the latent difference score model is to add a lagged effect of the observed variable measured at the prior occasion on the difference factor. The equation for the lagged effect model (depicted in Figure 9.4) can be stated as $\Delta\eta_t = y_{t-1} - y_t = \beta_{\Delta\eta_t, y_{t-1}} y_{t-1}$, with the path coefficient $\beta_{\Delta\eta, y_{t-1}}$ representing the relation between a variable's prior value and its difference score taken over the subsequent interval. This coefficient is often called a *self-feedback effect* or "proportional change effect." For example, lower vocabulary scores measured immediately prior to kindergarten might be associated with greater vocabulary gains between pre-kindergarten and first grade, suggesting that students with initially lower scores experience more rapid vocabulary growth than those with initially higher scores. As considered in Chapter 4, this is a typical pattern reflecting regression toward the mean, with higher initial scores associated with decreases and lower initial scores associated with increases. The self-feedback effect also can be related to the

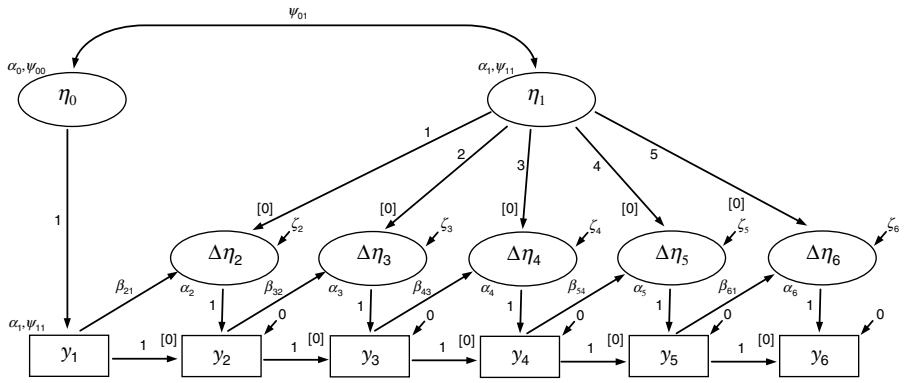

Figure 9.4 Dual Change Score Model.

dynamic systems concept of the damped oscillator model. When the self-feedback effect is incorporated into a latent difference score model with intercept and slope parameters, it is referred to as a *dual change score* model. The term "dual" is used because two components of change are represented – self-feedback and the rate of change estimated by the slope factor.

Example 9.4: Dual Change Score Model

Following Figure 9.4, a simple dual change score model of BMI was tested. In this model, each difference factor was regressed on the observed BMI score at each preceding time point. The model had an acceptable fit to the data according to most of the alternative fit indices, $\chi^2(10) = 2232.414$, $p < .001$, CFI = .962, SRMR = .045, but the RMSEA value of .204 indicated an unacceptable fit. Each of the self-feedback coefficients was significant, –.034, –.045, –.044, –.051, –.037, all *p*s < .001, signifying that higher scores on BMI at the preceding time point were associated with smaller increases in BMI across the subsequent interval. Alternatively stated, lower BMI scores at the preceding time point were associated with larger increases in BMI across the subsequent interval. These relationships are consistent with the pattern expected with regression toward the mean.

Covariates

Time-invariant and time-varying covariates can be used to explain difference scores or control for potentially confounding factors. Figure 9.5 illustrates both types of covariate in a single model. Generally, covariates serve to identify characteristics associated with individual variation in difference scores. For example, a researcher might be interested in how preschool vocabulary assessment is associated with reading performance differences across later grades. This would be an example of a time-invariant covariate, because the preschool vocabulary assessment does not change across grades.[4] In this example, the researcher would investigate whether children with higher vocabulary scores show larger increments from grade to grade throughout elementary school. Alternatively, a time-varying covariate, such as emotional adjustment at each grade, could be used to investigate whether better emotional adjustment is associated with larger increments in reading scores from grade to grade.

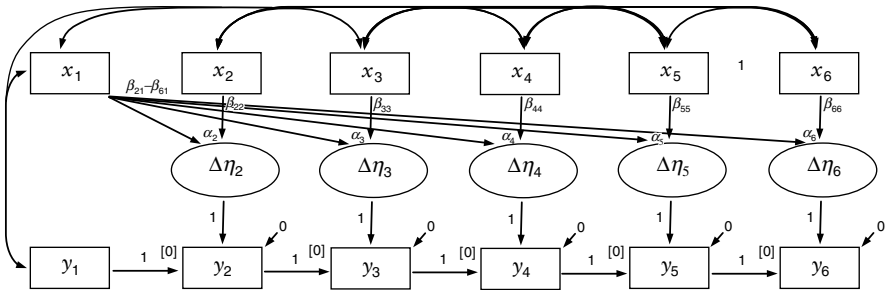

Figure 9.5 Latent Difference Score Model with Time-Invariant and Time-Varying Covariates.

Interpretation of covariate effects depends on the specification of the model. A time-invariant covariate predicting latent difference scores would provide information about which cases have lesser or greater increases or decreases across particular intervals. Similar to concerns raised with regard to latent growth curve models, means of the latent difference score factors will usually be more interpretable when the covariate has been centered. Such a model could be modified to examine the relationship of the covariate to average linear change across waves by constraining the path coefficients (β_{21} through β_{61} in Figure 9.5) and disturbances to be equal across time.

A time-varying covariate provides information about the synchronous relationship of the covariate to changes in the dependent variable at each interval. The primary motivation for such a model would likely be to investigate differences once the effect of the covariate is removed at each wave. As with time-varying covariates for growth curve models, the difference score must be interpreted in light of the scaling of the covariate. Centering the covariate will usually be more interpretable than not centering the covariate. With time-varying covariates, there are several possible options for scaling (see Chapter 7 for a discussion), such as computing differences from the baseline value and centering the baseline measure, $x^{*}_{ti} = x_{ti} - \bar{x}_1$. With this scaling, the estimate of the intercept for effect of the covariate on each difference score factor will be an adjusted mean of the difference score taken at the average value of the covariate relative to the average at baseline. Constraints can also be used in this model to estimate a single synchronous effect across waves if desired, and the longitudinal equivalence of the synchronous effect could be tested with a likelihood ratio test.

Paths for time-varying covariates also can be estimated as lagged effects, with the difference in the dependent variable predicted by the measurement of the predictor corresponding to the initial wave that the difference score is derived from (e.g., the covariate measured at Time 2 predicting the difference from Time 2 to Time 3). Each coefficient, in this case, would be equal to the difference score regression referred to as the unconditional change score model discussed in Chapter 4, with $y_t - y_{t-1}$ predicted by x_{t-1}. The addition of a dual change component (self-feedback or proportional change effect), involving prediction by the dependent variable measured at the prior time point, to a model in which the change score is predicted by the lagged covariate is unnecessary, however. The inclusion of a lagged covariate as a predictor of the change score while also controlling for the dependent variable measured at the prior time point is redundant with the unconditional change score model – the simple regression of the difference score regressed on the predictor measured at the prior time point (Chapter 4; see also Finkel, 1995; Plewis, 1985).

Including covariates in a latent difference model with added intercept and slope factors would provide tests of additional types of hypotheses. The effect of a time-varying

covariate on the slope factor with loadings set equal to 0, 1, 2, …, $T-1$ identifies characteristics of those more likely to show acceleration or deceleration in changes over time, thus explaining individual differences in the nonlinear effect. Constraining slope factor loadings to be equal to 1 would be an alternative strategy for estimating the effects of the covariate on average linear change across waves and would be conceptually equivalent to the effect of the covariate on the slope in a linear growth curve model. Inclusion of time-varying covariates in a model with added intercept and slope factor would allow for an estimation of the average intercept and slope factors adjusted for the covariate.

Example 9.5: Including Time-Invariant and Time-Varying Covariates

A basic latent difference score model with time-invariant covariate was tested to investigate whether age predicted differences in BMI across consecutive time points (following Figure 9.5 but omitting time-varying covariates). Age was centered to improve interpretability. Intercepts, disturbance variances, and residual covariances for the baseline measure and the difference factors were all freely estimated. The effect of age on the difference factor was allowed to vary at each time point (t_2 through t_6). The resulting model was just identified, so there was no fit information. Age was not associated with BMI scores at baseline, –.012, ns, but did significantly predict differences at each interval except between Waves 3 and 4, –.014, $p < .05$, –.021, $p < .001$, .002, ns, –.025, $p < .001$, –.027, $p < .001$. The standardized coefficients for the significant effects, which ranged between –.032 and –.056, were of small magnitude, however. The negative coefficients indicate that there were larger changes in BMI for younger participants. All of the residual variances remained significant after accounting for age, suggesting that there may be other variables that could explain individual variation in BMI difference scores. Note that if the intercepts, residual variances, and age effects on difference factors were held constant across time, results would closely approximate those obtained for the time-invariant covariates on the slope factor for the linear growth curve model in Example 7.3.

A second model demonstrated time-varying covariates by estimating the lagged effects of self-rated health measured at each $t_1 - t_5$ time point on differences in BMI at each subsequent interval $\Delta\eta_2$–$\Delta\eta_6$. To keep the example simple, age was omitted from the analysis. As with the previous model, intercepts, residual variances, and residual covariances estimated among baseline BMI and all difference factors. Self-rated health scores were centered around the self-rated health mean at baseline, $srh_t^* = srh_{ti} - \overline{srh}_1$. According to the alternative fit indices, this model fit the data well $\chi^2(24) = 153.885$, $p < .001$, CFI = .998, SRMR = .048, RMSEA = .032. The effect of self-rated health was significant at baseline –1.067, $p < .001$, indicating that those with higher self-rated health had lower BMI scores. The standardized coefficient for this path was –.239, suggesting a moderate effect. The lagged effects of self-rated health on the BMI difference factors were only significant for the first two intervals representing changes from t_1 to t_2 and from t_2 to t_3, .057, $p < .01$ and .–.056, $p < .05$. Interestingly, these effects were in opposite directions, suggesting that higher self-rated health was associated with greater increases in BMI across the first interval and smaller increases across the second interval. The remaining lagged effects were nonsignificant, –.026, .043, and .008.

Simultaneous Latent Difference Score Models

Two or more latent difference score models can be estimated simultaneously, also referred to as bivariate or multivariate latent difference score models (Figure 9.6). Let us first consider the basic specification of the latent difference score model to get a sense of the

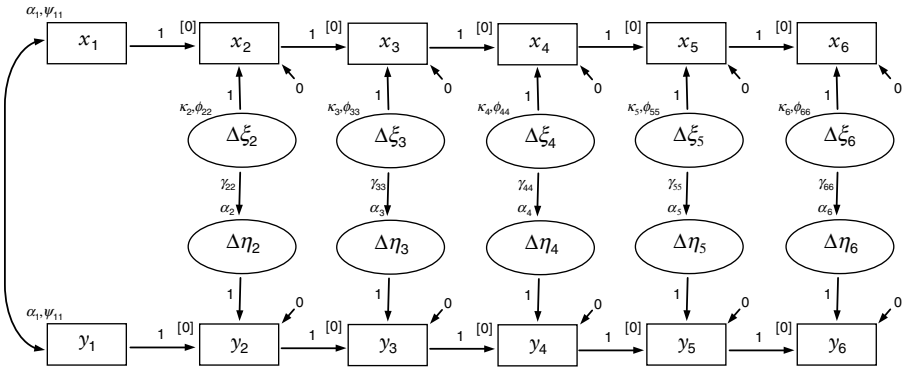

Figure 9.6 Simultaneous Latent Difference Score Model.

interpretation of such a model. Using the notion that a lagged time-varying covariate represents a difference score regression, the simultaneous latent difference score can be stated in terms of the effect of the difference score of a predictor variable on the difference score for an outcome, or Δy regressed on Δx. If path coefficients between the difference score factors at each occasion are held constant across time, then the simultaneous latent difference score model is equivalent to the fixed effects regression model covered in Chapter 5 (Allison, 2005).

Simultaneous models may include any of the alternative specifications described above, including the added intercept and slope factors, dual change components, lagged effects (referred to as *coupling effect*), and inclusion of covariates. Ghisletta and Ribaupierre (2005), for example, illustrate all of these components together, a specification dubbed the bivariate dual change score model (McArdle, 2001). Interpretation of any one of these effects becomes quite complex, as each kind of effect takes into account the other kinds of effects.

Latent Difference Scores with Multiple Indicators

Any of the above discussed models can be combined with multiple indicators at each time point (McArdle & Hamagami, 2004). The basic latent difference score model with multiple indicators is shown in Figure 9.7 as an illustration. As with other models, it is important to establish measurement invariance and, at minimum, longitudinal equality constraints are placed on intercepts and loadings. There are several advantages to the use of multiple indicators, but the advantages do not extend to the estimate of the average of the difference score factor, because the expected value of any observed variable is not biased by random measurement error.

Although multiple indicators have limited utility for estimating the average of the difference scores, there are several advantages to using multiple indicators at each occasion. One advantage is that correlated measurement residuals or method factors can be specified to account for method variance among repeated measurements of each indicator. These features will have little impact on mean differences, but they will improve estimates of the correlations between intercept and slope factors and between intercept and difference factors. Multiple measures of a construct at each occasion of course can add validity and reliability to each assessment in comparison to a single measurement at each occasion. Relatedly, and perhaps most importantly, the variance estimates of each

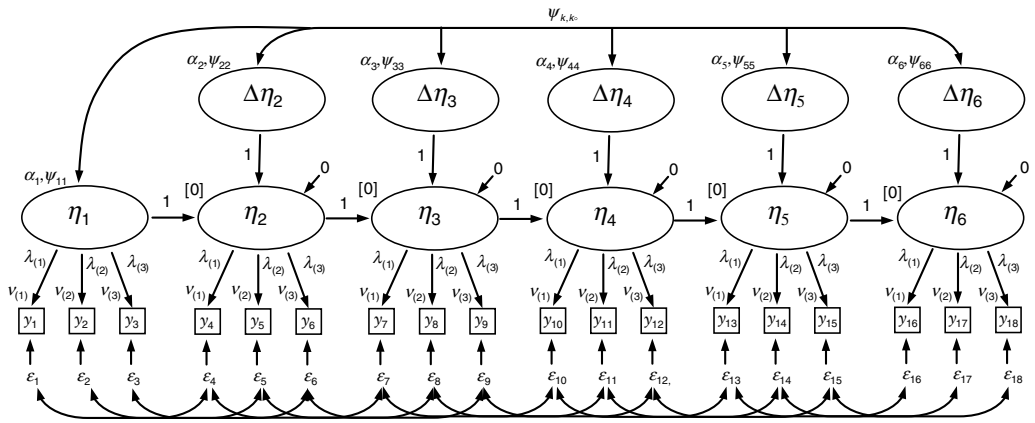

Figure 9.7 Latent Difference Score Model with Multiple Indicators.

construct are inflated by the presence of measurement error, and this inflation is compounded for any composite derived from individual measurements, such as a difference score. As a consequence, multiple indicators should improve the precision of the difference scores, increasing the power of significance tests of factor means and the effect of covariates. Not trivially, multiple indicators also add additional covariance elements and available degrees of freedom, making it possible to identify the latent difference model with fewer occasions.

As with other models, the factor identification method should be carefully considered when interpreting difference scores. Referent indicator identification will lead to difference score factors that are primarily based on the indicator variable.[5] The interpretation of the difference factors can be improved by using the effects coding identification method (Little, Slegers, & Card, 2006), which will lead to difference scores that reflect differences between weighted averages of the indicators at each consecutive time point. Parentheses in Figure 9.7 are used to indicate equality constraints on measurement intercepts and loadings over time. Loadings and measurement intercepts are shown for all items on a factor, with the assumption that the effects coding approach is used for identification. A latent difference model without multiple indicators, but with an observed variable based on a weighted composite using factor scores, would produce identical mean differences, however. Use of commonly employed equally weighted composite scores based on the sum or mean of indicators would produce slightly different results.

Example 9.6: Latent Difference Score Model with Multiple Indicators

One model was tested to demonstrate the latent difference score model with multiple indicators for each occasion. This analysis used the depression measure in the health and aging study. Three subscales, negative affect, positive affect, and somatic symptoms, were used as indicators of a latent factor at each of the six time points.[6] To incorporate information from each of the indicators, effects coding identification was used to define factor variances and means at each occasion via complex equality constraints. Equality constraints were imposed on factor loadings and measurement intercepts for each repeated indicator. Other parameter specifications followed Figure 9.7 with each autoregressive path set equal to 1 and each disturbance set equal to 0. The means and variances for the factor

at the first occasion and all difference score factors were freely estimated. Covariances among the first occasion factor and all difference score factors were estimated. In addition, all measurement residuals were allowed to correlate among all sets of repeated measurements to account for stable specific variance and improve model fit.

The fit of the model was good, $\chi^2(95) = 642.134$, $p < .001$, CFI = .986, SRMR = .042, RMSEA = .033. Note that the fit in this model mainly reflects the measurement portion of the model, as the model is otherwise saturated because all means, variances, and covariances for the difference score factors are estimated. The standardized loadings for each of the depression subscale indicators were generally of acceptable magnitude, ranging from .499 to .901. The mean of the baseline factor was .292, suggesting low overall depression on the 0 to 3 range of the subscale scores. The mean of the five difference score factors were not significant at every wave and did not change in a consistent direction, –.003, ns, .050, $p < .001$, –.016, $p < .01$, .005, ns, –.023, $p < .001$. Variances for each of the difference score factors were significant, indicating that there was individual variation in difference scores. Further analyses could investigate which individuals are more likely to change at particular intervals of the study.

Comments

Latent difference score models are closely related to ANOVA, latent growth curve models, and simplex models. With certain modifications of the specifications, the same hypotheses about change can be tested. On the one hand, this indicates a greater flexibility of latent difference score models, because the general framework can be used for this subset of hypotheses or for other hypotheses. The main contrast to these other models is that latent difference score models estimate level changes across each particular interval. On the other hand, a disadvantage is that latent difference score models may be difficult to estimate for some variants and may need equality or other constraints. Constraints on variances and covariances do not impact average mean estimates in the unconditional form of the model, but they may adversely affect model fit, self-feedback coefficients, and the effects of covariates. It is worth reiterating that the fit of the overall model is largely irrelevant, at least in the basic model form, because its fit does not speak to the validity of the model relative to other models.

Despite some of these limiting factors, the latent difference score model does provide a flexible approach that can be extended in a variety of ways. Perhaps most generally, the latent difference score model can be viewed as a method of estimating nonlinear change by estimating first differences. First differences give a local approximation of first derivatives, and therefore are related to the differential equation models. From this perspective, it should be possible to extend latent difference score models to assess higher-order differences that approximate higher-order derivatives. An additional advantage is that, with the inclusion of covariates, we can account for individual variation in change across specific intervals. The investigation of this type of hypothesis, however, needs to be distinguished from the investigation of causal precedence. Models with lagged time-varying covariates or simultaneous latent difference scores do not address the question of which variable may precede the other causally.

Recommended Readings

There are a variety of introductory chapters on latent difference score models. Perhaps the two most accessible and comprehensive chapters are by McArdle (2009) and Jajodia (2012). Each of these sources has several applied examples. The chapter by Jajodia also

demonstrates and explains a number of graphing approaches that can be used. Hamagami and McArdle (2001) have a fairly detailed exploration of missing data issues. There exist several published applied examples of latent difference score models as well (Ghisletta & Lindenberger, 2003; Ghisletta & Ribaupierre, 2005; Grimm et al., 2012), of which the paper by Ghisletta and Ribaupierre is perhaps the most extensive.

Notes

1 Aside from identifying the model, the inclusion of the intercept in Equation (9.2) also would alter the interpretation of the residual, because the intercept would be subtracted from the right-hand side. The residual in this case can no longer be interpreted as the difference score, and the average residual is not equal to the difference between the means of y_{t-1} and y_t.
2 Recall from Chapter 3 that if we set all the means of the difference factors to 0, we have an omnibus test that is the same as a repeated measures ANOVA.
3 The figure depicts a model specification of an intercept factor. I have depicted the model this way to highlight the resemblance to the growth curve model, but specification of a latent intercept factor in most SEM software programs is not necessary. The same parameter estimates can be obtained without the added intercept factor as long as the covariance between the first measured variable and the slope factor is estimated.
4 Despite the term time-invariant, the covariate may still involve a construct that changes over time. Covariates in a model are often measured at only one time point, either for theoretical reasons or because only a single measurement is available.
5 I use the word "primarily" here because the applicability of this statement partially depends on whether or not equality constraints are imposed on intercepts and loadings. As we saw in Chapter 1, equality constraints on the non-referent indicators complicates the interpretation of the factor. Even with equality constraints, the mean and variance of the factor are still closely connected to the mean and the variance of the indicator variable.
6 The subscale scores were used as indicators for convenience in this case, but these "item parcels" are not usually recommended (e.g., Bandalos & Finney, 2001). There are potential pitfalls when composite scores are used as indicators if the indicators are not strongly related to one another. The present use of parcels may be acceptable given that the depression subfactors are fairly well-established in the literature and are supported in these data. I used parcels in this illustration for convenience, because using individual items as indicators would require third-order factors that would make this example unnecessarily complicated for the present purposes.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary, NC: SAS Institute Inc.

Bandalos, D. L. & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides and R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.

Boker, S., & McArdle, J. J. (2005). Vector field plots. In P. Armitage & P. Colton (Eds.), *Encyclopedia of biostatistics* (Vol. 8, pp. 5700–5704). Chichester, UK: John Wiley & Sons.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis, forecasting and control*. San Francisco, CA: Holden-Day.

Finkel, S. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage.

Ghisletta, P., & De Ribaupierre, A. (2005). A dynamic investigation of cognitive dedifferentiation with control for retest: evidence from the Swiss Interdisciplinary Longitudinal Study on the Oldest Old. *Psychology and Aging*, *20*, 671–681.

Ghisletta, P., & Lindenberger, U. (2003). Static and dynamic longitudinal structural analyses of cognitive changes in old age. *Gerontology*, *50*, 12–16.

Grimm, K. J., An, Y., McArdle, J. J., Zonderman, A. B., & Resnick, S. M. (2012). Recent changes leading to subsequent changes: Extensions of multivariate latent difference score models. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 268–292.

Hamagami, F., & McArdle, J. J. (2001). Advanced studies of individual differences linear dynamic models for longitudinal data analysis. In G. Marcoulides & R. Schumacker (Eds.), *New developments and techniques in structural equation modeling New developments and techniques* (pp. 203–246). Mahwah, NJ: Erlbaum.

Jajodia, A. (2012). Dynamic structural equation models of change. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 291–328). New York: Routledge.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement*, *30*, 157–183.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change. Decade of behavior* (pp. 139–175). Washington, DC: American Psychological Association.

McArdle, J., & Hamagami, F. (2004). Methods for dynamic change hypotheses. In K. Montfort, J. H. L. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models* (pp. 295–335). Dordrecht, Netherlands: Kluwer Academic Publishers.

McArdle, J. J., & Nesselroade, J. R. (1994). Structuring data to study development and change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–267). Mahwah, NJ: Erlbaum.

Plewis, I. (1985). *Analysing change: Measurement and exploration using longitudinal data*. Chichester, UK: Wiley.

Keywords
latent difference score model, latent growth curve models, dual change score model

10 Latent Transition and Growth Mixture Models

SEM is generally concerned with continuous latent variables, but it is also possible to conceptualize latent variables as categorical. The basic multiple-indicator model in which the latent variable is categorical is usually referred to as a latent class model (Lazarsfeld & Henry, 1968; Goodman, 1974). Either exploratory or confirmatory approaches to latent class models are possible, analogous to the distinction between exploratory and confirmatory factor analysis. This chapter will focus only on confirmatory modeling where the number of factors and latent classes is assumed to be known or hypothesized a priori. A more general framework than the latent class model is the concept of finite mixture modeling, in which categorical latent variables can have single or multiple indicators (Lubke & Muthén, 2005; Muthén, 2001). Within such a framework, categorical latent variables may be included as predictors or outcomes in a larger structural equation model alongside observed and latent continuous variables. The flexibility gained by extending structural equation models to incorporate latent categorical models is considerable, allowing researchers to investigate a wealth of new longitudinal hypotheses.

Latent Class Analysis

Basic Concepts

Latent class analysis is based on the general concept of categorical latent variables that differ from continuous latent variables. With continuous latent variables, a set of indicators is used to estimate an unobserved value along a quantitative continuum. With categorical latent variables, a set of indicators is used to estimate membership in unobserved groups. Latent classes are generally assumed to be mutually exclusive and mutually exhaustive, so that all cases in theory are predicted to belong to one and only one latent class.[1] Latent class estimation is a process of estimating membership in an unknown group, which differs conceptually from the concept of the factor model. Factor analysis involving continuous latent variables could be said to cluster items, whereas latent class analysis involving categorical latent variables could be said to cluster cases. The analogy to factor analysis also can be extended to exploratory and confirmatory forms of latent class analysis (Finch & Bronk, 2011). Though the number of classes is unknown, a number ultimately must be specified in the application of the confirmatory model. Similarly, although exploratory forms allow all observed variables to be indicators of all latent class variables, the confirmatory model assumes that the latent structure is known or hypothesized a priori.

As a hypothetical example, consider a survey about political opinions. One way to organize the set of political survey items is in terms of an unobserved continuous variable of political liberalism and conservatism that varies along a single dimension. Another way to organize the set of political survey items is in terms of a categorical unobserved variable

that classifies individuals in terms of their patterns of responses. With latent class analysis, we would therefore group individuals in order to identify types of voters, perhaps obtaining patterns that reflected groups such as moderate liberal, radical progressive, apolitical, and libertarian. The classes are assumed to represent nominal categories of voters and do not represent a continuum.

The meaning of each latent class must be inferred from the data or theory, and it is up to the researcher to name and interpret them. Contrast the example of unknown political class described above from an example of a known political class, such as when the respondent can be assigned as a Democrat, Independent, or Republican based on self-declaration or registration documents, for instance. Latent class membership of any one individual is estimated in a probabilistic fashion. The true number of classes cannot be determined with certainty and must be decided upon, usually by comparing the fit of two or more models (more on this issue below).

Binary Indicators

Although observed indicators of latent class variables may be binary, ordinal, count, or continuous, it is instructive to consider the binary case. The depiction of a simple latent class model parallels the depiction of the continuous latent factor (Figure 10.1). Consider a political survey that includes a series of questions about support for various political issues, such as whether the voter supports prayer in school, more environmental regulation, less funding for government programs, and abortion rights. Underlying population groups can then be identified based on participants' responses to a set of survey questions. In Figure 10.1, the latent class variable is labeled as η^C, and the indicators are y_1 through y_4. Throughout, c will be used to represent an index for a particular class, with C designating the number of latent classes.[2] I use the superscript C for any latent variable η which is a latent class variable with C categories.

For binary indicators, the latent class model has two parameters of principal interest – the predicted probability of class membership of a case and the conditional probability that $y_{ij} = 1$ on the observed indicator that takes into account class membership, usually referred to as the response probability. The index i designates an individual case and j designates a particular indicator. The model shown in Figure 10.1 yields several parameters: thresholds for each of the four indicators and a factor mean estimate. Threshold estimates, τ_{jc}, are specific to each class. For a model with two classes, there is a single mean estimate, α_c, associated with membership in one of the classes. The other class serves as the referent class with an implied mean estimate of 0. With more than two classes, there is a factor mean estimate for all but one referent class. No loadings are estimated. The strength of the relation between the latent class variable and the indicators can be gleaned from the threshold estimates (more detail on this point is provided below). To reflect that the indicators are not directly linked to the latent class variable via traditional factor loadings as they are with continuous latent variables, I use dotted line arrows in the diagrams.

Class Membership Probabilities. A simple latent class model with two classes (e.g., "liberal" and "conservative") and a set of binary indicators yields a mean estimate for one latent class and thresholds for each of the indicators. The other latent class serves as the referent class for identification purposes, and its value is assumed 0. The referent category for latent classes is akin to the 0 to 1 comparison on the outcome in a logistic regression. The mean and intercept parameters from the latent class model are estimated from the observed values, similar to the estimation of the parameter values for a continuous latent variable with a set of indicators in standard SEM. Because the latent variable is binary in

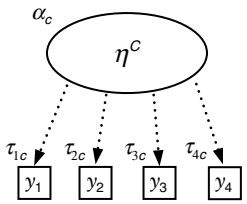

Figure 10.1 Latent Class Model with Binary Indicators.

our example, the unconditional probability of latent class membership is an exponential function of the latent class mean, α_c, just as the expected probability on the outcome can be expressed as an exponential function of the intercept in an intercept-only (i.e., the null model) logistic regression.

$$\hat{\pi}_c = P\left(\eta_i^C = c\right) = \frac{e^{\alpha_c}}{1 + e^{\alpha_c}}$$

In this equation, the probability that a randomly selected case belongs to class c of the latent class variable η_i^C is designated as $\hat{\pi}_c$. The symbol e is the mathematical constant equaling approximately 2.718. This transformation only applies to the simplest circumstance in which there are two classes.[3] To identify the model for more than two classes, there is one referent class with $C - 1$ class means estimated.

For more than two classes, the transformation can be extended to be a function of the sum of all C class means, where it is assumed that the reference class has a mean of 0 (with $e^0 = 1$) in the calculation.

$$\hat{\pi}_c = \frac{e^{\alpha_c}}{\sum_{c=1}^{C} e^{\alpha_c}}$$

The threshold parameters may be estimated uniquely in each class if the model is identified, but between-class constraints may also be imposed. As with logistic regression, the odds of membership in one class versus the referent class can be computed by exponentiating the latent class mean, $e^{\alpha_c} = \hat{\pi}_c / \hat{\pi}_0$, where $\hat{\pi}_0$ is the probability of membership in the referent class. Because the latent classes are mutually exclusive and exhaustive, the sum of the latent class membership probabilities should equal 1.

$$\sum_{c=1}^{C} \hat{\pi}_c = 1$$

This equation assumes that the indicators are fully accounted for by the latent class variable. In other words, there is no residual association among indicators, an assumption termed *conditional* or *local independence*.[4] For the model to be valid, we also assume that there are no predicted latent class membership probabilities equal to 1 or 0, so $0 < \hat{\pi}_c < 1$.

Response Probabilities. The second parameter of importance is the indicator threshold. The indicator threshold contains information about the conditional probability that a particular indicator will be equal to 1 given class membership. These conditional probabilities are referred to as *response probabilities*. In terms of responses to a survey, the response probability is the chance that an individual responds "yes" given that he or she

is the member of a particular class. With a continuous latent variable, the relationship of the indicator to the latent class variable can be conceived of as a regression of y_j on η. For a latent class model, however, the threshold is the parameter used to represent the strength of association of the latent variable to the indicator. The threshold is the intercept from a logistic regression model with y_j regressed on η^C. Just as the intercept in a logistic regression represents the log odds that $y = 1$ when $x = 0$, the threshold from the latent class analysis is the log odds that $y_j = 1$ for a particular value of the latent class variable, $\eta^C = c$. The threshold value for an indicator, τ_{jc}, can be converted to the conditional probability, $\hat{p}_{j|c}$, that the response is equal to 1 given class membership. The logistic cdf transformation can be used to obtain the conditional probability from the threshold (Finch & Bronk, 2011; Muthén, 2001),

$$P\left(y_{ij} = 1 \mid \eta^C = c\right) = \hat{p}_{ij|c} = \frac{e^{\tau_{jc}}}{1 + e^{\tau_{jc}}} \tag{10.1}$$

The complementary function to the exponent function, the natural logarithm, can be used to obtain the threshold again using the conditional probability.[5]

$$\tau_{jc} = \ln \frac{\hat{p}_{j|c}}{1 - \hat{p}_{i|c}}$$

As with binary variables in general, we have no unique information about variances, because the mean (proportion) and variance are not independent parameters. This is why Figure 10.1 for binary variable has no measurement residual associated with it.

The response probability is similar to the concept of the discrimination parameter from Item Response Theory (for more on the relationship between latent class models and IRT, see Agresti, 2013; Clogg, 1995). In the context of the political example, the conditional probability would represent the probability of a "yes" response to the statement "there should be more environmental regulation" if the class membership is liberal ($c = 1$) rather than conservative ($c = 0$). If the probability that the indicator is equal to 1 is very high (i.e., $\hat{p}_{j|c}$ is near 1 and τ_{jc} is a larger positive number) for one class compared to another class, the item discriminates well between the two latent classes. Low relative within-class probabilities also indicate good discrimination. If the probability that the indicator is equal to 1 is very low (i.e., $\hat{p}_{j|c}$ is near 0 and τ_j is a large negative number) compared to the other class, then this also indicates that the item discriminates well between the classes. If $\hat{p}_{j|c} = .5$, corresponding to $\tau_j = 0$, class membership is not related to the indicator, because a case has an equal chance of being in one class or another.[6]

The response probabilities from the latent class model provide the same kind of information that loadings provide in the continuous factor model. When indicators for a latent class variable have high or low response probabilities, they are good indicators of class membership, and, when all are good indicators, the pattern of responses of individuals in the class will tend to be similar, exhibiting high *homogeneity* (Collins & Lanza, 2010). Homogeneity suggests that individuals within a class are similar, and when the response pattern maximally differentiates between classes, there is high *latent class separation*.

The Latent Class Model. The statistical model for latent class analysis states that an observed pattern of responses is equal to the sum of the product of the class membership and the response probabilities across classes. A case has a particular pattern of values on the indicators if it falls into a particular cell of the contingency table formed by crossing all indicators. For a very simple example with two binary indicators, there will be a 2×2 contingency table formed by the two possible responses to two indicators, y_1 and y_2. For

the political survey example, one cell would contain all individuals who responded "no" to school prayer and "yes" to environmental regulation. With *J* binary indicators, there will be 2^J cells. The predicted probability that a case will be in one cell of the two-way design, $\hat{p}_{y1\times y2}$, is equal to the product of the probability of class membership, $\hat{\pi}_c$, and the conditional probabilities of the responses to the two indicators, $\hat{p}_{1|c}$ and $\hat{p}_{2|c}$,

$$\hat{p}_{y1\times y2} = \hat{\pi}_c \hat{p}_{1|c} \hat{p}_{2|c} \tag{10.2}$$

This simple example is just for illustrating the concept. In practice, the model would not be identified with only two indicators.

Based on the statistical model and assuming local independence, the overall (marginal) probability that an indicator y_j will be equal to 1 can be computed by summing the product of the probability membership in a given class *c* and the response probabilities across all of the *C* classes (e.g., Nylund, Asparouhov, & Muthén, 2007).

$$P\left(y_j = 1\right) = \sum_{c=1}^{C} \hat{\pi}_c \hat{p}_{j|c}$$

In other words, the latent class model states that the class membership and response probabilities account for the overall probability of the response when all classes are taken into account.

Equation (10.2) leads us to the general formula for the latent class model with any number of indicators and any number of classes:

$$P\left(\mathbf{Y} = \mathbf{y}\right) = \sum_{c=1}^{C} \hat{\pi}_c \prod_{j=1}^{J} \hat{p}_{j|c} \tag{10.3}$$

Y is a vector of possible values, **y** is a vector representing a particular observed pattern, and $P(\mathbf{Y} = \mathbf{y})$ is the probability of a particular response pattern across all of the binary *y* indicators. The symbol Π is the product operator, indicating the product of the conditional probabilities across all *J* indicators. The formula therefore states that a particular response pattern on the indicators is a function of the predicted latent class membership probability and the product of all the conditional probabilities for the values of the indicators.

Notice that the product on the right-hand side of Equation (10.2) or Equation (10.3) resembles the product of the conditional and marginal probabilities that appears in Bayes' theorem. Working backward from this general equation, the posterior probability of class membership for an individual case, $\hat{\pi}_{ic}$, can be generated using the class membership probability and item conditional probability (Clark & Muthén, 2009).

$$\hat{\pi}_{ic} = \frac{\hat{\pi}_c \prod_{j=1}^{J} \hat{p}_{j|c}}{P\left(\mathbf{Y} = \mathbf{y}\right)}$$

The probability of an individual's class membership is "posterior" because the average probability of class membership derived from the data is used to estimate probability of class membership for a particular case. The latent class probabilities for an individual case parallel individual factor scores with continuous latent variables (Muthén, 2001).

Entropy. Probability of individual class membership can then be used to summarize the overall accuracy of classification or class separation. Although there are several possible measures of accuracy, the *entropy* index, *E*, which represents a kind of average of

the natural log of all class membership probabilities, is the most frequently employed (Ramaswamy, DeSarbo, Reibstein, & Robinson, 1993):

$$E = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\left(-\hat{\pi}_{ic}\ln\hat{\pi}_{ic}\right)}{N\ln C}$$

The term "entropy," borrowed from physics, implies a tendency toward disorganization, but the entropy index for latent class variables is typically scored in the opposite direction, with a maximum of 1 indicating the highest accuracy of class membership assignment and 0 representing the accuracy of class membership assignment. The number of cases N and the number of classes C are both taken into account in the denominator, so this value is an estimate of average accuracy of classification for a given case in a given class.

The latent class model is not restricted to binary indicators, and concepts from the generalized linear model can be applied by positing a continuous unobserved distribution y^* that corresponds to the propensity for the binary or ordinal observed response y_{ij} (Clogg, 1995). I will not discuss ordinal or other non-continuous variable types here and, instead, turn attention to continuous indicators next.

Continuous Indicators

Categorical latent variables also may be defined by continuous indicators. Although such models are often referred to in a general sense as latent class models, the term *latent profile analysis* is more specifically applied to latent class models with continuous indicators (Lazarsfeld & Henry, 1968). Figure 10.2 depicts a latent profile model with four indicators. In contrast to Figure 10.1, this model shows measurement residuals, because residual variances for each indicator can be estimated when indicators are continuous. As with the binary case, there are no factor loadings estimated. Parameters include means for each latent class (other than the reference class), intercepts for each indicator, and measurement residuals for each indicator. Because indicators are continuous, intercepts have replaced the thresholds used with binary indicators. Any of the parameters may be constrained to be equal across classes, but, as long as the model is identified, parameters can be estimated separately for each class.

With continuous variables, the latent profile model can be described in terms of the conditional normal distribution of the indicators y_{ij} with mean μ_{jc}, and variance, σ^2_{jc}, given the latent class. For a simple example for an indicator y_j where there are two latent classes, the combined distribution for the indicator, $f(y_{ij} \mid \hat{\pi}_1, \mu_{j1}, \sigma^2_{j1}, \hat{\pi}_2, \mu_{j2}, \sigma^2_{j2})$, is a function of

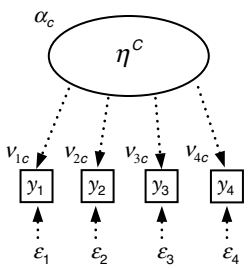

Figure 10.2 Latent Class Model with Continuous Indicators.

the separate distribution density for each latent class (numeric subscripts), weighted by the two latent class proportions, $\hat{\pi}_1$ and $\hat{\pi}_2$.

$$f\left(y_{ij} \mid \hat{\pi}_1, \mu_{j1}, \sigma^2_{j1}, \hat{\pi}_2, \mu_{j2}, \sigma^2_{j2}\right) = \hat{\pi}_1 f\left(\mu_{j1}, \sigma^2_{j1}\right) + \hat{\pi}_2 f\left(\mu_{j2}, \sigma^2_{j2}\right)$$

The symbol *f* is for the normal probability density function with specified mean and variance. The observed means and variances for each indicator for the full sample are fairly simple weighted composites of their class-specific values, weighted by the class proportions (Bauer & Curran, 2004; Muthén, 2002). For two classes (indicated by subscripts), with estimated class proportions equal to $\hat{\pi}_1$ and $\hat{\pi}_2 = 1 - \hat{\pi}_1$, the mean and variance of an indicator in the two classes are

$$\mu_{\text{mixture}} = \hat{\pi}_1 \mu_{1j} + \hat{\pi}_2 \mu_{2j} \tag{10.4}$$

$$\sigma^2_{\text{mixture}} = \hat{\pi}_1 \left(\mu^2_{1j} + \sigma^2_{1j}\right) + \hat{\pi}_2 \left(\mu^2_{2j} + \sigma^2_{2j}\right) - \mu^2_{\text{mixture}} \tag{10.5}$$

As there is a separate distribution for each class with a mean and variance estimate, the statistical model is a *mixture model* because a mixture of distributions is involved. The concept of the finite mixture model is of two normally distributed distributions for unobserved subpopulations that have a combined distribution that is not necessarily normally distributed. Mixture modeling concepts have been used to estimate parameters for separate classes in a variety of different settings (McLachlan & Peel, 2000).[7]

To refer more generally to several latent classes, a vector of class-specific means, $\boldsymbol{\nu}_c$, and a variance–covariance matrix, $\boldsymbol{\Theta}_c$, which is diagonal with local independence assumed, are used. As with categorical variables, the posterior estimates of the predicted membership probability can be computed by

$$\hat{\pi}_{ic} = \frac{\hat{\pi}_c f\left(\mathbf{y}_i \mid \boldsymbol{\mu}_c, \boldsymbol{\Theta}_c\right)}{\sum_{c=1}^{C} \hat{\pi}_c f\left(\mathbf{y}_i \mid \boldsymbol{\mu}_c, \boldsymbol{\Theta}_c\right)}$$

where a vector of observed values for the indicators for each case is $\mathbf{y}_i$, the vector of means for the indicators is $\boldsymbol{\mu}_c$, and the variance–covariance matrix is $\boldsymbol{\Theta}_c$ for each class. The numerator of the equation refers to the probability of membership class *c* and the denominator refers to the sum of the probabilities of membership across all classes. Broadly stated, the equation represents the predicted probability of an individual being a member of a particular class when we take into account the means and variances of the indicators within each class.

Structural Equation Mixture Models

A focus on multiple indicators in a more traditional latent class analysis has served to provide an introduction to some of the underlying principles of categorical latent variables. A more general framework, suggested by Muthén and colleagues (Clark & Muthén, 2009; Lubke & Muthén, 2005; Muthén, 2001, 2002; Muthén & Muthén, 2000), combines categorical and continuous latent variables in the same model. These *structural equation mixture models* open a variety of possible avenues for investigating hypotheses involving unknown groups in two important ways: (1) associations among categorical latent variables, continuous latent variables, continuous observed variables, or categorical observed variables can be examined; and (2) categorical latent variables can be used in a flexible manner that allows classification of observed or latent variables.

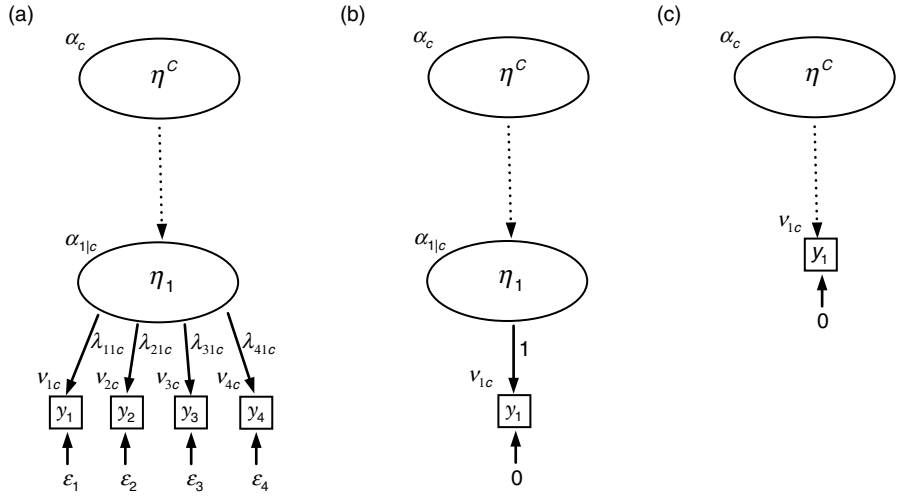

Figure 10.3 Latent Class Variable Prediction of (a) a Continuous Latent Variable; (b) a Single-Indicator Latent Variable; (c) an Observed Variable.

Applying this flexible framework to longitudinal data opens up several interesting modeling possibilities. Figure 10.3 presents three latent class specifications for a single wave of data that can be used as basic building blocks for several important longitudinal modeling approaches. In Figure 10.3a, a latent class variable η^C predicts a continuous latent variable η that has multiple indicators. This model could be viewed as a second-order factor model in which the first-order continuous latent variable is a single indicator of the second-order latent class variable. Indicators for the continuous latent variable, η_1, can be binary, ordinal, counts, or continuous as usual. The latent class factor in this model allows us to obtain separate estimates of the latent continuous variable parameters. Factor means, factor variances, measurement intercepts or thresholds, and measurement residuals for the continuous latent variable can be set equal across latent classes, allowing us to make statistical comparisons between the unobserved groups. Notice that when all of these parameters are equal across latent classes, there is only one latent class and the model simplifies to the common factor model.

Now consider a modification of the first model in which the continuous latent variable is defined only by a single indicator, which must be specified using standard constraints for identification (shown in Figure 10.3b). Because specification of a single-indicator latent variable simply equates an observed variable to the latent variable, the latent class could potentially be equivalently specified directly with only a single observed indicator, assuming it can be identified, and this reduces to the latent class specification to that shown in Figure 10.3c. Once we consider this range of possible specifications of a latent class variable, there is considerable flexibility for testing longitudinal analyses that involve unobserved groups.

Estimation, Model Identification, and Fit

The most common estimator for latent class models is maximum likelihood (ML) using an expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1997; Muthén & Shedden, 1999).[8] In the EM steps of the ML process, conditional expectations and the posterior class membership probabilities are computed in the

expectation step and parameter estimates are updated. The fit is then maximized through iterations in the maximization step. This process alternates between the two steps until an optimization criterion is reached. Estimation can be sensitive to start values and it is wise to retest any model with different start values to be certain that convergence was reached at a global solution not a local solution (Hipp & Bauer, 2006), a testing process that may be automated within the software program. A log-likelihood value obtained upon convergence is used to compute fit indices.

Because the latent classes are mutually exclusive and exhaustive, the class membership proportions sum to 1. Thus, one class proportion is determined when the others are estimated, so there are $C-1$ non-redundant membership probabilities (Clogg, 1995). Latent class mean estimates from the mixture model provide information about predicted membership probabilities, where the estimate associated with one class (typically the last number class) is set to 0 as the reference group. The mean for a given class represents the logit of the proportion of membership in its class, and the probability of membership in the referent class can be computed by subtracting the sum of the probabilities in the other classes from 1. For identification, the number of classes must be less than the number of indicators unless there are additional constraints (e.g., equal variances for the indicators across classes). Thus, with three indicators, only two classes can be specified without special restrictions.

For latent class models that have binary indicators and do not include covariates or latent variables, a likelihood ratio chi-square (sometimes called G^2) and Pearson chi-square are computed to assess fit. Degrees of freedom for these tests can be calculated by $df = \left[2^J - 1\right] - q$, where q is equal to the number of latent class means estimated $(C-1)$ plus the number of class-specific intercepts.[9] For example, with four binary indicators and two classes, there are six degrees of freedom. With three binary indicators and two classes, the model is just identified.

No chi-square model fit is available for latent class models with continuous indicators. Instead, for assessment of fit, one must rely on likelihood-based fit indices, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which are commonly used for evaluation of fit relative to comparison models. The sample size adjusted BIC (aBIC; Sclove, 1987) seems to perform better than other information criteria (e.g., Nylund et al., 2007).

$$\text{aBIC} = -2LL + q \cdot \ln\left(\frac{N+2}{24}\right)$$

The quantity –2LL is the –2 log likelihood value for the H0 model, q is the number of free parameters, and N is the sample size. The number of free parameters, q, is $2JC - (J-1)(C-1)$. Although the model may be theoretically identified, it may not be empirically identified and restrictions may be necessary for convergence. Models with a large number of binary indicators, fewer cases, large number of classes, low membership proportions in one or more classes, or sparse data (i.e., low frequency in the contingency table) may be more susceptible to convergence issues (Lubke & Muthén, 2005, 2007).

Determining the Number of Classes

Part of the process of latent class analysis involves deciding on the correct number of classes, sometimes called *class enumeration*. Although the researcher may have a priori hypotheses about the number of classes, comparisons are usually made among models

with different numbers of classes to provide evidence that the number of classes is correct. The difference in the log likelihoods for two models with a different number of classes is not distributed as chi-square, so an exact test to compare models does not exist. The adjusted BIC is commonly used for this purpose (lower values indicating better fit) and performs fairly well (Tofighi & Enders, 2008; Yang 2006), but a number of simulation studies suggest that more precise methods may be preferable. These methods are designed to compare two models that differ by only one latent class. Among the several proposed alternatives are a bootstrapped likelihood ratio test (Dziak, Lanza, & Tan, 2014; McLachlan & Peel, 2000; Nylund et al., 2007), the Lo–Mendell–Rubin adjusted likelihood ratio test, and the Vuong–Lo–Mendell–Rubin likelihood ratio test (Lo, Mendell, & Rubin, 2001; Vuong, 1989).[10]

An additional complication is that for every latent class model there is an equivalent continuous factor model that can account for the same covariance matrix, where the continuous factor model has one fewer factors than classes in the latent class model (e.g., Bauer & Curran, 2004; Bartholomew, 1987; Lubke & Neale, 2006; Molenaar & von Eye, 1994). The equivalence of these models has sparked considerable discussion among statisticians that is likely to be continued, but, for the applied researcher, there is currently no simple way to distinguish among the two types of models on an empirical basis. There is no choice for the researcher but to decide which type of model is most appropriate based on theoretical considerations in the context of the questions most of interest.

Example 10.1: Latent Class and Latent Profile Models

The depression measure from the social exchanges data set was used to create a latent class model. In order to illustrate a latent class analysis with binary indicators, four indicators of depression symptomatology (bothered by things, felt sad, felt blue, and felt depressed) were artificially dichotomized, with "never" responses set equal to 0 and any higher frequency set equal to 1. Syntax and data sets used in the examples are available at the website for the book. An initial model specified two latent classes. This model seemed to fit the data well, as the two chi-square values were just below the significance cutoff, Pearson $\chi^2(6) = 13.090$, $p = .042$, likelihood ratio $\chi^2(6) = 12.739$, $p = .047$. The aBIC was 2850.153, which cannot be interpreted without an alternative model to compare it to.

The intercept estimates for the four indicators in the first class were 1.134, 2.016, 1.663, and 1.908. Using the logistic transformation from Equation (10.1), the response probabilities for this class were .821, .954, .920, and .958, respectively. Though no clear cutoffs exist for evaluating these values, they suggest that there is a relatively high probability that individuals predicted to belong to Class 1 will report a particular depression symptom. The intercept estimates for the second class were –.392, –1.015, –.780, and –1.217, with response probabilities equal to .179, .046, .080, and .042, respectively. Notice that summing the response probabilities across the two classes for each time totals 1 within rounding error. These response probabilities suggest that those predicted to be in the second class are much less likely to report any particular symptom than those in the first class, and, taken together, the pattern of response probabilities suggest a strong relationship between indicators and the latent class variable.

The probability of class membership is also of interest. The mean for the first class was .301. The logistic transformation of this value indicated that the proportion of respondents predicted to belong to this class was .575 or 57.5%. The second class has a mean of 0 by default, because it is used as the referent class for scaling. Because the two classes are mutually exclusive and exhaustive, the proportion of cases in the second class is equal to .425 or 42.5%. The two classes are unobserved and an interpretation needs to be given to

the two classes. Based on observed proportions endorsing each of the symptoms and the response probabilities, it is likely that the first class is "depressed" or "more depressed" and the second class is "not depressed" or "less depressed." Further investigation of how these classes relate to other variables might provide additional information about the meaning of the two classes.

The four indicators may be better represented by a different number of classes. To investigate this, comparisons to models with different numbers of classes were pursued. Recall that the likelihood ratio tests cannot be used to compare the number of classes. Alternatively, the Lo–Mendell–Rubin adjusted likelihood ratio test (LMR) and the Vuong–Lo–Mendell–Rubin likelihood ratio test (VLMR) can be conducted to compare the fit of a model to another model with one fewer classes. As an initial test, the two-class model specified above was compared to a model with only one class. Results indicated that both tests were significant, LMR = 240.605, $p < .001$ and VLMR = 248.180, $p < .001$, suggesting that the two-class model was a better fit than the one-class model. A three-class model was attempted but did not converge without constraints on the thresholds.

A second model was tested to illustrate a latent profile analysis with continuous indicators. For this model, the health and aging data set was used so that the results could be presented for the latent transition model described below, a model that requires more time points than available from the social exchange data set. The latent profile model estimated positive affect classes using five continuous indicators, frequency ratings for the "happy," "enjoy," "satisfied," "joyful," and "pleased" questions. An initial model specified three classes. No chi-square is available for models with continuous indicators. The aBIC was equal to 5,252.331, but this value can only be used to judge relative fit. As a comparison, the aBIC for a two-class model with the same indicators was 6,575.164, a considerably higher value that indicates the two-class model had a poorer fit. The likelihood ratio tests for comparing the fit of this model to a model with two classes, suggested that the three-class model fit the data significantly better, LMR = 170.942, $p < .001$, VLMR = 176.372, $p < .001$. The mean of the first class was .202 and the mean of the second class was 1.315. Using the cumulative normal cdf transformation, the corresponding class membership proportions obtained were .206, .626, and .168 for the three classes. Table 10.1 presents the within-class intercepts for the five indicators. The observed means presented in the table are for a 5-point scale of agreement, with 5 representing the most positive. The intercepts appear to increase across classes, with the lowest positive affect rating in the first class and the higher positive affect ratings in the third class. This pattern seems to indicate that there are three groups of respondents, unhappy individuals, moderately happy individuals, and very happy individuals, though further investigations would be needed to obtain more information about the meaning of these classes.

Table 10.1 Within-Class Item Means for the Positive Affect Scale

	Observed means, $\bar{y}_j$	*Estimated within-class means*		
		Class 1	*Class 2*	*Class 3*
Happy	3.045	2.167	3.121	3.837
Enjoy	3.092	2.243	3.097	4.115
Satisfied	3.019	2.218	3.028	3.966
Joyful	3.016	2.103	3.024	4.101
Pleased	2.998	2.253	3.015	3.850

Note: responses to the scale are from 1 to 5, with higher values indicating more positive affect.

Latent Transition Analysis

The purpose of latent transition analysis is to study longitudinal changes in an individual's latent class membership over time. Latent class models provide information about the probability of transition from one qualitative state to another. Some possible applications of latent transition analysis might include transitions through Piagetian stages of child development (Jansen & Van der Maas, 2001), movement through stages of change in health behaviors as predicted by the Transtheoretical Model (Guo, Aveyard, Fielding, & Sutton, 2009), development of alcohol or drug abuse behavior patterns (Lanza & Bray, 2010), or structural changes in the labor force (Bassi, Hagenaars, Croon, & Vermunt, 2000). The aim is to discover transitions in membership from one unobserved group to another over time, contrasting with the aim of many other analyses to investigate increases or decreases in the frequency or intensity of a continuous construct over time. The latent transition model is an autoregressive model that uses membership in a latent class at one time point to predict the probability of membership in a latent class at a subsequent time point. The latent transition model with two classes and two time points resembles the autoregressive logistic model or the McNemar chi-square used to examine change in an observed binary variable over two time points, a test of the marginal homogeneity hypothesis (Agresti, 2013). When there are two classes, the latent transition model examines changes in an unobserved binary variable over time. For two time points but more than two classes, the latent transition model resembles the analysis of contingency tables for match pairs that can be used to investigate symmetry, quasi-symmetry, and marginal homogeneity through traditional nonparametric analyses (e.g., Friedman's test), Cochran–Mantel–Haenszel test, or loglinear models (Clogg, Eliason, & Greggo, 1990, discuss the relationship of the latent transition model to these analyses). It may be helpful to refer back to Chapter 3 where repeated measures tests of binary and ordinal variables were discussed.

Basic Concepts

To understand the concept of a latent transition, let us begin with a simple instance where there are only two time points for a two-category latent class variable with multiple indicators (Figure 10.4). Depicted here are binary indicators for the purpose of presenting the simplest example. The latent class variables are defined at the two time points by the same multiple indicators, which may be continuous or categorical as described in the preceding section. As with the cross-sectional latent class model, there are class membership probabilities, $\hat{\pi}_{tc}$, which can be called "latent status" probabilities in this context, and response probabilities, $\hat{p}_{ijt|c}$, for each indicator at each time point t.

In addition to these parameters, the transition probability gives the probability that a case will have class membership in one of the classes at the subsequent time point t given class membership at time $t-1$. The latent class means are conditioned on the latent class from the prior time point and are intercepts in a logistic regression, so their values can provide the probability estimates when transformed. As an example, suppose there are only two classes in our hypothetical political survey, liberal and conservative. Because the intercept is the value of the expected outcome variable when the predictor variable is equal to 0, in this example, the conditional factor mean, $\alpha_{2|1}$, corresponds to the probability of membership in the non-referent latent class (conservative) at Time 2 if the respondent has membership in the referent latent class (liberal) at Time 1. The logistic transformation of the conditional class mean returns the transition probability, $\hat{\pi}_{t|t-1}$, that a respondent will transition from liberal to conservative.

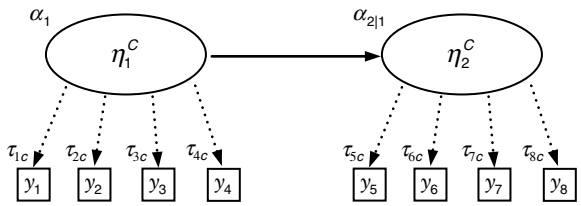

Figure 10.4 Simple Latent Transition Model with Two Time Points.

$$P\left(\eta_2^C = 1 \mid \eta_1^C = 0\right) = \hat{\pi}_{2|1} = \frac{e^{\alpha_{2|1}}}{1 + e^{\alpha_{2|1}}} \tag{10.6}$$

The parameter $\alpha_{2|1}$ is the estimated class specific mean for the latent class variable at Time 2 given a different class membership at Time 1. The formula follows the logit transformation to estimate the probability in the binary case and is parallel to Equation (10.1) for response probabilities with binary variables in the latent class model. For a latent transition model with two classes, a class mean of 0 is equal to the transition probability of .5 and indicates that there is a probability no greater than chance in changing membership from one occasion to the next given prior class membership.

For more than two classes, the transformation of the parameters follows a multinomial logistic model with a referent class for the latent class predictor at time $t-1$ and $C-1$ conditional means at time t. To obtain the transition probabilities with more than two classes, Equation (10.6) can be extended to the multinomial logistic transformation by summing the exponentiations of the additional class-specific means in the denominator.

$$P\left(\eta_t^C = c \mid \eta_{t-1}^C = 0\right) = \hat{\pi}_{tc|t-1} = \frac{e^{\alpha_{tc|t-1}}}{\sum_{c=1}^{C} e^{\alpha_{tc|t-1}}}$$

This formula derives the transition probabilities for a particular non-referent class when there are more than two classes at each time point, where the transition probability represents the transition from the referent class at the prior time point to some non-referent class c at the subsequent time point (symbolized here by the subscript $ct|t-1$). For more time points, the transition probability is derived from the conditional latent class mean at each time point after the first.

I have presented only the simplest circumstances, and it is easy to imagine that latent transitions can become more complicated quickly by adding additional classes and additional time points. It may be easiest, however, to just picture the latent transition model for two classes at two time points where the model is analogous to a logistic autoregression or marginal homogeneity model with unobserved categories. For now, I will spare you the burden of considering the statistical model generalization to continuous indicators, additional classes, and more occasions.

Example 10.2: Latent Transition Model for Two Time Points

To demonstrate just a few of the basic concepts of latent transition analysis, a latent transition model of two time points was tested, following Figure 10.4. There were four binary indicators of depression symptomatology at each time point. The chi-square values were significant, suggesting that the expected and observed frequencies did not match, Pearson

Table 10.2 Transition Probabilities for the Two-Wave Latent Transition Model

	Time 2	
Time 1	*Class 1* $\left(\eta_2^C = 1\right)$	*Class 2* $\left(\eta_2^C = 0\right)$
Class 1 $\left(\eta_1^C = 1\right)$	.677	.323
Class 2 $\left(\eta_1^C = 0\right)$	.236	.764

$\chi^2(244) = 304.541$, $p = .005$ and likelihood ratio $\chi^2(244) = 310.037$, $p = .002$. As there are no standard SEM alternative fit indices, it is difficult to determine whether this model is acceptable overall without comparisons to other models. Given the fairly large sample size ($N = 574$) and degrees of freedom, the overall fit may be adequate in this case. The separate latent class models suggest two classes fit better than one, however, so the lack of fit is unlikely to be due to incorrect number of classes.

The transition probabilities, which are presented in Table 10.2, are of most interest in this analysis. The table is a 2×2 contingency table with probabilities in each of the cells for the repeated measurement of the two-class latent variable. Columns represent the proportions predicted to belong to Class 1 and Class 2 at Time 2 according to predicted membership at Time 1 (rows). For example, the probability in the lower left quadrant indicates the predicted proportion of cases who were predicted to belong to Class 1 at Time 2 if they were initially predicted to belong to Class 2 at Time 1. Because the second (last) class is assigned a class mean of 0, this value represents the probability of transitioning from Class 2 to Class 1, the value designated $P(\eta_2^C = 1 \mid \eta_1^C = 0)$ in Equation (10.6). Thus, the logistic transformation of the conditional mean at Time 2, $\alpha_{2|1} = -1.177$, returns the proportion .236, equal to the transition probability in the lower left quadrant of the table and indicating that approximately 24% of cases from the non-depressed group transitioned to the depressed group over the one-year interval. The autoregressive estimate of 1.917 represents the slope for the logistic analysis of the binary latent class variable at Time 2 regressed on the binary latent class variable at Time 1. The odds ratio computed by raising the exponential constant to this value was 6.800, suggesting that an individual was nearly seven times more likely to be in the depressed group at Time 2 if they were in the depressed group at Time 1.

The Latent Markov Chain: a Quasi-Simplex Model for Latent Classes

Once the basic concepts of the latent transition model are understood as an autoregressive model of categorical variables, the model can be related to the autoregressive models with additional time points. With a set of continuous observed variables that have been repeatedly measured, stability or change can be modeled with the simplex or quasi-simplex model (Chapter 5). Markov chain model is a special name for a simplex model with binary measured variables in which the categories are known. If there are two latent classes defined by a single observed binary indicator at each time point, the model is called a *latent Markov model* or "hidden" Markov model (Baum, Petrie, Soules, & Weiss, 1970; Langeheine, 1994; Wiggins, 1973).[11] Recall that the quasi-simplex model is designed to take measurement error into account in the estimation of the autoregressive effects even though only a single indicator is used to define a latent continuous variable at each occasion. In a parallel fashion, the latent Markov model can be conceptualized as a quasi-simplex

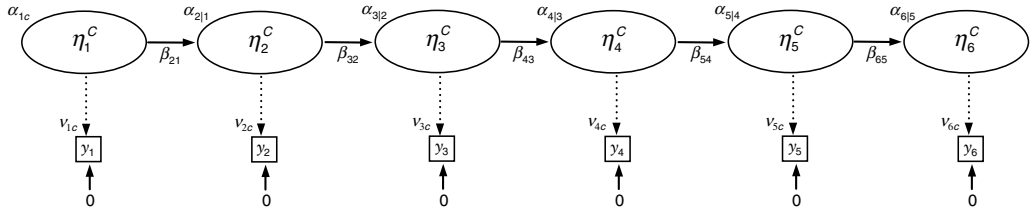

Figure 10.5 Latent (Hidden) Markov Chain Model.

analysis with latent classes defined by only a single indicator at each time point in which the observed variable is considered a fallible indicator of the true underlying class membership (Kaplan, 2009).

Figure 10.5 presents an illustration of the latent Markov chain model for six time points. This is a first-order (lag 1) model in which adjacent autoregressive parameters are estimated. Results from the latent Markov model produce several parameters of interest. The transition probabilities are likely to be a key interest. These values can be derived from the logit transformation of the conditional class mean, $\alpha_{t|t-1}$, taken at each wave (after baseline) using Equation (10.6). The transformation returns the estimated proportion of cases that change from the referent class to the non-referent class over the interval. Response probabilities also may be of interest and can be recovered from a logit transformation of the thresholds, τ_{tc}, using Equation (10.1). The response probabilities provide information about the strength of the relation between the observed binary variable and the binary latent class and, like the loading for a continuous latent variable, can be considered to be inversely related to the degree of measurement error. If response probabilities are all equal to 1, no measurement error is present and a model corresponds to a perfect Markov (or simplex) model with observed binary variables. The third parameter that may be of interest is the autoregressive coefficient, which can be converted to an odds ratio to derive the odds that the class membership at time t given membership at time $t-1$ by raising the exponent to the value of the path coefficient, $e^{\beta_{t,t-1}}$. The result gives the odds that a case is predicted to be in the non-referent class instead of the referent class if the case was in the non-referent class instead of the referent class at the previous time point. This value reflects autoregressive stability rather than transition.

Although Figure 10.5 depicts six time points, a model with only four time points could be identified. At minimum, scaling constraints will be needed by setting the first and last (or two other) within-class intercepts, parallel to the constraints used in the quasi-simplex measurement. If intercepts for the two classes are set to a large negative and large positive number (e.g., –15, 15), response probabilities will be equal to 0 and 1, much like the process of setting the loading and measurement residual for a single indicator of a continuous latent variable. One or all of the following longitudinal constraints also may be needed for identification: the conditional means are equal ($\alpha_{2|1} = \alpha_{3|2} = \ldots = \alpha_{t|t-1}$), the within-class thresholds are equal ($\tau_{1c} = \tau_{2c} = \ldots. = \tau_{tc}$), and the autoregressive parameters are set equal ($\beta_{21} = \beta_{32} = \ldots = \beta_{t,t-1}$). Note that I begin numbering here with $t = 1$, which differs from the growth curve notation.

The imposition of all of these constraints, where parameters are assumed to be equal over time, is often referred to as *stationarity*. Stationarity may not be appropriate if theory suggests transitions may be more likely at certain intervals than others. Because latent class models involve ML estimation, constraints on measurement residuals cannot be imposed directly. It may be desirable in many applications to assume measurement invariance. By

constraining the within-class intercepts to be equal over time, the response probabilities are assumed to be equal. This implies that the same relationship exists between the observed binary indicator and the binary latent class at all occasions. The stationarity model might be a reasonable baseline for testing whether transition probabilities are equal by comparing nested models with and without the latent class means constrained to be equal.

Example 10.3: Latent Markov Chain Model

A latent Markov chain model for self-reported diabetes diagnosis at six time points was estimated in the health and aging data set. The yes/no question about diagnosis may be subject to measurement error and the latent Markov model can be used to estimate the transition probabilities once measurement error has been removed. Measurement invariance was assumed by constraining the within-class intercepts and, thus, the response probabilities, to be equal over time. The Pearson chi-square and the likelihood ratio chi-square were both significant, Pearson $\chi^2(50) = 160.190.151$, $p < .001$, likelihood ratio $\chi^2(50) = 136.543$, $p < .001$, suggesting significant discrepancy between the obtained and expected frequencies. Lack of fit in this case was likely to be due to equality constraints or the assumption of the lag 1 autoregressive effects implied by the model. The sample-size-adjusted BIC (aBIC) was 16,662.570, a value not interpretable on its own but useful for comparison to an alternative model. The latent class mean at baseline was 2.313, which corresponds to a proportion of .910, which is nearly identical to the observed proportion at baseline of individuals reporting no diagnosis of diabetes. The conditional mean estimates for each subsequent occasion were –2.587, –2.310, –2.208, –2.338, and –2.395. Using the logistic transformation, these values correspond to transition probabilities of .070, .090, .099, .088, and .084. In other words, approximately 7% to 10% of the respondents who reported no prior diagnosis of diabetes reported a new diagnosis two years later (i.e., incidence). The autoregressive parameters for the five intervals were all significant 6.137, 5.650, 5.419, 5.277, and 5.123, $p < .001$, with large odds ratios, 462.663, 284.291, 225.653, 195.782, and 167.838, reflecting high stability.

A subsequent model constrained conditional means and autoregressive paths to be equal over time to test for parameter stationarity. This model had a poorer fit, Pearson $\chi^2(58) = 218.590$, $p < .001$, likelihood ratio $\chi^2(58) = 194.420$, $p < .001$, aBIC = 16,677.213. Comparison of the two likelihood ratio values gives a difference of 57.877 with 8 *df*, which would be significant at $p < .001$ according to the naïve test.

A Markov chain model with observed variables (i.e., a perfect simplex model) was estimated as a comparison to the first latent Markov model. The conditional means (intercepts/thresholds) for Wave 2 through Wave 6 were 3.173, 3.004, 2.823, 2.630, and 2.445, which, when converted, are equal to transition probabilities of .040, .047, .056, .067, and .080. These values were smaller than those obtained from the latent Markov model. The autoregressive effects for the observed variable Markov model were 4.420, 4.226, 4.093, 3.957, 3.797, with corresponding odds ratios equal to 83.096, 68.443, 59.919, 52.300, and 44.567, considerably smaller than those obtained with the latent Markov model. Lower stability is generally expected with the assumption of no measurement error implied by the use of the observed variables used in this model instead of latent class variables used in the prior model.

Latent Transition Models with Multiple Indicators

The multiple indicator latent class and Markov models can be integrated to form an autoregressive simplex model with multiple indicator latent class variables at each

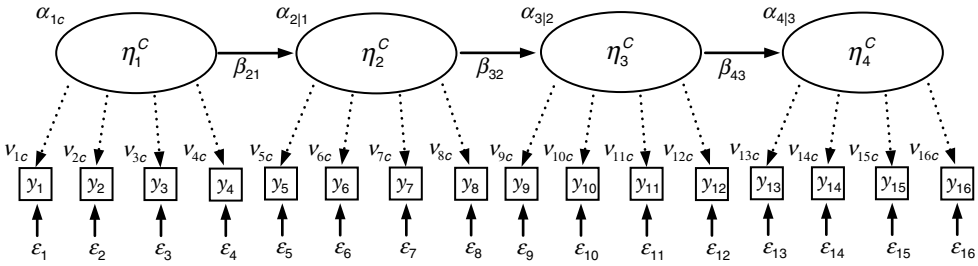

Figure 10.6 Latent Transition Model.

time point. An example of this model is shown in Figure 10.6. All of the necessary fundamental points related to the specification and interpretation of this model have already been discussed. As with the basic latent class model, indicators may be binary, ordinal, or continuous. Multiple indicators potentially add validity to the construct, improve reliability by estimating measurement error, and allow for tests of measurement invariance.

The latent transition modeling framework is sometimes more broadly termed *latent stage sequential modeling* (Collins & Wugalter, 1992; Lanza, 2003). The concept of the stage sequence is that an individual has a specific class membership pattern. For example, in a set of six binary latent classes, one individual's class membership might be 1, 2, 2, 1, 1, 1. This individual moves from Class 1 to Class 2 at the second occasion but then moves back to Class 1 at the fourth occasion. Theory may predict only certain types of transitions and researchers can constrain particular parameters to test them. Take alcoholism as an example. The Alcoholics Anonymous philosophy asserts that once alcoholism develops (or is manifested), a person cannot transition back to a non-alcoholic personality state. Restrictions on transitions can be made by fixing a subset of parameters to specific values and allowing estimation of others. Transitions can only occur in one direction, sometimes described as model with "no backsliding." To test such a model, parameter estimates (e.g., conditional means) can be set equal to 0 for any cases predicted to belong to Class 1 if they were predicted to belong to Class 2 at a prior time point.

The *mover–stayer model* is a specific type of the general state sequential model that classifies individuals in terms of their transition patterns over time (Blumen, Kogen, & McCarthy, 1955; van de Pol & Langeheine, 1989). "Movers" are individuals who change from one state to another, whereas "stayers" are individuals who do not change. Such a distinction can be useful for identifying individuals who do not recover from depression, for instance. Class membership in one of the other groups then can be predicted by covariates in an attempt to explain why individuals do or do not change. The mover–stayer model is estimated by specifying a superordinate latent class variable that categorizes cases into two groups – those that have a pattern that includes one or more transitions and those that have a fully stable pattern.

Example 10.4: Latent Transition Models with Multiple Indicators

A latent transition model was illustrated using six time points and multiple indicators at each wave. A depression latent class variable from the health and aging data set was specified at each wave, with three subscale scores (negative affect, positive affect, somatic symptoms) used as continuous indicators. The model follows the illustration in Figure 10.6 and

represents a hybrid of the multiple-indicator model from Example 10.2 and the latent Markov chain model from Example 10.1. Based on the results from the depression model in Example 10.1, a two-class model was assumed at each wave. Measurement invariance was assumed by constraining intercepts for the same subscales to be equal over time within each class.

Because this model had continuous indicators, there are no chi-square values to gauge fit. The aBIC value was 150,894.870, though this value is not interpretable on its own. With 35 free parameters and a large sample size ($N = 5{,}335$), it is not surprising that this value is large. The latent transition probabilities for those transitioning from the (presumably) non-depressed group to the depressed group were .170, .069, .102, .100, and .134 at Wave 2 through Wave 6, suggesting that between approximately 7% and 17% transitioned at any particular interval. The odds ratios for the five intervals were 26.629, 64.780, 82.682, 95.012, and 53.624, indicating that a respondent was anywhere between 26 times and 95 times more likely to be depressed if he or she was depressed at the prior time point two years earlier.

To illustrate a mover–stayer model, a simpler example was chosen. The binary latent class model from Example 10.1 with four indicators of two classes of depression using the social exchanges data set was expanded to three occasions. With three waves, there are $2^3 = 8$ possible patterns of transitions. This model is a relatively simple example in which there are only two classes at each wave. With three classes per wave, there would be $3^3 = 27$ possible patterns. In the two-class mover–stayer model of depression, there are two possible patterns for stayers. They could exhibit either a 1 1 1 pattern, indicating some depression at all three waves, or a 2 2 2 pattern, indicating no depression at all three waves. Movers could exhibit 25 other possible patterns. For example, the pattern 2 1 1 would indicate no depression at the first wave and some depression at the remaining two waves.

To specify the model, an additional latent class variable with two classes was estimated. The mover class allows transitions between any of the intervals, with free estimation of the within-class means and autoregression parameters. The stayer class sets all transition probabilities equal to 0 and all autoregressive parameters equal to 1. Constrained values are specified in terms of the logistic values, so large negative values (e.g., –15) are used for a probability of 0 and large positive values (e.g., 15) are used for a probability of 1.[12] Each occasion-specific latent class variable is regressed on the mover–stayer latent class variable, giving the odds that movers will be in the depressed class at a given time point.

The Pearson chi-square value was 446.627 with 3,690 degrees of freedom and the aBIC was 39,540.317. No likelihood ratio chi-square was computed because of sparse data. Examination of the estimated transition probabilities for each specific response pattern indicated several with a probability equal to 0, suggesting that the results from this model should be interpreted with caution. The proportion of cases in the mover class was estimated to be .343 and the proportion of cases in the stayer class was .657. These findings tell us that most individuals changed from depressed to non-depressed or from non-depressed to depressed at some point in the three-year span covered by the three waves. It was also the case that a higher proportion of movers (.398) than stayers (.271) were in the depressed class at Time 1. Within the mover class, the transition probabilities estimated by the within-class means were .430 between Time 1 and Time 2 and .412 between Time 2 and Time 3. There were 104 (18.1%) respondents predicted to belong to the depressed class at all three time points (stayers who were depressed).

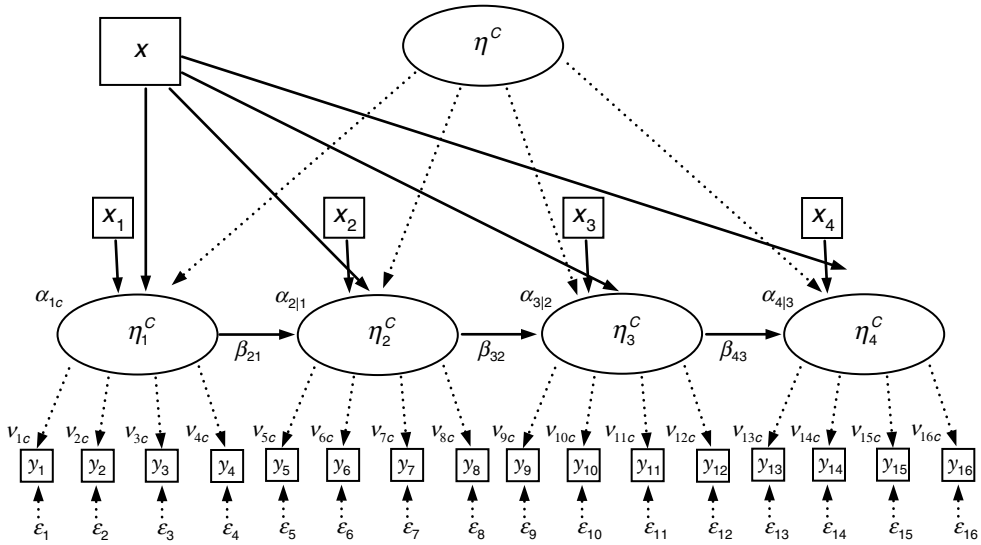

Figure 10.7 Latent Transition Model with Covariates. Note: Correlations among the covariates would be estimated in most applications, but their arrows are omitted in the figure to simply the presentation.

Covariates

Covariates may be included in latent transition models as time-invariant predictors of latent class variables at each occasion, time-varying covariates at each occasion, or time invariant latent class predictors in a mover–stayer model.[13] Covariates may be measured variables, continuous latent variables, or latent class variables. Figure 10.7 illustrates several possible ways covariates might be included in latent transition models. In all instances, the paths depicted for the covariate represent a binary or multinomial logistic effect, depending on the number of classes predicted, and odds ratios can be derived to describe the magnitude of the effect (Muthén, 2001). As with other models, scaling of covariates will be important for reasonable interpretations of the class proportions. Because class proportions are conditional on the covariate, they represent the expected proportion for a class when the covariate is at its zero point. In many applications, covariates may have zero values that are less meaningful (e.g., age at 0 years) or meaningless given the variable's potential range (e.g., value of 0 on a 1 to 7 scale). Consequently, centering the covariate may be desirable in most applications (see the more detailed discussion in Chapter 7).

Inclusion of covariates with latent class variables requires special attention to the accuracy of parameter estimates and their standard errors. The best strategy appears to be to include relevant covariates when investigating class membership and to use multiple pseudo-class draws (Bray, Lanza, & Tan, 2015). When decisions about class membership are made in a separate step excluding covariates, it can lead to problematic estimates of the relationship to auxiliary variables (Bolck, Croon, & Hagenaars, 2004; Clark & Muthén, 2009) and interpretation of classes can change when covariates are later added (Asparouhov & Muthén, 2013; Lanza, Tan, & Bray, 2013). Biases may still occur when variances differ across latent classes, and recent solutions to this problem use modifications of a three-step approach to testing (Asparouhov & Muthén, 2014; Bakk, Tekle, & Vermunt, 2013; Bolck et al., 2004).

Growth Mixture Modeling

Growth mixture models involve the combination of the latent growth curve model and latent class models, grouping individual trajectories by their patterns of change. Grouping of trajectories can be handled with an integrated approach using SEM software (Muthén & Shedden, 1999; Muthén, 2001) or a semiparametric process with other software (Nagin, 1999; Rindskopf, 1990). An important difference is that the semiparametric approach assumes no individual variance of parameters within class (see Shiyko, Ram, & Grimm, 2012 for a more detailed contrast of the approaches). Moreover, the SEM mixture modeling framework has a number of additional potential advantages, such as incorporation into larger structural models like second-order growth curve models.[14] Armed with an understanding of the basic concepts of latent growth curve models and latent class models, the growth mixture model is fairly easy to understand as it is a simple integration of the two approaches. The focus to this point in the chapter has been on latent class variables, so a review of the latent growth curve model will be useful before proceeding to the growth mixture model. I give only a very brief review of latent growth curve models, however, because a more complete introduction is given in Chapter 7.

Growth Curve Model Review

Recall that the linear latent growth curve model (Figure 7.3) uses two latent variables, one to represent the intercept, η_{0i}, and one to represent the slope, η_{1i}, to estimate individual values at baseline and individual change over time.

$$y_{ti} = \alpha_{0i} + \lambda_{t1}\eta_{1i} + \varepsilon_{ti}$$

This equation conceptually represents a regression equation for a line for each case. With the most common coding of loadings for the slope factor, $\lambda_{t1} = 0, 1, 2, 3, \ldots T-1$, the intercept value is the baseline value for an individual, but the coding can be changed to produce differing intercept interpretations. Note that I have switched to the numbering of time points, beginning with $t = 0$, that is used more conventionally with growth curve models. In the linear case, the slope is the linear increase or decrease in the value of the dependent variable for each case over time. The sample as whole can then be described by an average value for the intercept, representing the average baseline value, and an average slope, representing the average linear rate of change per unit of time (e.g., one wave of the study).

$$\eta_{0i} = \alpha_0 + \zeta_{0i}$$

$$\eta_{1i} = \alpha_1 + \zeta_{1i}$$

Combining the two equations forms a single equation for the latent growth model is that stated as

$$y_{ti} = \eta_0 + \eta_1\lambda_{t1} + \varepsilon_{ti}$$

with the means, variances, and covariances for the factors, $\alpha_0 = E(\eta_{0i})$, and $\alpha_1 = E(\eta_{1i})$, $\text{Var}(\eta_{0i}) = \psi_{00}$, $\text{Var}(\eta_{1i}) = \psi_{11}$, and $\text{Cov}(\eta_{0i},\eta_{1i}) = \psi_{01}$.

The latent growth curve model can be conceptualized at the individual case level, allowing for grouping according to unobserved classes. The classes characterize individual trajectories by their intercepts and slopes and may have a variety of interpretations. For example, trajectories from a growth model of changes in individual income that begin in early adulthood may form distinct groupings according to starting incomes and rate of

change over time. Some may have incomes that start out low and increase at a slow pace (e.g., those with only a high school diploma or equivalent or a four-year degree in liberal arts). Others may have incomes that start out low but increase relatively rapidly (e.g., those with a four-year degree in engineering or computer science). Still others may start with a relatively high salary and increase even more rapidly (e.g., those with a graduate education in a professional field).

Specification of the Growth Mixture Model

Figure 10.8 presents an illustration of the growth mixture model. In a similar fashion to how latent class factor models are specified, paths between the latent class variable and the intercept and slope factors are depicted with a dotted line. The dotted line is meant to convey that the model is specified by separate growth models within each of the *C* latent classes. The general model is stated in terms of the sum of the within-class growth curve models, each weighted by the probability of class membership (Bollen & Curran, 2006):

$$y_{tic} = \sum_{c=1}^{C} \hat{\pi}_{ic} \left(\eta_{0ic} + \lambda_{tc} \eta_{1ic} + \varepsilon_{tic} \right)$$

As the equation indicates, an observed score at a particular time point for an individual within a class, y_{tic}, is equal to the sum across classes of the growth curve (in parentheses) weighted by that case's estimated probability of membership in the class, $\hat{\pi}_{ic}$. The average intercept and slope, α_{0c} and α_{1c}, as well as the variances and covariances of the intercept and slope, ψ_{00c}, ψ_{11c}, and ψ_{01c}, can be estimated within each class with a parallel weighting to that in Equation (10.4) and Equation (10.5). In a more general matrix notation, the model's mixture parameters are weighted sums of the within-class parameters.

$$\boldsymbol{\alpha} = \sum_{c=1}^{C} \hat{\pi}_{c} \left[\boldsymbol{\Lambda}_{c} \boldsymbol{\alpha}_{c} \right]$$

$$\sum = \sum_{c=1}^{C} \hat{\pi}_{c} \left[\boldsymbol{\Lambda}_{c} \boldsymbol{\Psi}_{c} \boldsymbol{\Lambda}_{c}' + \boldsymbol{\Theta}_{c} \right]$$

On the left, the vector $\boldsymbol{\alpha}$ contains the mixture estimates of the factor means (e.g., α_0 and α_1) for the full sample and $\boldsymbol{\Sigma}$ contains the mixture variance–covariance for the full sample. The parameters on the right in square brackets are the familiar matrices for the structural model given any particular growth curve constraints desired (see Chapter 1).

Several different latent class specifications may be of interest. As with multigroup SEM, equality constraints on parameters are possible to test equivalence across unobserved classes. Presumably, latent classes are extracted because some aspect of the means and/or variances of intercepts and slopes differ, but equality constraints can be used to explore particular parameter differences, such as whether the average slope differs in two classes. Between-class constraints, such as requiring slope variances to be equal, may be needed for convergence as well. Another option may be to constrain slopes and/or intercepts to be non-varying within class, resulting in a model that equates to Nagin's specification (Nagin, 1999).

The growth curve model specifications are not limited to linear models. Any of the nonlinear models described in Chapter 8 are theoretically possible, including a latent basis model with freely estimated loadings that estimate a flexible form. Each latent class may imply different functional forms as a result of differing parameter estimates within a given specified nonlinear model (Grimm, Ram, & Estabrook, 2010). The possibility of different function forms is an interesting feature that moves beyond single-group growth curve

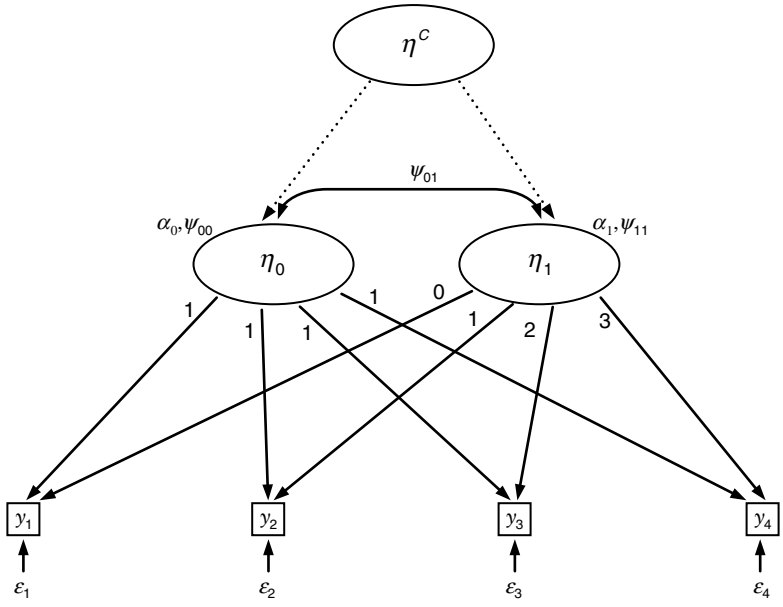

Figure 10.8 Growth Mixture Model.

models in that it allows the researcher to explore whether theoretically different trajectory shapes may apply to subgroups.

Example 10.5: Growth Mixture Model

A growth mixture model based on the linear model of BMI used in Example 7.1 was tested, specifying two latent classes. The model, which follows Figure 10.8, allowed separate estimates of the intercept and slope factor means in the two groups. The intercept and slope variances, their covariance, and measurement residual variances (for same items) were assumed equal across groups for identification. The aBIC was 137,410.379, and the likelihood ratio tests comparing a model with only one latent class were not significant, LMR = 1.185, $p = .802$, VLMR = 1.231, $p = .794$, suggesting that no unobserved classes of trajectories existed for this model. It is unlikely that a model with additional classes would fit better. There is no possibility of overextraction in this situation, because only one class was supported. The intercept and slopes for the two classes, 28.674 and –.112 for Class 1, respectively, and 26.527 and .165 for Class 2, respectively, did not appear to differ substantially, providing further supporting evidence that there were not two separate classes.

Comments

More general precautions related to latent class and latent profile models also apply to growth mixture models. Careful analysis is needed to be certain that the correct number of classes have been identified. The nested statistical comparisons can be used to aid this determination (e.g., Vuong, 1989). Once the number of classes has been determined, interpretation must be given to the latent classes in the final model, and plotting within-class trajectories is essential for the optimal interpretation. A continuum of growth rather than discrete classes may equivalently account for the data, and researchers should use theory

to guide decisions about which is most appropriate. Nonnormal distributions of growth parameters (i.e., violations of within-class normality assumptions) may also be responsible for incorrect conclusions about the number of latent classes, including whether or not latent classes exist (Bauer & Curran, 2003). These authors raise the specter that it may be impossible to distinguish between nonnormal distributions that are mixtures and true qualitatively different latent classes (see also Bauer, 2007). A latent class model will fit better than a single-group model, for example, a problem known as *overextraction*. This problem may be addressable to some extent by better scrutinizing the sample distribution, but distinguishing between true mixture distributions and homogenous nonnormal distributions also may require additional data and relevant theory (Muthén, 2003; Rindskopf, 2003). Recent work using a skew *t*-distribution estimation approach may provide a valuable method that may help avoid overextraction due to nonnormality (Asparouhov & Muthén, 2014; Lee & McLachlan, 2014).

Additional precautions are needed in ascertaining whether the model is correctly specified, particularly with regard to measurement residual correlations (Bauer & Curran, 2004). If correlated errors exist within class (e.g., lag 1 autocorrelations), the true number of latent classes may be fewer than the number of latent classes observed. Determining the number of classes in a growth mixture model requires the same precaution discussed with regard to inclusion of covariates with latent class models. Class membership may be affected by the presence of covariates and this may be especially concerning when relations are nonlinear. These issues highlight the need for thoroughly investigating plausible alternative models.

Ram and Grimm (2009) suggest four steps for testing growth mixture models: defining the problem, model specification, model estimation, and model selection and interpretation. As have others, they suggest that the growth mixture modeling process is one that is bound to be iterative, involving tests of initial hypotheses and adjustments based on the results. This suggested process is a reasonable overarching strategy for testing growth mixture models. The nature of the objective to discover unobserved classes, suggests a flexible approach that involves modifications and retesting. Emphasizing that the unobserved groups are not known but derived from the data and the researcher's judgment process, Ram and Grimm refer to growth mixture modeling as "a constrained exploratory technique" (p. 572).

Extensions

This chapter represents only a basic introduction to latent class models for longitudinal data. To provide an accessible introduction, I have focused primarily on a few latent classes and either binary or continuous indicators. More classes are often appropriate or feasible and other types of indicators are possible, including ordinal and count variables. In addition to the categorical latent class variables discussed, models related to latent class analysis may also involve ordinal classifications instead of nominal categories (Clogg, 1995; Croon, 2002), and this concept can be applied to longitudinal analysis in a state-sequential ordinal model (Kim & Bökenholt, 2000). In addition to these extensions, there are likely to be a variety of other possible model structures involving categorical latent class variables that have yet to be explored by researchers. Growth curve models and latent difference models using latent classification at each occasion and cross-lagged panel models with latent class variables are a few possibilities that come to mind.

Comments

Latent class models, which entail categorical latent variables, have two general functions – to estimate categorical values assuming classification measurement error and to discover discrete categories of cases that underlie a set of observations. The first function represents

a notion that many with a background rooted in continuous latent variables may find difficult to fully grasp at first. Latent class variables with multiple binary indicators or repeated measurements of a single binary indicator are conceptualized as a method of correcting for misclassification due to measurement error, to the extent that a true underlying membership in an unobserved group exists. To this end, Markov chain models and latent transition analyses provide a logical and useful extension of the psychometric underpinnings of continuous latent variables to longitudinal analysis of categorical latent variables.

The second function of latent class models, discovering unknown classes, gives researchers a tool for investigating theories about stages or emergent states, and, in the longitudinal context, transitions through these states over time. The statistical rationale for such models entails a philosophy that there are true qualitative states that can be discovered. In this regard, a more serious concern has been raised that has to do with the heart of the interpretation of mixture models. Not all authors subscribe to the existence or value of qualitative states and argue that observed classes may be equally accounted for by quantitative variation in known or unknown variables along a continuum. Bauer (2007, p.777), in fact, states "By this reasoning, latent trajectory classes are, in most cases, nothing more than artificial categories rendered from the true continuum of change." The reasoning is not based on aesthetic or philosophical differences but on the ability of continuous factor models to equivalently account for the same data accounted for by latent class factors. In response to this perspective, Muthén (2003) argues that such equivalences simply reflect alternative understandings of the same underlying reality, urging researchers to use theory, additional data, and the ability to explain novel phenomena as a guide.

In the end, the differences of opinion about whether categorical latent variables offer anything beyond what continuous latent variables already offer ultimately will press researchers to develop more specific theory and more sophisticated analytic approaches. As we have seen throughout this text, there are many ways to analyze longitudinal data and, in many cases, seemingly different analyses equivalently account for the same data. In all of these instances, a researcher may find one way of interpreting, presenting, or applying the results preferable over another given the theoretical grounding, objectives, or the perceived utility of the approach. This does not in any way imply that the method of analysis never matters. Many times it does, even if just for the sake of greater clarity of presentation. And when an equivalent analysis exists, it may only hold in a limited set of circumstances. But it always is the responsibility of the researcher to be well-informed about the details of an analysis technique and to be certain that the interpretation precisely characterizes the results, making it clear to readers when alternative interpretations exist.

Recommended Readings

An excellent accessible introduction to the general concepts of latent class analysis and latent transition analysis can be found in Collins and Lanza (2010). Clogg (1995) also has a very lucid presentation of the fundamental concepts of latent class analysis. Neither of these sources discusses the SEM approach at length, and there are few general discussions of latent class analysis with SEM that do not focus exclusively or primarily on growth mixture models. Papers by Lubke and Muthén (2005) and Finch and Bronk (2011) are two exceptions that describe the basics of latent class and latent profile analysis within the SEM framework. To learn more about the similarities and contrasts between categorical and continuous latent variables, see Lubke and Muthén (2005), Bauer and Curran (2004), and Bartholomew and Knott (1999). Muthén (2001, 2002) presents overviews of the general mixture modeling integrative framework for combining categorical and continuous latent variables. And papers by Bauer and Curran (Bauer, 2007; Bauer & Curran, 2003, 2004) discuss many of the potential pitfalls of mixture modeling. There

are many more introductions to and examples of growth mixture models. The chapter by Shiyko, Ram, and Grimm (2012) introduces the basic concepts, supplies key references, and compares and contrasts the SEM-based growth mixture model and the semiparametric approach. Grimm, Ram, and Estabrook (2010) provide an illustration of nonlinear growth mixture models with details on implementation in two software programs.

Notes

1 Alternatives to the mutually exclusive assumption allowing cross-classified or partially classified cases have been explored, however (see for example, Erosheva, Fienberg, & Joutard, 2007).
2 Notation varies considerably across authors. I have attempted to be consistent with the rest of the text rather than follow any one of the existing notation systems.
3 The logistic transformation can be stated in one of two ways. An equivalent formula to obtain the probability is $\hat{\pi}_c = 1/\left(1+e^{-\alpha_c}\right)$.
4 Local independence can be relaxed under some circumstances (see Asparouhov & Muthén, 2011).
5 The odds ratio also can be used with $e^{-\tau}$ to describe the odds that $y_j = 1$ given that $\eta^C = c$ (Collins & Lanza, 2010). Chapter 1 explains the interpretation of thresholds for the binary ML estimation in the continuous latent variable case.
6 If there are three unordered possible outcomes for the indicator y_j, then the transformation is a multinomial logit and an average probability of .333 instead of .5 indicates the no discrimination among latent classes, with $\hat{p}_{jlc} = 1/(1+e^{-\tau_1}+e^{-\tau_2})$. The log transformation is then modified accordingly to be $\tau_{jc} = \ln[2\hat{p}_{jlc}/(1-\hat{p}_{jlc})]$.
7 The term "mixture" is not uniquely applied to latent class or structural equation models. The mixture distribution concept has been around for a long time (Pearson, 1895) and is much more broadly applied in statistics.
8 Most software programs use a maximum likelihood estimation via the expectation maximization (EM) algorithm, but a Bayesian process is also possible. There are a variety of software programs that estimate latent class models (Lanza et al., 2007; Vermunt & Magidson, 2005), but currently only Mplus (Muthén & Muthén, 1998–2007) and Mx (Boker et al., 2012) integrate latent class variables within the larger SEM framework (the so-called mixture modeling approach). Mplus uses the maximum likelihood-EM approach with a robust standard error adjustment as the default.
9 The likelihood ratio chi-square may be problematic for sparse data, where the Pearson chi-square is sometimes substituted. The likelihood ratio chi-square is not valid for comparing different number of classes (Lanza, Bray, & Collins, 2013).
10 These tests, which compare the specified model to a model with one fewer classes, have been automated in current version of Mplus (Asparouhov & Muthén, 2013).
11 Models with two or more classes with single ordinal or continuous indicators also are conceivable but stray from the principal conceptualization of the Markov model as a method of estimating measurement error in observed binary variables.
12 Mplus software also allows the user to specify values in terms of probabilities instead of logistic parameter values.
13 Variables other than the latent class variables in the model, including outcomes, are sometimes referred to as *auxiliary variables*.
14 Some researchers have also employed a two-step process of estimating an overall growth curve model and then using other methods of clustering, such as discriminant analysis, cluster analysis, or graphical methods, to classify the trajectories. It has been argued that such a two-step process is flawed because the original model parameters are estimated assuming a single group (Jedidi, Jagpal, & DeSarbo, 1997).

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.
Asparouhov, T., & Muthén, B. (2013). *Using Mplus TECH11 and TECH14 to test the number of latent classes*. Retrieved from http://statmodel2.com/examples/webnotes/webnote14.pdf.

Asparouhov, T., & Muthén, B. (2014). *Structural equation models and mixture models with continuous non-normal skewed distributions: Mplus Web Note, 19*. Retrieved from http://www.statmodel.com/download/Skew.pdf.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological methodology*, *43*, 272–311.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Griffin.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bassi, F., Hagenaars, J. A., Croon, M. A., & Vermunt, J. K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors an application to unemployment data. *Sociological Methods & Research*, *29*, 230–268.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, *42*, 757–786.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*, 338–363.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological Methods*, *9*, 3–29.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164–171.

Blumen, I., Kogan, M., & McCarthy, P. J. (1955). *The industrial mobility of labor as a probability process*. Ithaca, NY: Cornell University Press.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... Brandmaier, A. (2012) *OpenMx version 1.3*. Retrieved from. http://openmx.psyc.virginia.edu.

Bolck, A., Croon, M. A., & Hagenaars, J. A (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis 12*, 3–27.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. New York: Wiley.

Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 1–11.

Clark, S. L., & Muthén, B. (2009). *Relating latent class analysis results to variables not included in the analysis*. Retrieved from http://statmodel2.com/download/relatinglca.pdf.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum.

Clogg, C. C., Eliason, S. R., & Grego, J. M. (1990). Models for the analysis of change in discrete variables. In A. Von Eye (Ed.), *Statistical methods in longitudinal research*, (Vol. 2, pp. 409–441). San Diego, CA: Academic Press.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). New York: Wiley.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*, 131–157.

Croon, (2002). Ordering the classes. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 137–162). Cambridge, UK: Cambridge University Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect Size, Statistical Power, and Sample Size Requirements for the Bootstrap Likelihood Ratio Test in Latent Class Analysis. *Structural Equation modeling: A Multidisciplinary Journal*, *21*, 534–552.

Erosheva, E. A., Fienberg, S. E., & Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, *1*, 346–385.

Finch, W. H., & Bronk, K. C. (2011). Conducting confirmatory latent class analysis using M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 132–151.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. *American Journal of Sociology*, *79*, 1179–1259.

Grimm, K. J., Ram, N., & Estabrook, R. (2010). Nonlinear structured growth mixture models in Mplus and OpenMx. *Multivariate Behavioral Research*, *45*, 887–909.

Guo, B., Aveyard, P., Fielding, A., & Sutton, S. (2009). Using latent class and latent transition analysis to examine the transtheoretical model staging algorithm and sequential stage transition in adolescent smoking. *Substance Use & Misuse*, *44*, 2028–2042.

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, *11*, 36–53.

Jansen, B. R., & Van der Maas, H. L. (2001). Evidence for the phase transition from Rule I to Rule II on the balance scale task. *Developmental Review*, *21*, 450–494.

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*, 39–59.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: Sage.

Kim, J. S., & Böckenholt, U. (2000). Modeling stage-sequential change in ordered categorical responses. *Psychological Methods*, *5*, 380–400.

Langeheine R. (1994). Latent variable Markov models. In A. von Eye, C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 373–395). Thousand Oaks, CA: Sage.

Lanza, S. T. (2003). *Latent stage sequence analysis* (Unpublished doctoral dissertation). Pennsylvania State University.

Lanza, S. T., & Bray, B. C. (2010). Transitions in drug use among high-risk women: An application of latent class and latent transition analysis. *Advances and Applications in Statistical Sciences*, *3*, 203–235.

Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology* (2nd ed., Vol. 2, pp. 691–716). Hoboken, NJ: Wiley.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling: An Interdisciplinary Journal*, *14*, 671–694.

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 1–26.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lee S., & McLachlan G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Journal of Statistics and Computing*, *24*, 181–202.

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*, 767–778.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*, 21–39.

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 26–47.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41*, 499–532.

McLachlan, G., & Krishnan, T. *The EM Algorithm and Extensions*. New York: John Wiley & Sons.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Molenaar, P. C., & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 226–242). Thousand Oaks, CA: Sage.

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class-latent growth modeling. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). Washington, DC: American Psychological Association.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–118.

Muthén, B. (2003). Statistical and substantive checking in growth mixture modeling: Comment on Bauer and Curran. *Psychological Methods*, *8*, 369–377.
Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*, 882–891.
Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*, 463–469.
Muthén, L. K. and Muthén, B. O. (1998–2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological Methods*, *4*, 139–157.
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 535–569.
Pearson, K. (1895). Contributions to the theory of mathematical evolution, II: skew variation. *Philosophical Transactions of the Royal Society of London A*, *186*, 343–414.
Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*, *33*, 565–576.
Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, *12*, 103–124.
Rindskopf, D. (1990). Testing developmental models using latent class analysis. In A. von Eye (Ed.), *Statistical methods in longitudinal research: Time series and categorical lont:itudinal data* (Vol. 2, pp. 443–469). Boston: Academic Press.
Rindskopf, D. (2003). Mixture or homogeneous? Comment on Bauer and Curran. *Psychological Methods*, *8*, 364–368.
Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*, 333–343.
Shiyko, M. P., Ram, N., & Grimm, K. J. (2012). An overview of growth mixture modeling: A simple nonlinear application in OpenMx. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 532–546). New York: Guilford Press.
Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock (Ed.), *Mixture models in latent variable research* (pp. 317–341). Greenwich, CT: Information Age.
van de Pol, F., & Langeheine, R. (1989, October). Mixed Markov Models, mover-stayer models and the EM algorithm. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 485–495). Amsterdam: North-Holland.
van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, *18*, 213–247.
Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 User Manual*. Belmont MA.: Statistical Innovations Inc.
Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.
Wiggins, L. M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Amsterdam: Elsevier.
Yang, C. (2006). Evaluating latent class analyses in qualitative phenotype identification. *Computational Statistics & Data Analysis*, *50*, 1090–1104.
Zhu, H.-T., & Lee, S.-Y. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, *66*, 133–152.

Keywords
latent class analysis, latent transition analysis, growth mixture models, mixture, Markov model, simplex model

11 Time Series Models

Time series analysis focuses on longitudinal data collected over many time points. The primary reason for studying many time points, a *time series*, is to understand regular periodic cycles as well as trends. Only when there are many time points, does it become possible to distinguish regular cycles and trends from unsystematic fluctuations, or noise. This is a key advantage over other longitudinal methods. Although there are many stated benefits of times series analysis, I will highlight two – more accurate forecasting and enhanced investigation of causal hypotheses.

Identifying and removing trends and cycles in time series analysis helps find a stable and reliable underlying process that is not subject to transient conditions, a valuable asset for forecasting and prediction (Chatfield, 2000). Removal of trends and cyclical patterns from the data also can aid the estimation of the causal effects of covariates. Because regular cycles are related to a variable's prior values, sometimes in complex ways. Taking values from prior time points into account beyond the values from the immediately preceding time point extends the analysis of autoregression and cross-lagged effects, thereby expanding our tool box for investigating causal processes (West & Hepworth, 1991). For example, consumer confidence and consumer prices may correspond closely over time, but it may not be that consumer prices directly impact consumer confidence. This confounding may be hard to see if a third variable tends to co-occur with both consumer confidence and consumer prices. Other economic factors, such as economic expansion may affect both.

The immediate objective of this chapter is to provide a basic understanding of how SEM can be used for time series analyses. In service of this objective, the chapter focuses much more on the identification stage than on forecasting or modeling covariates. The identification stage is used to discover trends and cycles in the data in order to stabilize the time series and isolate occasion-specific variance (Box, Jenkins, & Reinsel, 2008). This identification process is of greater concern to social scientists than forecasting typically, and is a necessary foundation that is needed before one can pursue investigation of causal processes when covariates are included. This latter component is also missing to a great extent from this chapter, because the inclusion of covariates to investigate causal processes is discussed at length throughout the text.

The reader will recognize many topics from elsewhere in the book. As will become apparent shortly, time series analysis includes a host of familiar elements of other longitudinal structural equation models. In fact, except for a few new topics, time series will seem like a natural extension of just about every model discussed to this point, including ANOVA models, autoregression and simplex models, growth curve models, and latent difference score models.

Time Series Models for Multiple Cases

Conventional time series analysis is almost always a study of the history of a single case. Aggregate or individual data is tracked over many time points. SEM, however, is generally

an analysis of multiple cases. Time series concepts can be applied quite readily to SEM with multiple cases to model trends and cycles over an extended series. I first discuss analysis of multiple cases and then return later to the analysis of single cases. There are some important differences in the two circumstances, but there are perhaps more similarities than differences.

Overview

Autocorrelation. Much of time series analysis is concerned with identifying cycles and then modeling them in order to eliminate them and stabilize the series, removing one aspect of the fluctuation in values over time. In single case analyses, an important purpose of this process is a statistical one, that of eliminating serial dependency. A time series analysis regresses y_t on t and this almost inevitably leads to serial dependency that violates the independence of residual assumption of regression.[1] The process of stabilizing the data, however, also leads to discovery of the patterns and process of change over time. And, by removing cycles, the remainder is occasion-specific variance, which is variance that is easier to explain in isolation of preceding values and natural random fluctuations.

Cycles may be due to systematic or unsystematic factors. Several definitions, which have not been discussed thus far in this volume, are needed in order to understand time series analysis. *Period* and *frequency* describe cycles and are inversely related to one another. A period is the number of time units it takes for a value to repeat, and frequency is the number of complete cycles per unit of time. For example, if sales dramatically increase each December, the period is 12, assuming a monthly interval. Frequency is thus 1/12 or approximately .083, because only .083 of the cycle is completed in a month. Peak-to-peak a*mplitude* is the magnitude of difference between the lowest and highest values (i.e., the peaks) within a given cycle. A linearly increasing or decreasing trend has no cyclical pattern, so there is zero frequency and an infinite period.

When cyclical patterns exist, whether they are due to systematic factors or not, the data will have an autocorrelation or autoregressive structure.[2] Cyclical patterns can be accounted for by explicitly modeling the autocorrelation through autoregression. Non-cyclic fluctuations, such as spikes, also occur, and these phenomena can be reduced by the use of moving averages. Trends, involving increasing or decreasing patterns, are taken into account by the process of *integrating* which entails transforming the data by taking differences or employing other transformations. The autoregressive, integrated, and moving average processes comprise the ARIMA approach to time series (Box & Jenkins, 1970), the focus of this chapter.

Stationarity. The concept of stationarity is central to ARIMA models. The term *stationarity* implies that some parameter is unchanging over time and may be applied to the mean, the variance, or autocorrelations. Strictly speaking, assumptions and constraints involve autocovariances rather than autocorrelation, but I refer to autocorrelations often because they are easier to conceptualize than their unstandardized counterparts. Stationarity is achieved to the extent that autocorrelation, trends, and non-regular fluctuations have been eliminated from the series. The term *white noise* is applied to the observations when it can be established that there is no change in the mean (mean is 0 on average) and no autocorrelation over time, and the process of producing these stationarity conditions is called *prewhitening*. When a process is nonstationary, the mean, variance, or autocovariances may increase, decrease, or just fluctuate over time. More precisely, the stationarity assumption for the mean implies that the expected value of the observation, $E(y_t)$, is not a function of (i.e., does not change with) time. Similarly, variances $\mathrm{Var}(y_t)$, of the observations are not a function

of time either. The expectation and the variance here are taken over the series rather than across cases. Covariances among observations over time also must not be a function of time. For any lag of order ℓ, the covariance, $\mathrm{Cov}(y_t, y_{t+\ell})$ should be the same across time points for any give lag ℓ. In other words, if $\ell = 1$, the covariance of consecutive time points for the segment of the series, say t_0 through t_{19}, should be the same as the covariance of consecutive time points for the segment of the series, say t_{20} through t_{39}. Similarly, covariances must be equal for all segments of the full series and equal for any given lag. If the distance between time points varies in the study, expectation of equal autocovariances would not be reasonable without taking the interval variation into account algebraically. The covariance stationarity assumption is a mathematical necessity for single-case single time series, because there must be a constant that describes the covariance of the lagged variables in the series. Note that that covariance assumption generalizes to $\ell = 0$, so that the variances also are assumed to be unchanging with respect to time.

For conventional time series analysis of a single case, establishing stationarity signals that independence of errors and constant variance assumptions from regression analysis have been met. Strict stationarity also requires that the joint distribution does not change over time, which is reflective of the normality assumption, but a weaker version of stationarity without the joint probability stipulation is usually assumed for time series analysis.

For time series structural equation models with multiple cases, the stationarity assumptions are no longer a mathematical necessity to identify the model, because we can estimate means, variances, and autocorrelations across cases. We do not bias parameter estimates just because stationarity is not met. Stationarity is important for other conceptual reasons, however. When trends exist in the data, correspondence with another variable may be obscured by a trend that is common to the two variables rather than a causal relationship between the two variables. Accounting for trends and cyclical patterns captures some aspects of the historical factors that may be confounding causes of any corresponding cycles or trends. The ability of time series models to estimate or remove these patterns from the relationships among potential causes and effects is one distinct advantage over other longitudinal analysis methods discussed in this book, such as cross-lagged panel models or growth curve models of shorter series.

Investigation of mean, variance, and covariance stationarity of time series models can make use of the familiar model fitting approach using likelihood ratio (chi-square difference) tests of longitudinal invariance (see Chapter 2 for a general overview). These tests are preliminary investigations of stationarity that will need to be repeated after estimating the moving average, integrative, and autoregressive components to check that the proper process has been modeled. The development of the final time series model is a process of settling on the appropriate moving average, integrating, and autoregressive orders.

Example 11.1: Stationarity

The first set of illustrations of time series analyses make use of the diabetes data set, which has not been used yet. The data set, described in greater detail at the beginning of the book. Example Data Sets section at the beginning of the book, contains 24 days of daily diary data from a sample of 129 patients with Type II diabetes. The focus of the examples is a daily assessment of positive affect measured by the Positive and Negative Affect Scale (Watson & Clark, 1994). Syntax and data sets used in the examples are available at the website for the book.

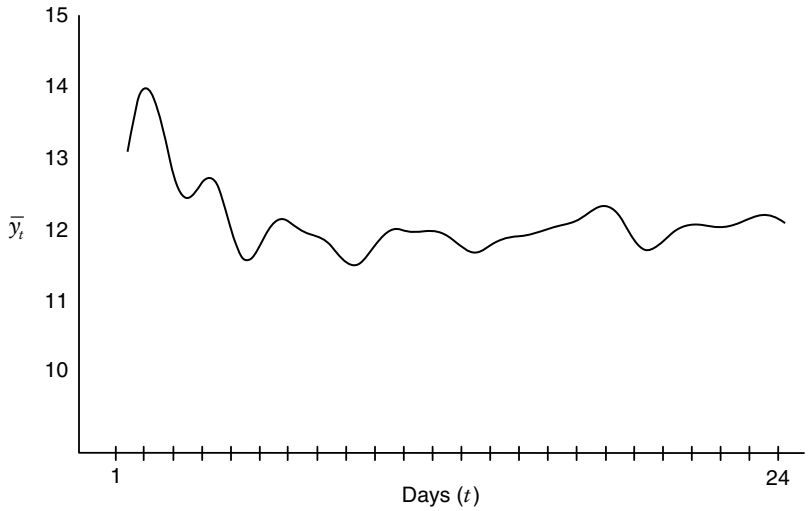

Figure 11.1 Plot of Positive Affect Means From the Diabetes Diary Data.

Before conducting statistical tests of stationarity, the means of positive affect ratings over 24 days were plotted against time (Figure 11.1). The plot suggests variable changes from day to day as well as some overall declining trend across the 24 days. A test of mean stationarity was then used to assess whether this variability over time was significant. The mean stationarity test compares a model in which means were estimated freely for each wave to a model in which means were constrained to be equal at each wave. A significant difference in fit indicates changes over time that may be due to daily fluctuations, linear or nonlinear trends, or cycles. The model included freely estimated correlations among time points, although this specification will not impact the conclusions of the model comparisons as long as the same autocorrelation structure is imposed on the models that are compared. The freely estimated model was just identified. Mean estimates varied between approximately 11.5 and 14, with a value of 13.157 on the first day and a value of 12.102 on the last day. Results for the model with means constrained across time points indicated a significant chi-square, $\chi^2(23) = 97.948$, $p < .001$, although the relative fit indices indicated a very good overall fit, CFI = .983, SRMR = .050. This was a moderate difference in chi-square compared with the just identified model, $w = .086$, $\Delta Mc = .063$. There is no definitive criterion for deciding that the degree of departure from stationarity is important, of course, but the plot of means and significance test suggest that there may be daily fluctuations and periodic cycles that could be reduced by using moving averages, differencing, or another transformation.

Stationarity of autocovariances was tested for three lag lengths ($\ell = 1$, $\ell = 2$, and $\ell = 3$) as a convenient illustration. In practice, it may be advisable to test for stationarity at several more or all lag lengths. For each lag length, two models were compared, one in which the autocovariances for the particular lag tested were allowed to be freely estimated across time points and one in which autocovariances for that lag were constrained to be equal across time points. For each model comparison, the autocovariances for the lower-order autocovariances were constrained to be equal. Table 11.1 shows that for all lags tested, stationarity is not supported. The researcher must decide, however, whether the magnitude of the stationarity violation is sufficient to cause concern. In this instance, the magnitude increase in the chi-square value when imposing the constraint was small to moderate,

Table 11.1 Stationarity Tests for Autocovariances of Lags 1, 2, and 3 ($N = 129$) for the Positive Affect Diabetes Data

	Free				*Equal*				*Difference*				
	χ^2	*df*	*CFI*	*SRMR*	χ^2	*df*	*CFI*	*SRMR*	$\Delta\chi^2$	Δdf	*p*	*w*	ΔMc
Lag 1	3,137.050	253	.356	.646	3,294.317	275	.369	.650	157.267	22	<.001	.112	.009
Lag 2	2,556.315	253	.485	.628	2,607.953	274	.478	.628	51.638	21	<.001	.065	.003
Lag 3	2,252.656	253	.553	.612	2,350.085	273	.536	.612	97.429	20	<.001	.092	.011

Cohen's $w = .112$, .134, and .092, and $\Delta Mc = .009$, .003, and .011. This is only a sample of tests that could be conducted. A full set of covariance constraints could be tested in a single omnibus test, or a more theoretically driven approach could be used, imposing separate equality constraints for particular lag length segments (e.g., by week).

An investigation of variance stationarity used the likelihood ratio test to compare models with and without constraints on observed variances. In this test, means were allowed to be freely estimated across time points and all autocovariances were freely estimated. The less restricted comparison model was just identified, which allows for a convenient assessment of the relative fit of the restricted model. When all variances were constrained to be equal, the chi-square was significant, $\chi^2(23) = 92.224$, $p < .001$, although the relative fit indices still indicated a reasonably good fit, CFI = .985 and SRMR = .092. The magnitude of the chi-square difference was small to moderate, $w = .084$ and $\Delta Mc = .059$. Given the three sets of stationarity tests, stationarity does not seem to be tenable for means, variances, or covariances, with none of these parameters showing independence of time. It would therefore be reasonable to pursue specification of one or more aspects of the ARIMA model to reduce or eliminate variation in these parameters over time.

Autocorrelation

Examination of the pattern of autocorrelations for different orders can be informative about the process and is a good preliminary step for understanding the data. Although free correlation estimates of all possible lags may be simple to obtain in some programs, obtaining a single correlation estimate for each particular lag length of interest by constraining covariances to be equal over time is useful for discerning systematic patterns. To obtain single correlation estimates for each lag, however, equality constraints are also needed on variances over time. The autocorrelation is the autocovariance standardized by the variance. If we assume the variance is stationary, with all $\mathrm{Var}(y_t)$ equal, then the autocovariance for any lag ℓ is equal to the covariance divided by the variance.

$$\rho_\ell = \mathrm{Cor}(y_t, y_{t-\ell}) = \frac{\mathrm{Cov}(y_t, y_{t-\ell})}{\mathrm{Var}(y_t)}$$

The covariance is between pairs of observations for some specific interval apart, $\mathrm{Cov}(y_t, y_{t-\ell})$ obtained as a single estimate using equality constraints on the variances and the covariances. When stationarity of variances has not been established, equality constraints would be problematic, because differences in variances could be mistaken for differences in autocorrelations.

A plot of these autocorrelations is called an *autocorrelation function* (ACF) or "correlogram," which can be useful for understanding the pattern of the series. The ACF is a simple bar graph with lag ℓ on the *x*-axis and the value of the correlation on the *y*-axis. A common pattern is for the magnitude of the correlation to decline with increasing lag lengths (i.e., shorter bars). The autocorrelation of the largest magnitude is usually $\ell = 1$, with higher lags having values closer to zero. If the autocorrelation for the series decays rapidly after the first lag, then the series is mean stationary. Slower decays or non-monotonic changes are more ambiguous, however, and can indicate a linear trend, a nonlinear trend, or a cycle. The ACF is also sometimes used to diagnose the appropriate autoregressive and moving average orders (Hamilton, 1994). In a pure first-order autoregressive process, for instance, the signs of the autocorrelation will alternate at each increasing lag length and will decay toward zero. But, in practice, pure first-order autoregressive processes or

definitive trends in the ACF will be rare, and the plot of the autocorrelations only serves as supplemental exploratory information.

In examining the autocorrelation structure of a time series, it is important to be open to the possibility that there may be an autocorrelation pattern different from what is expected. In standard time series analysis, it is relatively simple to estimate many or all possible autocorrelations to obtain information about the cyclical pattern of autocorrelation. Within an SEM framework, it is not typically as convenient to specify all possible autocorrelation lags when there are many time points, so the researcher may want to investigate at least the first several orders. It is also important to consider any higher-order lags that may be theoretically relevant. For example, researchers investigating daily alcohol use would want to consider whether stronger associations exist six or seven days apart because drinking behavior on weekdays differs from drinking behavior on weekends for many individuals.

An alternative plot, which I will only briefly mention, is the spectral density function (sdf). The sdf is sometimes plotted and inspected instead of or in addition to the ACF. The sdf is simply a mathematical transformation of the autocorrelation, however, so provides no additional information – it is just an alternative method of plotting autocorrelation data (Box et al., 2008). The sdf is computed using a variable frequency, *f*, that ranges between 0 and .5.

$$\text{sdf} = 2\left[1 + 2\sum \rho_{\ell} \cos(2\pi f \ell)\right] \tag{11.1}$$

On the right-hand side of the equation is the summation across products for each of the possible lags, with correlation for a particular lag, ρ_{ℓ}, the cosine function, cos, the mathematical constant, the frequency, *f*, and the lag length, ℓ. The sdf sums across each of set of autocorrelation at each lag, and this value is plotted on the *y*-axis as a function of the frequency, *f*, on the *x*-axis. Positive autocorrelations will exhibit a decreasing density function with the highest values for low frequencies. Conversely, negative autocorrelations will exhibit an increasing density function, with the highest values for high frequencies.

Example 11.2: Autocorrelation

For the positive affect measure in the diabetes data set, I estimated correlations for lag length up to seven days. Variances were set equal across time points and covariances were set equal for each lag length. These constraints produced a single correlation value for each lag. The ACF is presented in Figure 11.2a and shows a steady decreasing autocorrelation across lag lengths, with the highest value for lag 1, equal to .504, and the lowest value for lag 7, equal to .161. Because of the equality constraints, the autocovariance can be used to compute the autocorrelation value with a simple ratio, $\rho_1 = \text{Cov}(y_t, y_{t-1}) / \text{Var}(y_t) = 7.903 / 15.668 = .504$

The sdf was computed for several frequency values between 0 and .5, and a smoothed plot of the values was generated for Figure 11.2b. The sdf should be based on all possible lag lengths, and this figure is only an approximation based on seven lags. The decreasing autocorrelations produced a decreasing sdf graph, consistent with expectations. Both figures indicate that higher scores on positive affect at one point were associated with higher scores on positive affect at a later time point, where the largest autocorrelations were for next day and a decreasing relationship with additional days between assessments. The autocorrelation patterns were also consistent with earlier information, suggesting that these data are not mean stationary.

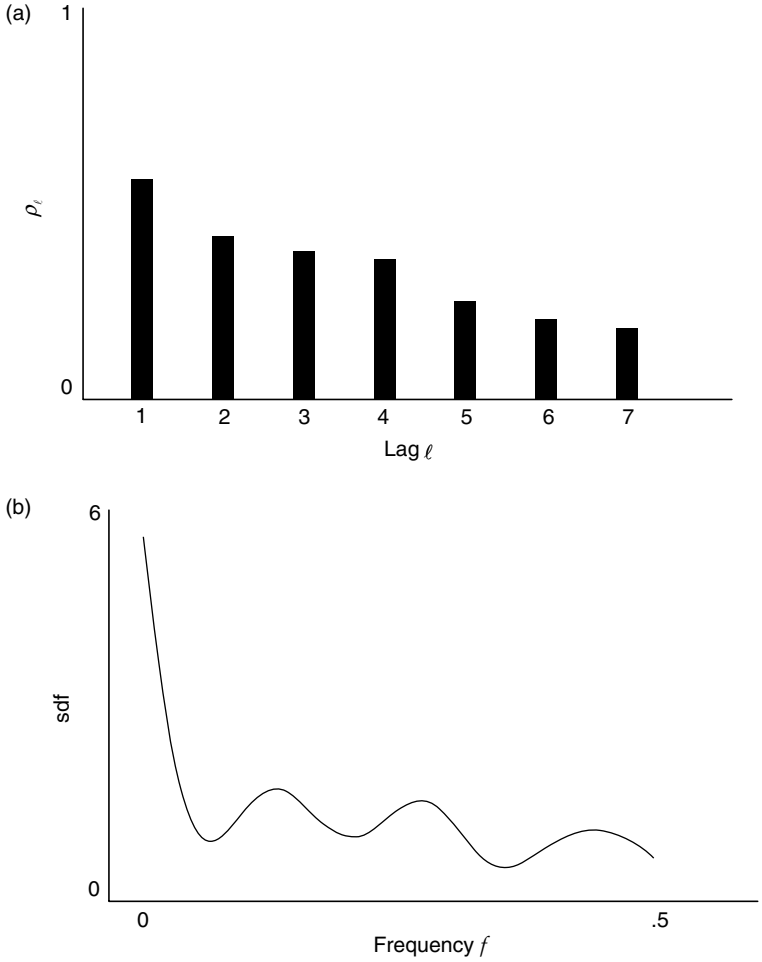

Figure 11.2 (a) Autocorrelation Function (Correlogram) for Seven Lags of the Diabetes Positive Affect Data; (b) Spectral Density Function (sdf) Based on Seven Lags.

Autoregressive Process

An autoregression involves the prediction of an observation at a later time point, y_t, from an observation at an earlier time point, time, $y_{t-\ell}$. Most commonly, the autoregression is for adjacent time points, but longer lag lengths can also be modeled. Autoregressive paths are central to several longitudinal models discussed earlier, including basic models of change (Chapter 4), cross-lagged panel models and simplex models (Chapter 5), and latent difference score models (Chapter 9). Estimating autoregression paths accomplishes several things. First, it provides descriptive information about the stability or inertia of the variable from one time point to the next. Second, it allows the researcher to investigate whether seasonality or cycles exist such that values of the variable are connected over two or three or more waves (i.e., annual increases in consumer purchases during winter holidays). Third, autoregression is a mechanism for taking into account prior values much in the same way that prior values are taken into account with cross-lagged panel models; so that inclusion of lagged covariates can be interpreted as prediction of change in values

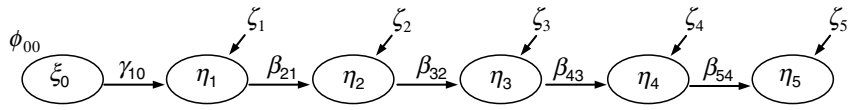

Figure 11.3 Autoregressive (Simplex) Model.

over time. Autoregressive paths model the stability of the time series, and, in the first-order autoregressive process, each time point is a function of the prior time point and disturbance term. The autoregressive process of time series analysis is the same as the process estimated in simplex models (Biesanz, 2012; Sivo & Willson, 2000).

A simple autoregressive model with six time points as typically specified is shown in Figure 11.3. This model corresponds to an autoregressive lag 1 or *AR(1)* model, because regressive paths are estimated only between adjacent time points. The model depicts latent variables at each time point, but each occasion also could be an observed variable. Occasions are enumerated throughout the chapter as $t = 0, 1, 2, 3, \ldots, T-1$. I depart from the notation used in much of the rest of the book and use the full LISREL notation that distinguishes between exogenous, ξ_t, and endogenous, η_t, factors for greater clarity in models presented later. This expanded notation is reviewed in Appendix A. Another new symbol is introduced in the figure. The symbol for the structural path between an exogenous and endogenous variable differs from the all-*y* notation, given here as γ_{10}, which is equivalent to any other structural path except in name. Although latent variable numbers usually begin with 1, I use 0 as the first number of the exogenous factor, which provides greater numeric consistency in the models presented later.

Specifications. For compatibility with other time series concepts, we will specify the first-order autoregressive model in a slightly different, but equivalent, manner. In this model, the disturbances are explicitly defined as latent variables. Specifying the autoregressive model in this manner is useful for incorporation into an ARMA or ARIMA time series model. The structural disturbances are described as "random shocks," "noise," or "innovation" to stress that the goal of time series models is to account for occasion-specific variation. Multiple indicators will be added to the model later, but, for now, assume that ξ_0 and η_1 through η_5 are all single-indicator latent variables, with loadings set equal to 1 and the measurement residual set equal to 0 (not depicted in the figure), producing a model that is the same as a model using only observed scores. ξ_0 is simply the first observed data point, but it is numbered separately for notational convenience and to distinguish it from the exogenous latent variables ξ_1 through ξ_5.[3]

When stationarity is assumed for all parameters, a general equation for the model in Figure 11.4 simplifies to

$$\eta_t = \beta_{t,t-1}\eta_{t-1} + \xi_t \tag{11.2}$$

if ξ_0 from the figure is considered to be one of the η_t latent variables referred to in Equation (11.2) and γ_{10} in the figure is considered to be one of the $\beta_{1,t-1}$ autoregressive paths referred to in Equation (11.2). With these provisos, Equation (11.2) is directly analogous to the equation commonly given for the AR(1) time series model (e.g., Box et al., 2008). The equation is a simple regression where ξ_t functions as a disturbance or random shock.

In Figure 11.4, the random shock variables represent disturbances and are all assumed uncorrelated with one another, $\phi_{t,t-1} = \mathrm{Cov}(\xi_t, \xi_{t-1}) = 0$. All structural disturbances, ζ_t, are set equal to 0 so that the variance of the random shock variables can be estimated. Although the autoregressive process is often given as in Equation (11.2), omitting the

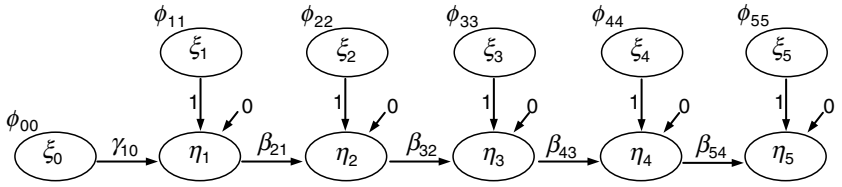

Figure 11.4 First-Order Autoregressive Model with Random Shock Factors, AR(1).

intercept, means will be of interest in most of the applications discussed in this chapter. To model means, we simply estimate a mean structure specifying that the intercepts, α_t, be estimated for each autoregression. If the autoregression intercepts are estimated, then the means for the random shock variables cannot be estimated, $\kappa_t = E[\xi_t] = 0$. To identify the intercepts in a model that uses single-indicator latent variables at each occasion, the measurement intercepts, ν_t, must be set to 0.

Stationarity. For a stationary model that follows assumptions of conventional time series analysis, all autoregressive paths are equal across waves, all random shock variables have equal variances, $\mathrm{Var}(\xi_t) = \phi_{(t)}$, all means of random shock variables are assumed zero, $\kappa_t = E[\xi_t] = 0$, all intercepts are equal, $\alpha_{(t)}$, and all errors of the random shock variables are assumed uncorrelated, $\phi_{t,t-1} = \mathrm{Cov}(\xi_t, \xi_{t-1}) = 0$. When equality constraints are imposed, the parameters associated with ξ_0 should be estimated separately, as neither its mean nor variance have the same interpretation as the other latent variables. du Toit and Browne (2001) show that separate estimation of the variance for the first wave helps account for the autoregressive processes preceding the first observed time point of the study. To the extent that stationarity has not been established, a single autoregressive estimate derived from equality constrained paths will produce biased estimates. Biesanz (2012) suggests the specification of a "phantom" variable to represent stable variance for the variable prior to the observed series. The phantom variable partitions the variance at baseline into observed variance and unobserved variance due to stable aspects of the variable prior to the study. The result is an equivalent model in all respects except that it allows for equality constraints on the variances across all observed time points including the first. This specification may help ensure that the model implied variances across all waves will be stationary.

Tests of stationarity of autoregressive parameters can be conducted using longitudinal invariance tests comparing models with and without equality constraints imposed on the autoregressive paths. And, because stationarity of variance is an assumption that involves conditional variances in the context of an autoregressive process, nested models can also be used to retest stationarity of variance once the autoregressive process has been modeled. The (conditional) means can then be plotted to get a sense of the pattern of the time series after removing the inertial component. This type of model retesting is common in time series analysis to help discover the most appropriate autoregressive lag.

Higher-Order Autoregressive Processes. For the first-order (lag 1) autoregression, the autocorrelation is equal to the autoregression estimate, thus $\psi^*_{t,t-1} = \beta^*_{t,t-1}$, where $\psi^*_{t,t-1}$ is the correlation and $\beta^*_{t,t-1}$ is the standardized regression estimate. When a lag length greater than 1 is used, each path estimate becomes partial with respect to the lower-order lags. The second-order autoregression, or AR(2), model specifies an additional autoregressive path between occasions separated by two time points, $\beta_{t,t-2}$, giving a partial regression coefficient with respect to $\beta_{t,t-1}$. The equation below parallels the conventional time series autoregression model.

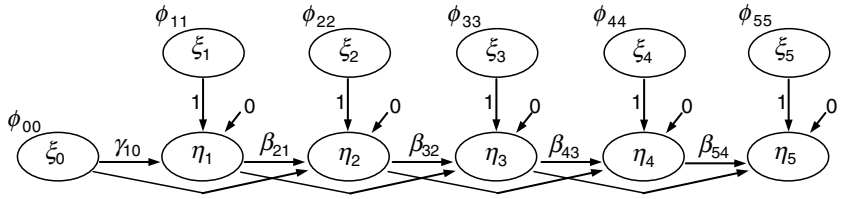

Figure 11.5 Second-Order Autoregression Model, AR(2).

$$\eta_t = \beta_{t,t-1}\eta_{t-1} + \beta_{t,t-2}\eta_{t-2} + \xi_t$$

As in the lag 1 autoregression equation, stationarity is assumed in this second-order autoregression equation, with all autoregression parameters at the same lag and all random shock variables equal. As with the first-order autoregression structural equation model, however, stationarity does not need to be imposed. Figure 11.5 illustrates the AR(2) model without longitudinal equality constraints on parameters.

Because each observation is a function of two prior time points, the lag 2 autocorrelation will not be equal to the lag 2 autoregressive estimate. The lag 2 standardized autoregression is now a function of the square of the lag 1 autocorrelation, assuming stationary variance and autocovariances,

$$\beta^*_{t,t-2} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

with ρ_1 equal to the lag 1 autocorrelation and ρ_2 equal to the lag 2 autocorrelation. Or, more generally, the higher-order standardized autoregression is

$$\beta^*_{t,t-\ell} = \frac{\rho_\ell - \rho_1^\ell}{1 - \rho_1^\ell} \tag{11.3}$$

where ρ_1^ℓ is first-order autocorrelation raised to the power of the lag, ℓ. This function illustrates that the autoregression coefficients will decay toward zero at higher orders.

Example 11.3: Autoregressive Process

The plausibility of a first-order autoregressive process was explored using the diabetes positive affect data. An initial model was tested with autoregressive estimates constrained to be equal over time, estimating the mean structure with the mean for the first time point, κ_0, and intercepts for all subsequent time points, α_1 through α_{23}, freely estimated. Variances of the random shock factors (ξ_1 through ξ_{23}) also were constrained to be equal in this model in order to obtain an independent estimate of the autoregressive parameter, but the constraint is not necessary.

The autoregressive estimate was significant, $\beta = .795$, $p < .001$, and the standardized estimate, $\beta^* = .747$, suggested positive affect was highly stable ($R^2 = .558$). The overall model fit poorly, $\chi^2\,(297) = 1909.884$, $p < .001$, CFI = .640, SRMR = .469. The poor fit of the model may be due to incorrect equality constraints on variances or autoregressive parameters, or it may indicate that a first-order autoregressive process is not sufficient for these data. Freeing the constraints on the autoregressive parameters and retaining all other specifications, resulted in a significant decrease in chi-square value, $\Delta\chi^2(22) = 74.628$,

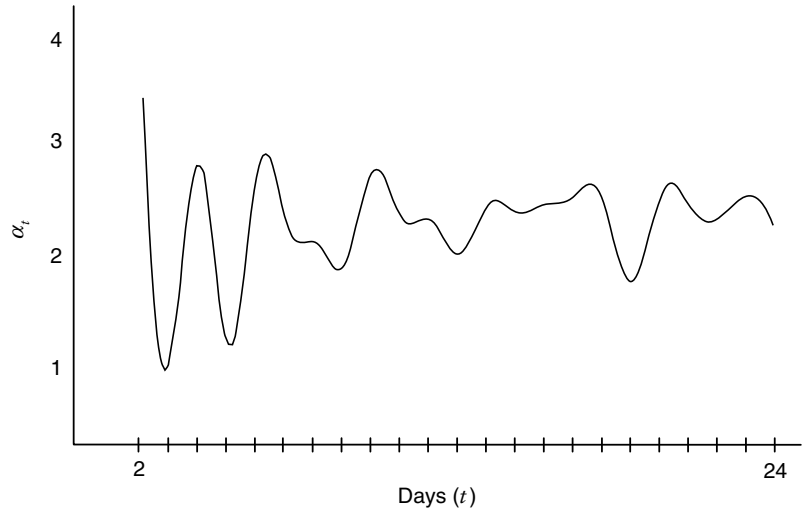

Figure 11.6 Conditional Positive Affect Means (Intercepts) for First-Order Stationary Autoregressive Model.

$p < .001$. Although this was a significant improvement in fit, it was not of large magnitude, $w = .077$, $\Delta Mc = .012$. Although the results suggest stationarity of the autoregressive process is not strictly met, assuming stationarity would not result in large biases.

The intercept values were plotted to examine whether mean stationarity was apparent after accounting for the first-order autoregressive process (Figure 11.6). Values were taken from the initial model with invariant autoregressive paths. Initial inspection of the figure suggests that, after accounting for autoregression process, the means are much more stable over time when compared with Figure 11.1. Intercept values, which represent the expected value when the prior observation is equal to 0, ranged between 3.470 (second day) and 1.311 (fifth day). (Centering around the first time point would produce more interpretable intercepts, but would not change the pattern.) Most of the downward trend seems to have been eliminated after accounting for the first-order autoregressive process. A test of mean invariance for this model compared with the initial model with autoregressive and variance parameters invariant did indicate that the means were not stationary, $\chi^2(320) = 2{,}223.745$, $\Delta\chi^2(23) = 313.861$, $p < .001$, with magnitude of the difference in chi-square that was more substantial, $w = .154$, $\Delta Mc = .055$.

A model test of stationarity of the disturbances (shock factors) allowed these parameters to vary over time. Compared with the first model with all parameters equal, the fit of the model improved significantly, $\chi^2(275) = 1{,}835.115$, $\Delta\chi^2$ (22) =74.769, $p < .001$, though this was a fairly modest improvement in fit, $w = .077$, $\Delta Mc = .012$. None of the tests, therefore, suggested that means or residual variances (shock factors) were stationary after accounting for the first-order autoregressive process. Higher order autoregressive processes should also be tested, but I will leave these tests for the reader to explore.

Moving Average Process

Overview. Due to systematic or unsystematic factors, a time series will fluctuate from one time point to the next, creating a jagged plot of the values. The appearance of this pattern is almost universal when data are observed over an extended period. More spikes of

increase or decrease on any given occasion may also be observed. This type of volatility can detract from observing other underlying processes. A moving average smooths the curve over time and eliminates some of the erratic fluctuations or "noise." Moving averages are commonly used with economic statistics, such as use of the four-week moving average of unemployment applications, to obtain a more stable estimate of the current rate. Calculated by grouping the means for two or more adjacent time points together in succession, moving forward a time point for each occasion, the means stabilize because redundant information is involved in each average, reducing the impact of any aberrant individual values. The term "moving average" is a misnomer, because it is usually computed as a sum rather than an average. The time series implementation does not create averages with weights that total 1. Although often associated with the goal of reducing erratic fluctuations, moving averages also tend to reduce the magnitude of cyclical changes (i.e., reduce amplitude) and flatten trend lines.

The moving average can be specified within an SEM framework by constructing a series of latent variables and constraining paths. A first-order moving average model, or *MA(1)*, can be written as

$$\xi_t = \eta_t - \gamma_{t,t-1}\xi_{t-1} \tag{11.4}$$

where ξ_t is the estimate of random shock variable for a time point, η_t is an observed score at a particular time point *t*. As before, we assume a single-indicator with loading equal to 1 and a measurement residual variance equal to 0, but we will expand η_t to represent latent variable with multiple indicators later.[4]

More generally, the moving average can be a function of any number of lags, expanding the equation to be

$$\xi_t = \eta_t - \gamma_{t,t-1}\xi_{t-1} - \cdots - \gamma_{t,t-\ell}\xi_{t-\ell}$$

Simple algebraic manipulation puts Equation (11.4) into a more informative format for specifying a structural equation model that can represent a moving average process,

$$\eta_t = \xi_t + \gamma_{t,t-1}\xi_{t-1} \tag{11.5}$$

This equation implies the model structure shown in Figure 11.7 in which the measured variable η_t is predicted by the exogenous variable ξ_{t-1} from the prior time point and the exogenous variable ξ_t from the same time point. There is an implicit coefficient equal to 1 for η_t regressed on ξ_t, $\gamma_{tt} = 1$. Following Equation (11.5) gives the observed mean and the moving average estimates. If $E(\xi_t) = \kappa_t$, $E(\xi_{t-1}) = \kappa_{t-1}$, and $E(\eta_t) = \bar{y}_t$, then the observed mean can be computed as $\bar{y}_t = \kappa_t + \gamma_{t,t-1}\kappa_{t-1}$.

The variable η_t thus represents a weighted combination of the current and prior values. The weight is determined by the estimate of γ_{t-1}, where a value greater than 1 gives the mean for $t-1$ greater weight than time *t*, and a weight less than 1 gives the mean at *t* greater weight. In time series analysis, the γ weight is assumed to be equal across all time points as a condition of stationarity. If γ_{t-1} were equal to 1, then the two means would be equally weighted.

Specification. Figure 11.7 represents a pure lag 1 moving average model for six time points that parallels Equation (11.5). It is a "pure" moving average model because there is not an accompanying autoregression process. I will discuss models that combine various processes in the same model later. The full notation system that distinguishes between exogenous and endogenous variables is again used for clarity. The endogenous variables,

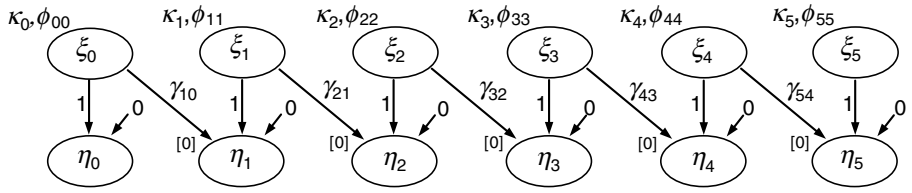

Figure 11.7 Pure First-Order Moving Average Model with Estimated Means, MA(1).

η_t, are observed scores, whereas the exogenous variables, ξ_t, capture the moving average by combining each pair of adjacent observations. The structural paths between each lagged exogenous variable and the endogenous variable are estimated by $\gamma_{t,t-1}$, while the synchronous relationship between each exogenous and endogenous pair has its value set equal to 1. This parameterization produces an estimate of $\gamma_{t,t-1}$ that is the relative weight of the moving average. The symbols κ_t and ϕ_{tt} are the mean and variance for each moving average. The figure shows the model without equality constraints on parameters, but, if stationarity is assumed, a single variance, $\phi_{(t)} = \text{Var}(\xi_1) = \text{Var}(\xi_2) = \ldots = \text{Var}(\xi_{T-1})$, and a single moving average parameter, $\gamma_{21} = \gamma_{32} = \ldots = \gamma_{T-1,T-2}$, are estimated for all time points.

There are two possible specifications for a pure moving average process model. The first approach is to set all means for the ξ_t exogenous variables to 0 and estimate intercepts for the η_t endogenous factors. This specification follows conventional time series analysis in which the shock factors are assumed to have a mean of 0. The second approach is to freely estimate all of the means for the ξ_t exogenous variables and set all of the intercepts for the η_t endogenous factors to 0. The two specification approaches produce the same fit indices and estimate for the moving average parameter, $\gamma_{t,t-1}$. Freely estimating the means for the exogenous factors is useful so that the means can be reexamined after changes in the model have been made. The values for κ_t, the means of ξ_t, provide smoothed values for the series, whereas the intercepts α_t for the regression of η_t, if estimated instead, provide the original observed means.

The covariances among the exogenous shock factors, $\text{Cov}(\xi_t, \xi_{t-\ell}) = \phi_{t,t-\ell} = 0$, are not estimated in the model. Although it may be theoretically identified, adding these parameters will change the estimates for the means and the moving average parameters. When the prior time point is taken into account in predicting η_t by including correlations among the exogenous factors, the interpretations of $\gamma_{t,t-1}$ and ξ_t are altered. Estimating the correlations in this case accounts for the autoregressive process, which would not be desired for the pure moving average model. Because the covariances are assumed to be equal to 0, however, the fit of the model will likely suffer to the extent that there is any autocorrelation process that should be taken into accounted.

As with autoregressive models, plots of means and stationarity tests should be conducted to check whether the appropriate moving average process has been specified.

Example 11.4: Moving Average Process

To investigate the plausibility of a first-order moving average process for the diabetes positive affect data, a series of models were tested. The initial moving average model specified exogenous factors as shown in Figure 11.7, with these factors uncorrelated and the moving average parameter set equal across all waves. Means of the exogenous factors were freely estimated at each wave, but variances for exogenous factors and all moving average parameters were constrained equal across waves.

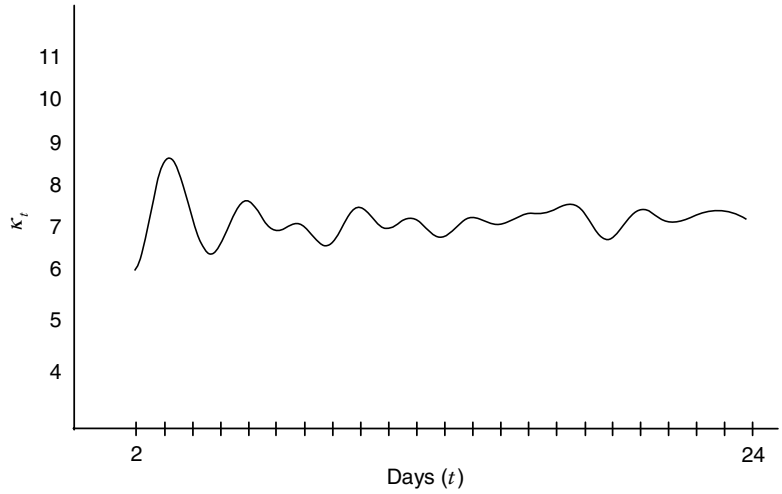

Figure 11.8 Positive Affect Means From First-Order Moving Average Model.

The model of the first-order moving average process had very poor fit, $\chi^2(275) = 3{,}331.513$, $p < .001$, CFI = .317, SRMR = .647. The poor fit of this model is not surprising given the assumed stationarity and independence of the exogenous factors. In practice, the moving average model would frequently incorporate an autoregressive component. The moving average estimate was .551, which gives the proportion of weight given to the $t-1$ observation in computing the "moving average" for each wave. The means of the exogenous factors, which fell between 6.625 and 8.742 did not suggest an overall decline as they did in the unconditional tests. The observed means can be derived from these moving average values following Equation (11.5). For example, taking the observed mean at $t = 3$, $\bar{y}_3 = E(\eta_3) = 12.853$, can be computed from the moving average estimates at $t = 2$ and $t = 3$, $\kappa_2 = E(\xi_2) = 8.037$ and $\kappa_3 = E(\xi_3) = 8.037$, $\bar{y}_3 = \kappa_3 + \gamma_{(t,t-1)}\kappa_2 = 12.854 = (8.742)(.551) + 8.037$. Figure 11.8 shows that modeling the moving average process has smoothed the series and eliminated the declining trend in comparison to Figure 11.1. One can see that even by just incorporation of the moving average component without any autoregressive or integrative components the series can be detrended and daily fluctuations can be smoothed out.

A subsequent model constrained exogenous means to be equal across time points to determine whether there was any remaining nonstationarity in the means. It should be noted that it does not make sense to include the mean for the first time point in this test. Because there is no prior time point to integrate into the mean of the first factor, it is not a moving average. The resulting model had a poorer chi-square than the initial moving average model, $\chi^2(297) = 3{,}362.804$, $p < .001$, CFI = .315, SRMR = .648, but the difference was not significant, $\Delta\chi^2(22) = 31.291$, $p = .093$, and was fairly minute, $w = .050$, $\Delta Mc = .001$. This suggests that incorporation of the moving average process led to approximately stationary means.

Models with Autoregressive and Moving Average Processes: the ARMA Model

Moving averages are often modeled in conjunction with autoregression processes and can be specified simply by estimating paths between the endogenous variables at the desirable

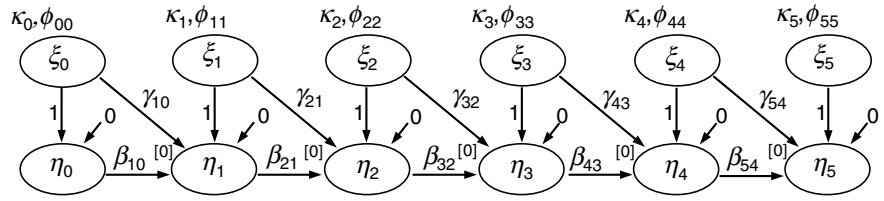

Figure 11.9 Combined First-Order Autoregressive and First-Order Moving Average Model, ARMA(1,1).

lag. Figure 11.9 is a hybrid of the pure first-order autoregression process and the pure first-order moving average process, known as the *ARMA(1,1)* model. Combining the two processes allows the researcher greater flexibility in sorting out random noise from systematic change. The model depicts the parameters as freely estimated across time points, but stationarity may be assumed for variances, autoregression parameters, moving average parameters, or means as desired.

Although moving averages involve means and correlation/regression involves association, moving averages are not independent from autocorrelation and autoregression. Higher-order moving averages imply lower-order autocorrelations, and higher-order autocorrelations imply lower-order moving averages. It can be shown that the first-order autocorrelation is a direct function of the first-order moving average (see Hamilton, 1994, for a derivation). Equation (11.6) shows that the autocorrelation, ρ_1 is equal to a ratio of the moving average parameter, $\gamma_{t,t-1}$, to one plus its squared value.

$$\rho_{t,t-1} = \frac{\gamma_{t,t-1}}{1+\gamma_{t,t-1}^2} \tag{11.6}$$

Larger moving average parameters are associated with higher autocorrelations. If the moving average order is 0, then the autocorrelation implied is 0. A moving average parameter of .5 implies an autocorrelation value of .4. When the moving average parameter is equal to 1, the autocorrelation will be .5.

The dependence of moving averages and autocorrelations suggests that it may be difficult to distinguish the two processes when modeling only one at a time, a phenomenon known as the *Slutsky–Yule effect*. Changes in the average that may be caused by random fluctuations from occasion to occasion may create the appearance of cyclic patterns that can be mistaken for meaningful patterns. Are apparent drought cycles a result of normal variation in weather patterns or the result of some regularly occurring and explainable cause? Either the autoregression or the moving average process need not be first order, and the correct process is usually concluded only after a series of tests.

Example 11.5: ARMA Model

Using the diabetes positive affect data, an ARMA(1,1) model was tested, setting variances for the exogenous variables, the moving average paths, and the autoregressive paths equal across time points. Means of the exogenous variables, κ_t, were freely estimated for each time point in order to examine whether means were stationary. The resulting model had improved fit from the moving average only model but was not acceptable overall, $\chi^2(297) = 1278.593$, $p < .001$, CFI = .781, SRMR = .300. The autoregressive parameter,

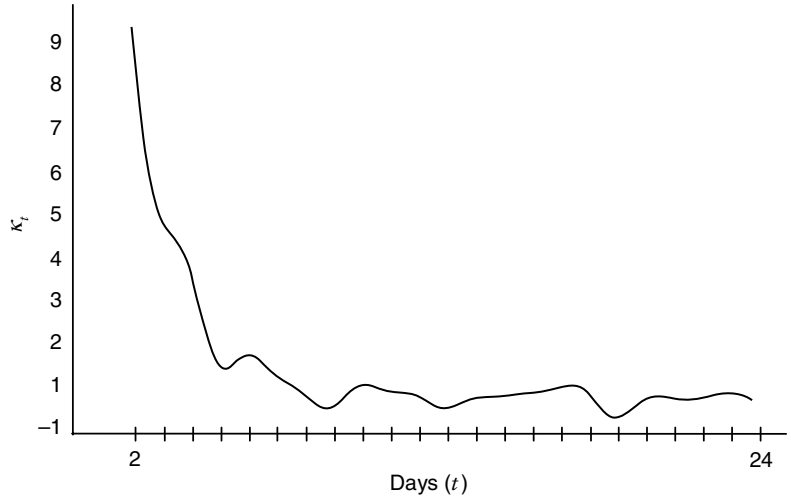

Figure 11.10 Means from the Diabetes ARMA(1,1) Model.

$\beta_{(t,t-1),}$ was equal to 1.001, and the moving average parameter, $\gamma_{(t,t-1)}$, was equal to –.655, a negative value that was not present in the pure moving average model. The κ_t means are plotted in Figure 11.10, omitting the first time point, which is unadjusted. Compared with the moving average only model, an interesting pattern emerges. There is now a steep decline over the first few occasions. This suggests that the ARMA(1,1) model is not appropriate and may be overcorrecting or oversmoothing the data.

Integrated Process

Overview. An alternative process for removing trends is "differencing," which, in its simplest form, just takes the difference between successive time points. If differences are taken between adjacent time points, $t-1$and t, it is a first-order difference. The general process of detrending the data is an integrating process and can involve other mathematical functions such as logarithmic, sine, or cosine transformations (Bowerman & O'Connell, 1993; Mills, 1990). If an integrated process is added to autoregressive and moving average processes, the model is referred to as an ARIMA. If each process is first order, then the model is *ARIMA(1,1,1)*. For higher-order processes, the order of each process is given in parentheses in the order of the three processes: autoregressive, integration, and moving average.

Detrending can be accomplished in a structural equation model either by preprocessing the data in an initial step and then analyzing the detrended data or by incorporating differencing into the model specifications. I will not discuss the preprocessing approach and will leave discussion of transformations other than differences to other authors (e.g., Gottman, 1981; Hamilton, 1994; Mills, 1990). Modeling *integration* processes is sometimes recommended as the first stage analysis before examining autoregressive and moving average processes (Box et al., 2008). Differencing processes also are closely related to moving average processes and autoregressive processes. Although moving average process is primarily associated with smoothing and autoregressive processes are associated with modeling cycles, the three types of processes are not easily separable. The potential redundancy of these processes suggests that by employing one we may be canceling the operation of the other. When

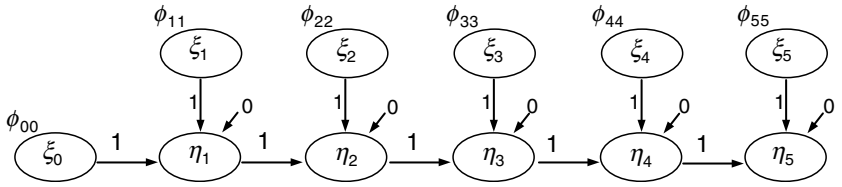

Figure 11.11 First-Order Differencing Model.

applied to integrative processes, the redundancy is referred to as "overdifferencing," which can create trends in the data that did not exist initially (Bell, 1987). The overdifferencing issue is part of a larger concern about overfitting the time series model by using processes that are of a higher order than required by the data, resulting in poorer forecasting than had the appropriate lower-order model been used (Box et al., 2008). Differencing may be needed to detrend the data and achieve stationarity, but, if stationarity is met with autoregressive and moving average processes, then differencing is not required (Hershberger, Molenaar, & Corneal, 1996).

Specification. Computation of any difference score, $\xi_t = \eta_t - \eta_{t-1}$ can be translated into a latent variable model specification, which becomes clear when these terms are rearranged.

$$\eta_t = \eta_{t-1} + \xi_t \tag{11.7}$$

The equation closely resembles Equation (11.2) for the autoregressive process, except that the autoregressive path is set equal to 1 (recall from Chapter 4). Also compare this equation to Equation (11.5) for the moving average process. The differencing process is the same model if the moving average parameter, $\gamma_{t,t-1}$, is equal to 1 (i.e., both $\gamma_{tt} = 1$ and $\gamma_{t,t-1} = 1$).

Figure 11.11 depicts a first-order differencing model for six waves in which each of the immediately preceding scores are subtracted. Each exogenous variable, ξ_1 through ξ_5, represents difference scores with estimated mean and variance, κ_t and ϕ_t. To assess the whether the data have been detrended, plots can be inspected and mean stationarity tests can be conducted. The pure differencing model is the same as the latent difference score model (Hamagami & McArdle, 2001; see also Chapter 9). Means can be assumed stationary or can be freely estimated to investigate stationarity once differencing has been incorporated. The data have been detrended to the extent that there is not an increasing or decreasing pattern across the series, but means may not be entirely stationary even if no linear trend exists.

First-order differencing will eliminate linear trends, but not higher-order trends. More complex differencing models may be needed to eliminate nonlinear trends. One approach is to apply differencing multiple times (i.e., polynomial differencing). This could be accomplished by transformations prior to the analysis, but higher-order differencing can also be achieved through model specification. Although the addition of a second-order difference factor on top of the structure shown in Figure 11.11 would produce a higher-order difference transformation, estimating third-order or higher-order factors would become cumbersome and might lead to estimation difficulties. A more general approach is to alter the specifications that would equivalently produce a higher-order difference.

The second-order difference is a difference between the two adjacent first-order differences,

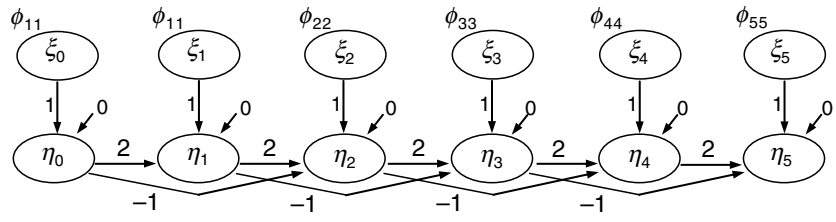

Figure 11.12 Second-Order Differencing Model.

$$\begin{aligned}\xi_t &= (\eta_t - \eta_{t-1}) - (\eta_{t-1} - \eta_{t-2}) \\ &= \eta_t - 2\eta_{t-1} + \eta_{t-2}\end{aligned}$$

Solving for η_t reveals the regression weights desired for estimating a second-order difference with a simple modification of the model specified in Figure 11.11.

$$\eta_t = \xi + 2\eta_{t-1} - \eta_{t-2} \tag{11.8}$$

Thus, the second-order difference can be estimated by setting the lag 2 autoregressive path to −1 and the lag 1 autoregressive path to 2, matching the weights in Equation (11.8). Figure 11.12 illustrates the specification of a second-order difference where the exogenous variable, ξ_t, is a random shock after two successive observations have been differenced.

Example 11.6: Differencing

The diabetes data were again analyzed, this time to illustrate differencing. Specification followed Figure 11.11. Recall that the raw data plotted in Figure 11.1 had an apparent decreasing trend in positive affect over the 24 days. Differencing should remove such a trend. In the initial model, means of the exogenous (difference, shock) factors were freely estimated but their variances were constrained to be equal over time. The model had very poor fit, $\chi^2(299) = 2{,}244.811$, $p < .001$, CFI = .565, SRMR = 1.407. A comparison with a model in which the means were constrained to be equal across waves resulted in a slightly poorer fit, which differed significantly from the initial model, $\Delta\chi^2(22) = 76.644$, $p < .001$, but was a fairly modest difference, $w = .076$, ΔMc = .008. A plot of the means in Figure 11.13 suggested little evident trend remained after differencing.

ARIMA Model

The autoregressive, differencing, and moving average components can be combined into a single model. Figure 11.14 summarizes the specification for the combined model. The lower two rows of ellipses represent the differencing process, whereas the upper two rows represent the combined autoregressive and moving average processes. The middle row of ellipses have variables named η_t^{AR} to distinguish them from the observed variables η_t on the lowest row. Square brackets are used to indicate mean or intercept values set equal to 0. Variances of the ξ_t variables may be set equal for stationarity. Moving average and autoregressive parameters may also be constrained to be equal over time. Mean stationarity can be tested by comparing a model with exogenous means, κ_t, estimated to a model in

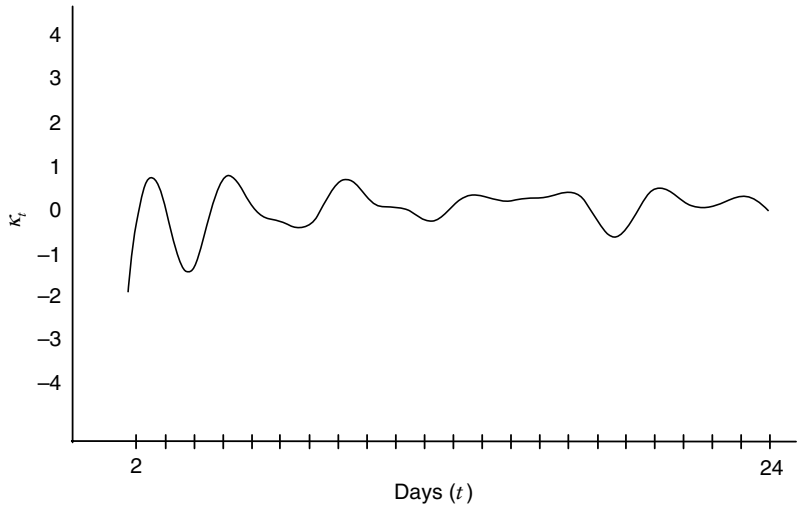

Figure 11.13 Positive Affect Means From the First-Order Differencing Model.

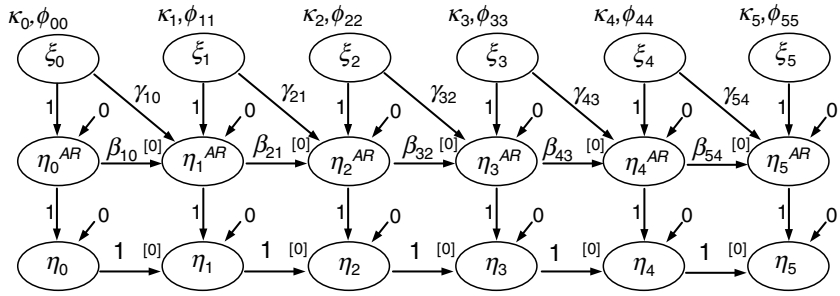

Figure 11.14 Combined First-Order Autoregressive, First-Order Differencing, and First-Order Moving Average Model, ARIMA(1,1,1).

which they are free to vary. Given a considerable number of constraints and the complexity of the model, convergence issues may arise and starting values may be needed. All three processes often may not need to be modeled simultaneously in practice.

Example 11.7: ARIMA(1,1,1)

A model of positive affect combining autoregressive, differencing, and moving average was tested, specified as shown in Figure 11.14 with exogenous variances, moving average parameters, and autoregressive parameters all held equal over time. Exogenous means were freely estimated in order to examine the means for stationarity. The model fit better than previous models but did not meet the criteria for acceptable fit, $\chi^2(297) = 1{,}268.768$, CFI = .783, SRMR = .312. The moving average parameter was –.699, $p < .001$, and the autoregressive parameter was .085, $p < .01$. The moving average parameter was similar to that obtained in the ARMA model (Example 11.5), but the autoregressive parameter differed considerably. A test of a mean stationarity by constraining the exogenous factor means to be equal resulted in a substantially poorer fit, $\chi^2(319) = 1{,}904.384$, CFI = .646,

SRMR = .435, $\Delta\chi^2(24) = 635.616$, $p < .001$, $w = .215$, $\Delta Mc = .177$, suggesting that mean stationarity was not achieved with this model. The plot of the means closely resembled that obtained with the ARMA model (Figure 11.10), so a plot will not be presented.

Comments

Many models may not require each of the ARIMA components, and any combination of autoregressive, integrated, or moving average processes may be necessary for a given data set. It is clear from the diabetes affect data that both differencing and moving averages would not be appropriate for these data, as inspection of means and mean invariance tests from the ARMA and ARIMA models suggest. Both are likely to overfit the data. The pure autoregressive model came the closest to mean stationarity, and, although significant, the magnitude of the chi-square difference tests was fairly small. Theory and substantive knowledge will be needed to help determine whether the violations of stationarity are of sufficient magnitude to cause concern. Although I did not explore higher-order autoregressive processes, inclusion of a second-order autoregressive process might further stabilize the means. Primary focus has been placed on the means, but variances also appeared to violate stationarity and deserve attention during model fitting as well.

Time Series Models for a Single Case

To this point, the models I have illustrated are appropriate for applications that involve multiple cases. These models are natural extensions of the other longitudinal structural equation models presented in this book, with each element given a new interpretation in light of time series concepts. Time series applications are traditionally applied to analysis of a single case observed over many time points, often with economic or historical data because they are likely to have a sufficient number of time points. Time series analysis may be applied to aggregate data, such as economic indicators or national political trends, or individual cases, such as psychological diary data or medical records. Lately, there has been rapid growth in interesting intensive longitudinal studies that feature diary data with many time points and few cases (Bolger & Laurenceau, 2013; Mehl & Conner, 2012). A large number of time points were suggested by Box and Jenkins (1970), who said at minimum that $T \geq 50$, and this has become the standard recommendation for required sample size for conventional time series analysis. With rare exceptions (e.g., Molenaar, 1985), SEM has been applied to data with multiple cases, with a usual minimum recommendation of $N = 100$ under ideal conditions (e.g., Anderson & Gerbing, 1984). In recent years, considerable development work has taken place to extend early work on single-case analyses using SEM to time series models. Many if not all of the concepts discussed for multiple case time series models can be applied to the single case with a few additional considerations.

Each of the algebraic equations above could be shown with a subscript, *i*, for individual cases in addition to a subscript *t*, for time points. Means can be conceptualized as expected values taken across individuals as well as time points. For example, the general ARMA model for a single case is

$$\xi_{ti} = \alpha_{ti} + \beta_{ti,ti-1}\eta_{ti-1} + \beta_{ti,ti-2}\eta_{ti-2} \cdots \beta_{ti,ti-k}\eta_{ti-k} - \gamma_{ti,ti-1}\xi_{ti-1} - \cdots - \gamma_{ti,ti-\ell}\xi_{ti-\ell}\eta_{ti} \tag{11.9}$$

with $E(\alpha_{ti}) = \alpha_t$, given stationarity, $\alpha_t = \alpha_1 = \alpha_2 = \cdots = \alpha_T$. Variances for the white noise component, which represent differences across cases, $\mathrm{Var}(\xi_{ti}) = \phi_{ti}$, are identified

because values over time are set equal, $\phi = \phi_1 = \phi_2 = \cdots = \phi_T$, assuming stationarity and independence, $\mathrm{Cov}(\xi_t, \xi_{t-k}) = 0$.

Maximum likelihood (ML) estimates from time series of a single case can be derived using standard SEM software with raw data from a single series as input (du Toit & Browne, 2001, 2007; Hamaker, Dolan, & Molenaar, 2003; Singer, 2010). Equation (11.10) gives the full information maximum likelihood (FIML) log likelihood function maximized for each case (Arbuckle, 1996; Enders, 2001; see also Chapter 13).

$$\log L(\theta) = \frac{-N(p+q)}{2}\log(2\pi) - \left(\frac{N}{2}\right)\log\left|\Sigma(\theta)\right| - \left(\frac{1}{2}\right)\sum_{i=1}^{N}\mathbf{z}_i'\Sigma^{-1}(\theta)\mathbf{z}_i \tag{11.10}$$

The function implies fitting to an individual covariance matrix of vector $\mathbf{z}_i$ consisting of any set of variables in deviation form under the assumption of multivariate normality. FIML, developed for a sample with multiple cases with missing data, can be used for repeated measures of a single time series. The full ML function for an individual covariance matrix of a single case with repeated measures is then

$$F_{FIML} = \sum_{i=1}^{N}\left(K_i + \log\left|\Sigma_i(\theta)\right| + \mathbf{z}_i'\Sigma^{-1}(\theta)\mathbf{z}_i\right)$$

with constant K_i based on the number of completed observations for case *i*. On the surface, it would seem that a covariance matrix of variables over time constructed from $N = 1$ would lead to violation of independence of error assumptions, but it can be shown that accurate estimates can be obtained and are equivalent to those obtained from traditional ARIMA model estimates. Independence can be met in the SEM model, because the white noise is modeled by an exogenous latent variable, ξ_t, appearing on the left side of Equation (11.9), that is constrained to be uncorrelated with other white noise variables over time (Hamaker et al., 2003).

A series of simulations (Hamaker et al., 2003; Voelkle, Oud, van Oertzen, & Lindenberger, 2012) have shown that the FIML estimates closely match those obtained from other approaches for time series models, such as traditional ARIMA estimation with Mélard's algorithm and Kalman filtering (Zhang, Hamaker, & Nesselroade, 2008) or an SEM approach using block-Toeplitz covariance matrices (Hamaker, Dolan, & Molenaar, 2002; van Buuren, 1997).[5] Convergence issues may arise with small *N* and small *T* in practice, but the simulation work by Hamaker and colleagues (Hamaker et al., 2002, 2003) and by Voelkle and colleagues (2012) show that parameter estimates, standard errors, and model fit closely match traditional ARMA estimation methods (e.g., Mélard's algorithm) when $T \geq 50$. Raw data can be used as input in most programs with no special data preparation or features, as long as the FIML estimator typically used for missing data estimation is available. One exception is that the SEM program also must be able to estimate the model for only a single case in which the variance across individuals is constant. Most programs, however, allow for the default variance check to be turned off.

Example 11.8: Single Cases ARIMA (1,1,1)

The data set for single case analyses used 164 months of the Consumer Confidence Index, an assessment issued by the Consumer Confidence Board based on attitudes and interest in purchasing goods. The example data sets are described in the Example Data Sets section at the beginning of the book. To illustrate, I estimated several models using SEM software and conventional software for ARIMA analysis. Syntax and data sets used in the examples are available at the website for the book. Results for all the models are

Table 11.2 Time Series Estimates for Consumer Confidence Index Data ($T = 164$, $N = 1$)

	SEM		*Conventional ARIMA*	
	Est	*SE*	*Est*	*SE*
AR(1)				
$\beta_{(t,t-1)}$	.994	.005	.998	.002
$\phi_{(tt)}$	40.114		40.682	
AIC	1,082.881		1,079.655	
BIC	1,076.002		1,082.755	
MA(1)				
$\gamma_{(t,t-1)}$	.918	.023	−.955	.029
$\phi_{(tt)}$	2,993.099		2,868.251	
AIC	1,785.808		1,774.525	
BIC	1,779.808		1,777.625	
I(1)				
κ	−.425	.497	−.425	N/A
$\phi_{(tt)}$	40.282		40.463	
AIC	1,077.578		1,065.737	
BIC	1,071.340		1,065.737	
ARIMA(1,1,1)				
κ	−.424	.509	−.425	N/A
$\gamma_{(t,t-1)}$	.714	.351	−.708	.705
$\beta_{(t,t-1)}$	−.672	.375	−.665	.744
$\phi_{(tt)}$	40.156		40.828	
AIC	1,087.068		1,069.196	
BIC	1,077.068		1,075.383	

Note: AR(1) = first-order autoregressive model, MA(1) = first-order moving average model, I(1) = first-order integrated model, AIC = Akaike's Information Criterion, BIC = Bayesian Information Criterion. N/A = not available. SEM estimates were obtained using Mplus Version 7.11 and conventional ARIMA estimates were obtained using SAS 9.2.

summarized in Table 11.2. For the sake of brevity, only key parameters are presented in each analysis. Standard errors are presented for these parameters where available. Results from the SEM approach appear in the left-hand column, and results obtained with conventional software using an ARIMA procedure appear on the right. Model fit of the models is compared using Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwartz, 1978). Lower values on these indices indicate better fit.

General Model Specifications. Specifications for each of the models tested have already been depicted in the figures presented earlier. For the notational convenience of generalizing to multiple indicator variables, observed variables in the figures are represented by single-indicator latent variables, but most software packages allow direct paths among observed variables. Stationarity was imposed for all parameters with means of the exogenous variables, k_1–k_{164}, set equal to 0. Estimation of means is possible, but the estimation of additional parameters in the SEM model may reduce the certainty of convergence. If mean structures are estimated in the SEM model, all means/intercepts should be set equal across time (mean stationarity) so that only one parameter is estimated. This specification will produce results that correspond more closely to results that are obtained from the traditional ARIMA analysis when the constant is included. Values on average should be very similar (Voelkle et al., 2012), but may differ for an individual analysis because of slightly different ML estimation algorithms used in the two packages.

AR(1). Specification of the first-order autoregression model follows the illustration in Figure 11.4. The variance of the first time point for the first measured exogenous variable, φ_{00}, is estimated freely and separately. Results from all models show that estimates from SEM and ARIMA procedure closely match. Parameter estimates suggest a significant autoregressive path of .994 and variance for the series equal to 40.114. The results suggest a strong inertia in consumer confidence from one month to the next.

MA(1). A separate first-order moving average model was tested following specifications in Figure 11.7 but with stationarity constraints on parameters. For this model, means for κ_1–κ_{163} were set equal to 0. Variances of the exogenous latent variables were set equal, ϕ_{11}–$\phi_{163,163}$, with a separate estimate of the variance at the first time point, ϕ_{00}. As seen in Table 11.2, estimates from the two analysis approaches closely match, although the moving average parameter is of opposite sign. This is due to the alternative parameterization of moving averages described earlier (Note 3). If the synchronous path, $\gamma_{t,t}$ was set equal to –1 instead of 1, the obtained estimates for the two approaches would be in the same direction. The estimate for the moving average parameter was .918, suggesting a similar weighting in combining the two consecutive values.

I(1). Although differencing or other transformations could be made prior to analysis, I illustrate how first-order differencing, or I(1) for the integrated process, estimates comparable to those obtained with traditional ARIMA analyses can be obtained with the single case SEM model as specified in Figure 11.11. Results in Table 11.2 are again similar for the two approaches. The mean estimate after differencing is –.425, which is small in comparison to its standard error, indicating that the average difference was not significantly different from zero.

ARIMA(1,1,1). Finally, the autoregressive, integrated, moving average processes were all combined in a single model. Specifications for the final model are depicted in Figure 11.14, but this model follows the basic elements of each of the separate models. As with the results from the separate models, the full ARIMA(1,1,1) model produced similar parameter estimates for both modeling approaches. Compared with the results from the separate models, however, the autoregressive and moving average parameter estimates were considerably different in the combined model. This is due to the non-independence of the three processes. Whereas the autoregressive parameter in the AR(1) model was .994, it was –.672 in the ARIMA(1,1,1) model. Moving average estimates changed from .918 in the MA(1) model to .714 in the full ARIMA model. These estimates could be more accurate to the extent that a first-order difference is appropriate, but they may be less accurate to the extent that overfitting or overdifferencing plays a role. A plot of the means is shown in Figure 11.15. This model was tested with pedagogical goals in mind and, in practice, further work would be needed to find the most appropriate processes for the data.

Comments

It should be noted that the conceptualization and testing of stationarity is considerably different for times series models with multiple cases from that for time series with a single case. The stationarity that I have discussed in connection with multiple-case time series analysis incorporates values based on many cases in the data set, where a statistical value, such as the mean, at a particular time point is an ensemble average. Single-case time series analysis involves inferences about the population of observations

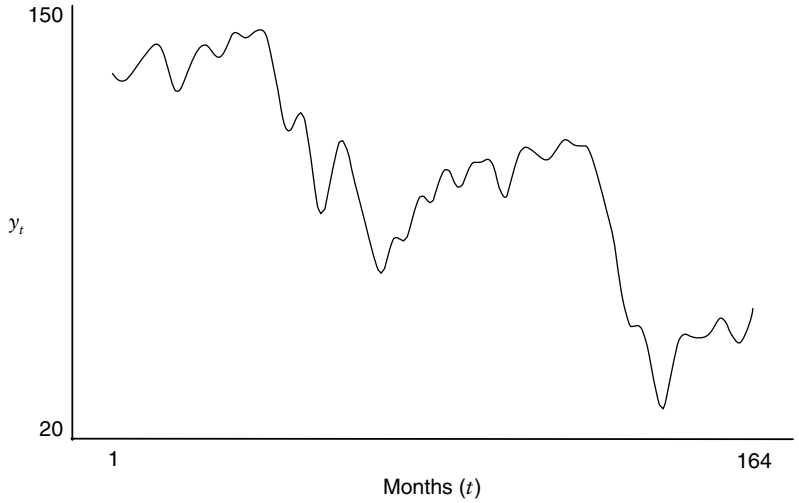

Figure 11.15 Plot of the Exogenous Means for the ARIMA(1,1,1) Model.

across an infinite series, where $T \rightarrow \infty$. The ergodic theorem states that the limit of the statistic from the single-case series will converge on the value of the ensemble statistic for an infinite series. The condition is that the series must be stationary and ergodic, where *ergodicity* refers to the condition that the autocovariance approaches zero as the lag increases (Hamilton, 1994). This statistical principle may be difficult to swallow for social scientists concerned about generalizability beyond a single case, but it is easier to imagine if observations represent an aggregate value based on many smaller units, such as monthly consumer confidence or quarterly national employment rates. With a structural modeling approach to time series analysis, the researcher has the best of both worlds with the ability to move between the analysis of a single case and the analysis of multiple cases.

Ergodicity can be evaluated with tests such as the Kolmogorov–Smirnov test, which can be conducted with single-case structural equation time series models (Voelkle et al., 2012). With multiple cases, estimates of means, variances, or autocovariances for each time point are derived from the probability distribution for a cross-sectional sample of cases for each time point and the need to test for ergodicity may be less important with larger sample sizes. Single-case analysis does not have probability information from a cross-sectional sample to borrow from in assessing stationarity. Information about stationarity must be derived from probability information in the time series for a single case. This is one reason a sufficient number of time points (e.g., ≥ 50) is recommended in the single-case time series (Box & Jenkins, 1970). Without a sufficient number of observations over time, there is insufficient probability distribution information to derive stable estimates of mean and variances (i.e., first and second moments). Voelkle and colleagues suggest that single-case SEM time series models will produce the best estimates when $T \geq 100$.

Multiple Indicator Time Series Models

Each of the models above can be extended to include multiple indicators for each time point. There are several advantages to incorporating latent variables into time series

models. One advantage is the availability of measurement invariance tests and constraints as detailed in Chapter 2. Invariance tests establish that changes in reliability are not mistaken for changes in the constructs of interest. Another important advantage is that latent variables can correct for measurement error. Although measurement error does not impact mean estimates, autoregression and autocorrelation estimates are affected by measurement error. Because time series models may involve several of these components, mean estimates may be distorted by measurement error because the moving average, integrated, and autoregressive processes are interdependent. Such distortions could lead to choosing the incorrect model. Inclusion of correlated measurement residuals among repeated indicators is an additional advantage that can substantially improve the accuracy of estimates of the true autoregressive process (see Chapter 4).

Specification of the multiple indicator model is a simple modification of any of the models illustrated above, where ξ_0 and η_1 through η_T are estimated by multiple indicators instead of single indicators. Identification of the latent variables requires a scaling constraint, using one of the approaches discussed in Chapter 1. The scaling constraints can have impacts on the mean estimates for the factor at each occasion and this may have complex effects on time series models when several processes are modeled. No identification method is more correct than another, but interpretations must be made consistent with the way the mean and variance of each latent variable are identified.

Multiple indicator models are also possible for single case time series models, where the specifications are closely related to the *p*-technique dynamic factor model (Hershberger, Molenaar, & Corneal, 1996). I do not illustrate multiple indicators for single-case time series models, but details on multiple indicator latent variables with single-cases analysis can be found in the work of Molenaar and colleagues (Molenaar, 1985; Molenaar, de Gooijer, & Schmitz, 1992).

Example 11.9: Multiple Indicator Model

A multiple indicator time series model using the diabetes positive affect data was tested to illustrate an ARMA model with multiple indicators. Three questions from the positive affect subscale were used for each affect factor, measured at each of the 24 time points. To demonstrate the general modeling technique a first-order autoregressive, first-order moving average, or ARMA(1,1), model was specified. This model followed the same structure as Figure 11.9. Each factor was identified with a referent indicator, with the loading set equal to 1 and the intercept equal to 0. Measurement invariance was assumed for the other parameters. Loadings for the two other loadings were set equal to their counterparts across all waves. Similarly, measurement intercepts for the two other indicators were set equal to their counterparts across all waves. Correlated measurement residuals were freely estimated each of the three indicators with their repeated measurements at all other time points. Stationarity was assumed for all variance parameters.

The fit of the model was poor, $\chi^2(1{,}769) = 8{,}713.285$, CFI = .637, SRMR = .224. Estimates of the autoregressive effect and moving average effects were .997 and –.607, both ps < .001. These parameters were similar to those obtained from the ARMA(1,1) model of the composite variables described in Example 11.5 although the model is based on latent variables at each occasion rather than observed variables using a composite index. A model with stationary means had a significantly poorer fit, χ^2 $(1{,}791) = 9{,}042.567$, p < .001, $\Delta\chi^2(22) = 392.282$, p < .001, though not a large magnitude difference, w = .161, ΔMc = .001. The test suggests that the means were not strictly stationary over time. Figure 11.16 plots the means. The relatively smooth line with no discernable trends suggests that the series is close to mean stationarity.

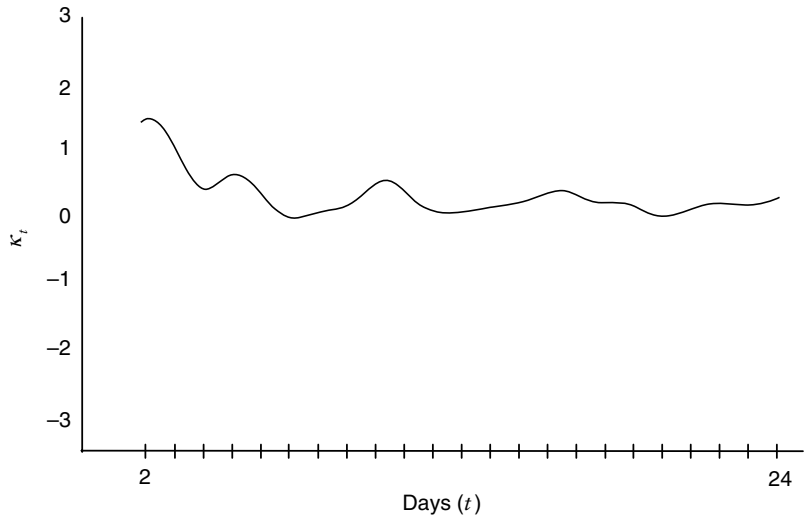

Figure 11.16 Plot of Exogenous Means From the Multiple Indicator ARMA(1,1) model.

Inclusion of Covariates

Covariates can be included in time series structural equation models in several ways. The most obvious approach would be to construct a concomitant time series model for another variable (i.e., multivariate time series) and then examine the correlation or predictive paths among random shock factors. Time series models have relative strength to answer causal questions in comparison to other analysis approaches used with shorter longitudinal designs, because spurious associations between variables due to corresponding trends or cycles among two variables hypothesized to be causally related are removed. Directional paths can be included at a simultaneous time point or a lagged time point. The determination of which lag length to use, however, should be made on the basis of theoretically reasonable criteria (Gollob & Reichardt, 1991). For example, if supportive social exchanges have a causal effect on positive affect among diabetic patients, then such an effect may be most likely on the same day or an adjacent day rather than several days or weeks prior. Inclusion of predictors in this manner has considerable methodological strengths, because prior values of the dependent variable are taken into account in a similar way to that in which cross-lagged panel models attempt to take prior time points into account (Chapter 4).

Time-varying covariates could also be incorporated by including predictive paths between a covariate and each observation or latent dependent variable in the time series model. As with latent growth cure models, the primary advantage of inclusion of time-varying covariates is to model the variable of interest after adjusting for the covariate. One particular application of use of a time-varying covariate is to investigate the effects of an intervention as with interrupted time series models (Meidinger, McDowell, McCleary, & Hay, 1980; Ostrom, 1990). A series of dummy variables can be employed to represent segments of the series before and after the introduction of an intervention. In addition to mean, variance, and covariance comparisons via equality constraints, an SEM approach can be especially useful for investigating differences in autoregressive, integrative, and moving average processes for points before and after the intervention.

Because time series can be analyzed within a structural modeling approach, a wide variety of options exist for including the time series in a larger model. The ability to estimate latent variables and therefore mitigate the attenuating effects of measurement error suggests that more accurate estimates of time series forecasts can be obtained. The SEM approach to time series also allows for the incorporation of time series into more complex models, examining the effects on distal outcomes or mediating pathways.

This chapter includes only a very basic introduction to time series analysis using an SEM approach. There are many valuable approaches, such as forecasting, not addressed here. Establishing a basic foundation of the essential time series within an SEM framework, however, is the first step in extending this approach to other realms.

Recommended Readings

Several books provide a comprehensive introduction to conventional time series analysis. Gottman's book (1981) is a clear and accessible discussion of many of the difficult concepts of time series analysis. Although fairly technical, the essential reference for time series analysis is the classic by Box and Jenkins (1970; Box et al., 2008). A very comprehensive but even more technical reference is by Hamilton (1994). There are several valuable introductions to logic, estimation, and specification of time series structural equation models for single-case designs with many time points (Hamaker et al., 2003; van Buuren, 1997; Voelkle et al., 2012).

Notes

1 The serial dependency problem is handled differently in growth curve analysis by the modeling of a multilevel error structure.

2 It is important to point out that the concept of autocorrelation differs in the multiple-case and the single-case sense. In the multiple-case sense, autocorrelation is the association taken across multiple units whereas autocorrelation in the single-case sense is assessed over a series of time points for just one case.

3 My notation differs from that of other sources (e.g., Hamaker, Dolan, & Molenaar, 2003). By specifying each measured variable as a latent variable with single indicator, there is greater flexibility for showing consistency with the AR model. The present notation makes it easier to generalize to models that incorporate multiple indicators.

4 It may help the reader to note that the formula for a moving average differs among authors. Some authors (e.g, Gottman 1981; Hamilton, 1994) express the observed value at time t as a sum of the random shock at time t and the weighted value of the prior random shock from time $t-1$ as I do in Equation (11.5). Other authors (e.g., Box & Jenkins, 1970), however, subtract the weighted value of the shock from the prior time point. For example, the Box–Jenkins expression for weighted averages is $\tilde{z}_t = a_t - \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$ for q time points. a_t is the random shock for a particular time point, a_{t-1} is the random shock for the prior time point, and θ_t is the weight parameter for the prior shock. The choice is an arbitrary one that does not impact any other aspects of the model, but the discrepancy could raise unnecessary concerns if SEM estimates are directly compared to ARIMA estimates obtained from other software (e.g., SAS's PROC ARIMA). The SEM moving average estimate will be of opposite sign from any software that uses the same parameterization as the Box and Jenkins formulation. The sign of the moving average parameter estimate from SEM output can be made to match the estimate obtained from traditional software by setting the value of the value of parameter for η_t regressed on ξ_t depicted in Figure 11.7 (which we could call $\gamma_{t,t}$) to –1 instead of to +1. In neither case is the moving average truly an "average," because the weights are not chosen to sum to 1. As with the autoregressive model, the intercept (constant) term can be added to these equations. In addition to these variants on the moving average specification, Browne and Nesslroade (2005) and Sivo and Willson (2000) specify

moving averages in terms of the value at time *t* plus observed values at subsequent time points instead of observed values at prior time points. This variation on moving averages is a more substantive difference, but in practice, with many time points, it may not make a large difference for the moving average estimates.

5 The Toeplitz covariance matrix approach, which involves analysis of a precomputed autocovariance matrix with values that converge toward zero for higher-order autocovariances rather than raw data, is less convenient than the more recent formulations of ARMA models (du Toit & Browne, 2001, 2007; Hamaker et al., 2002, 2003). In addition, the Toeplitz approach produces moment estimates rather than ML estimates, and the moving average estimates are biased.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Anderson, J. C., & Gerbing, D. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155–173.

Arbuckle, J. L. (1996) Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bell, W. (1987). A note on overdifferencing and the equivalence of seasonal time series models with monthly means and models with (0, 1, 1) 12 seasonal parts when Θ= 1. *Journal of Business & Economic Statistics*, *5*, 383–387.

Biesanz, J. C. (2012). Autoregressive longitudinal models. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 459–471). New York: Guilford.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York: Guilford Press.

Bowerman, B. L., & O'Connell, R. T. (1993). *Forecasting and time series: An applied approach* (3rd ed.). Belmont, CA: Duxbury Press,

Box, G., & Jenkins, G. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). *Time series analysis: Forecasting and control*. New York: Wiley.

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of ARMA time series models. In A. Maydeu–Olivares & J. J. McArdle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (pp. 415–452). Mahwah, NJ: Lawrence Erlbaum Associates.

Chatfield, C. (2000). *Time-series forecasting*. Boca Raton, FL: Chapman & Hall/CRC Press.

du Toit, S. H. C. & Browne, M. W. (2001) The covariance structure of a vector time series. In R. Cudeck, R., S. H. C. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 279–314). Chicago: Scientific Software International.

du Toit, S. H. C., & Browne, M. W. (2007). Structural equation modeling of multivariate time series. *Multivariate Behavioral Research*, *42*, 67–101.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128–141

Gollob, H. F., & Reichardt, C. S. (1991). Interpreting and estimating indirect effects assuming time lags really matter. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 243–259). Washington, DC: American Psychological Association.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge, UK: Cambridge University Press.

Hamagami, F., & McArdle, J. J. (2001). Advanced studies of individual differences linear dynamic models for longitudinal data analysis. In G. Marcoulides & R. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 203–246). Mahwah, NJ: Erlbaum.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2002). On the nature of SEM estimates of ARMA parameters. *Structural Equation Modeling*, *9*, 347–368.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. *Structural Equation Modeling*, *10*, 352–379.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Hershberger, S. L., Molenaar, P. C. M., & Corneal, S. E. (1996). A hierarchy of univariate and multivariate structural time series models. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 159–194). Mahwah, NJ: Lawrence Erlbaum.

Mehl, M. R., & Conner, T. S. (Eds.) (2012). *Handbook of research methods for studying daily life*. New York: Guilford Press.

Meidinger, E., McDowall, D., McCleary, R., & Hay, R. A., Jr (1980). *Interrupted time series analysis*. Beverly Hills, CA: Sage.

Mills, T. C. (1990). *Time series techniques for economists*. Cambridge, MA: Cambridge University Press.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202.

Molenaar, P. C. M., de Gooijer, J. G., & Schmitz, B. (1992). Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika*, *57*, 333–349.

Ostrom, C. W. (1990). *Time series analysis: Regression techniques*. Beverly Hills, CA: Sage.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Singer, H. (2010). SEM modeling with singular moment matrices Part I: ML-estimation of time series. *The Journal of Mathematical Sociology*, *34*, 301–320.

Sivo, S. A., & Willson, V. L. (2000). Modelling causal error structures in longitudinal panel data: A Monte Carlo study. *Structural Equation Modeling*, *7*, 174–205.

van Buuren, S. (1997). Fitting ARMA time series by structural equation modeling. *Psychometrika*, *62*, 215–236.

Watson, D., & Clark, L. A. (1994). *THE PANAS-X Manual for the Positive and Negative Affect Schedule – Expanded Form*. Unpublished manuscript, University of Iowa.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, *59*, 611–662.

Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum likelihood dynamic factor modeling for arbitrary N and T using SEM. *Structural Equation Modeling*, *19*, 329–350

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling, 15*, 377–402

Keywords

time series, ARIMA, ARMA, autoregression, autoregressive, moving average, integrated, differencing

12 Survival Analysis Models

Survival analysis is used to investigate prediction of a discrete event, such as a heart attack, death, product failure, marriage, college graduation, or company bankruptcy. It answers not only questions about the prediction of whether an event occurs but also when the event occurs. Survival analysis, which also goes by the interchangeable names of *event history analysis* and *failure time analysis*, addresses biases that may result when the period of observation ends before it is known for all cases whether or not the event will occur (Sørensen, 1977; Tuma & Hannan, 1978). If a certain model of car is recalled because of problems with the ignition, we cannot accurately estimate the proportion of cars that will have ignition failures or the predictors of those failures, because we have not examined the full life of the car. Later heart attacks are not known in a medical study that ends after five years. The event may occur just after the end of the observation period for some cases, long after the observation period for other cases, or never for other cases, resulting in biases in means or regression estimates and their standard errors if standard logistic, probit, ordinary least squares, or standard structural equation models were used. When the outcome is known for certain for all cases in a study, however, as would be true if the full life of the car or person has been observed, then there is no need for survival analysis.

Although survival analyses have been applied in many different disciplines, their implementation with SEM has received much less attention. Nonetheless, survival analyses can be tested using structural equation models, offering many of its usual advantages, such as incorporation of latent variables, complex path models, missing data estimation, or multiple survival processes, among others. This chapter introduces basic survival analysis concepts and two common approaches to estimating survival analysis with structural equation models.

Basic Concepts

Survival analysis requires information about whether and when an event occurs. I will focus throughout on the situation when there is only one possible event outcome (i.e., binary case) and the event occurs only once (or only its first occurrence is of interest). Survival analyses are also possible for multiple outcome categories (competing risks) or repeat events, variations that could be addressed with SEM as well. In either case, the event in question is not a variable of an ordinal nature, such as the level of health, an attitude measure, or financial success.

There are two general approaches to survival analysis commonly used in the social sciences, *discrete-time survival analysis* and *continuous-time survival analysis*, distinguished by the precision of the measurement of the timing of the event. When the time to the event is not known precisely and is measured in crude increments, such as whether the event occurred between interview intervals of a panel study, discrete time survival analysis is

used. When more precise timing of the event is known, such as the days, weeks, or months until the event occurs, continuous time survival analysis is used. The most common form of continuous-time survival analysis is *Cox regression,* which is the method discussed in this chapter. The distinction between discrete and continuous time is less than definitive and represents a matter of degree. With greater measurement accuracy of the time of the event, further statistical precision can be gained. Generally, when there is considerable variation in the time of the event across cases, as with age of first marriage, even if measured somewhat crudely, the time until the event is considered to be continuous. In contrast, when there are many cases that share the same event time, such as the number of academic years it takes to complete college, the time until the event is considered to be discrete.

Censoring

A key concept of survival analysis involves the premature end to the observation period for any case. A case has been *censored* if the observation ends before it is known whether the event will occur or not. Because the uncertainty falls at the right side of the figurative time line, this type of censoring is known as *right censoring*. Right censoring is the primary focus of most commonly employed survival analysis methods. If it is not known whether an event occurred prior to the beginning of the study, such as a prior pregnancy in a study of teenage pregnancies, the term *left censoring* is used. *Interval censoring* refers to uncertainty about the exact timing of the event, where it is only known that the event occurred between two ends of a time interval. Many applications of discrete-time survival analysis involve an event in which the timing of the event varies within a particular interval (e.g., reported diagnosis that occurred sometime since the prior annual interview) and are therefore interval censored. Such data are commonly treated as discrete even though this method assumes no variation in the timing of the event occurring within a particular interval. Although methods specific to addressing left and interval censored data have been developed (e.g., Collett, 1994; Farrington, 1996; Finkelstein, 1986; Kim, 2003), left censoring and interval censoring will not be covered here.

In many studies, the observation period ends for all cases at the same time, so the right censoring occurs singly. Many studies involve censoring at different times for different cases, however. Attrition implies a certain kind of right censoring if the event has not occurred prior to dropping out of the study. Because different individuals drop out from the study at different times, the time of right censoring varies across cases. Whenever the censoring is unrelated to the probability of the occurrence of the event after controlling for other factors in the model, it is considered to be *uninformative censoring* (Cox & Oakes, 1984). When the probability of the event's occurrence is independently related to whether or not the case has been censored, it is considered to be *informative censoring*. For example, if those who drop out of a study about psychiatric hospitalization are more likely to be hospitalized after the study loses track of them, then those who stay in the study may be at less risk of hospitalization than those who drop out. Informative censoring is based on the same principles as the missing at random definition or the ignorable/nonignorable distinctions used more generally (Rubin, 1976; see Chapter 13 for more detail). The idea is that if those who are censored differ from those who are not censored, once variables in the model are taken into account, biases may result. That censoring is uninformative is an implicit assumption in most general survival analysis methods but cannot be known from the observed data. In the psychiatric hospitalization example, unless further tracking is done or hospitalization records are obtained for dates after participants drop out, there is no information available to determine if those who drop out are at greater risk of psychiatric hospitalization than those who complete the study.

Hazard

Survival analysis is concerned with the likelihood that an event will occur in a particular time period. It is easiest to consider the discrete case in which there are several separate intervals of observation, such as annual interviews in a typical longitudinal panel survey. The concept of the *hazard rate* (or just "hazard") involves the risk that an individual experiences an event for one of the intervals of the study. Expressed as a proportion, the hazard rate is calculated by using the total number of cases at risk as the denominator. The total number at risk excludes individuals who have already experienced the event and individuals who have been censored.

Using a more formal statement, the hazard function is the probability that an event will occur for a case at a certain time point, given that it has not already occurred for that case. If t is a particular time point for the study, the hazard at a particular time point for an individual case is h_{ti}, and the time that the event occurred is $\mathcal{T}_i$. The hazard is

$$h_{ti} = P\left(\mathcal{T}_i = t \mid \mathcal{T}_i \geq t\right) \tag{12.1}$$

which simply states that the hazard is a conditional probability of the event occurring for any case given it had not occurred earlier.[1] To estimate the expected probability that a randomly selected individual experiences the event for a certain period, we can compute the proportion of cases experiencing the event divided by the number of cases that are at risk for experiencing the event for that period. The number of cases "at risk" for a given time point, or *risk pool*, is the number of cases who have not experienced the event and have not been censored (e.g., have not dropped out).

As a brief illustration, say the event of interest is a reported heart attack in an annual longitudinal health survey. If 10 cases report a heart attack in Year 3 of the study and there were 250 who responded to the survey at that wave, then the hazard proportion for that wave would be 10/250 = .04. If everyone was at equal risk, a randomly chosen respondent would have a 4% chance of reporting heart attack during that year. Hazard plots, with the hazard on the *y*-axis and time period on the *x*-axis, can be useful to visualize whether there are increasing or decreasing trends in the risk over time or whether particular periods have particularly high or low risk. Variations on the hazard plot include use of the cumulative hazard function or a smoothed hazard function instead of the raw hazard. A plot of the cumulative hazard function can be a helpful aid for visualizing whether the hazard rate accelerates or decelerates over time. A constant hazard rate over time would have a positive linear slope, whereas an accelerating or decelerating rate would have a nonlinear cumulative hazard plot.

Survival

The survival probability refers to the chance that a case will not have experienced the event up to and including a certain interval. Thus, the *survivor or survival function* is the cumulative proportion of those still in the risk pool and has a complementary relationship to the hazard proportion.

$$\begin{aligned} S_{ti} &= P\left(\mathcal{T}_i > t\right) \\ &= \prod_{1}^{t}\left(1 - h_{ti}\right) \end{aligned} \tag{12.2}$$

The last line of Equation (12.2) indicates that the probability of survival can be computed from the products (the Π symbol) of the preceding probabilities of not experiencing the event. As with the hazard, the probability of a randomly selected case not experiencing

Table 12.1 Observed Hazard and Survival Proportions for a Major Health Diagnosis in the Social Exchanges Data Set

t	*No diabetes*	*Diabetes*	*Censor*	*Hazard*	*Survival*
1	5,671	0	0	0	1.000
2	5,170	338	163	.061	.939
3	4,707	245	240	.049	.892
4	4,227	258	260	.058	.841
5	3,678	258	358	.066	.786
6	3,106	217	477	.065	.734

the event until a certain point is computed from the proportion of cases that have not experienced the event but are still in the risk pool. In other words, the number of cases that have experienced the event and those that are censored are removed from the total number of cases when the survival proportion for a period is calculated. If no cases are censored until the end of the study, then the denominator is simply the total number of cases in the study at the beginning.

This method of computing the survivor function is the same as the Kaplan–Meier estimate (Kaplan & Meier, 1958) for the discrete case (Efron, 1988). The *median survival time*, in which the survival risk dips below .50, gives the expected time it takes for the event to occur, also a useful tool for characterizing the survivor function. To examine the rate of depletion of the risk pool, a survival curve, using the survivor function, log of the survivor function, $-\ln[S_{ti}]$, or the log-log of the survivor function, $\ln[-\ln(S_{ti})]$ on the *y*-axis, may be plotted against time, with the survivor function on the *y*-axis and time intervals on the *x*-axis. The natural log, or log base *e,* is given by ln. The latter two plots are more often used in the continuous time case to investigate distributional assumptions. A summary of the hazard and survival probabilities at each time point, called a *life table*, is a standard step for describing any event data.

Example 12.1: Life Table with Hazard and Survival Proportions

Table 12.1 is a life table that presents data on diabetes diagnosis from a new version of the health and aging data set (described at the beginning of the book Example Data Sets section at the beginning of the book). This version of the data set has a different sample size from the version used in other chapters because those who did not report diabetes at the first wave were eliminated. Intermittent missing data in which an individual did not respond to the survey at one wave but responded at later waves also were eliminated (see Chapter 13 for more information about this type of missing data and methods of handling it). The main reason for eliminating missing data in this instance is that it is unknown if an individual was diagnosed during the period with the missing information. Syntax and data sets used in the examples are available at the website for the book.

The first row of Table 12.1 shows that there were 5,671 respondents who had not reported diabetes at baseline. The next row shows that, at the second interview, 5,170 did not report diabetes, 338 reported diabetes, and 163 were missing. The risk pool at $t = 2$ was thus $5{,}170 + 338 = 5{,}508$, so the hazard rate at this wave was $338/5{,}508 = .061$. Similarly at $t = 3$, the hazard rate was calculated by dividing the risk pool by the number of new cases reporting diabetes, or $245/(4{,}707+245) = .049$. (Note: in this chapter I begin the first time point with $t = 0$ to be consistent with usage in this domain.) This latter hazard rate indicates that during the third interview there was a 4.9% chance that a randomly chosen individual

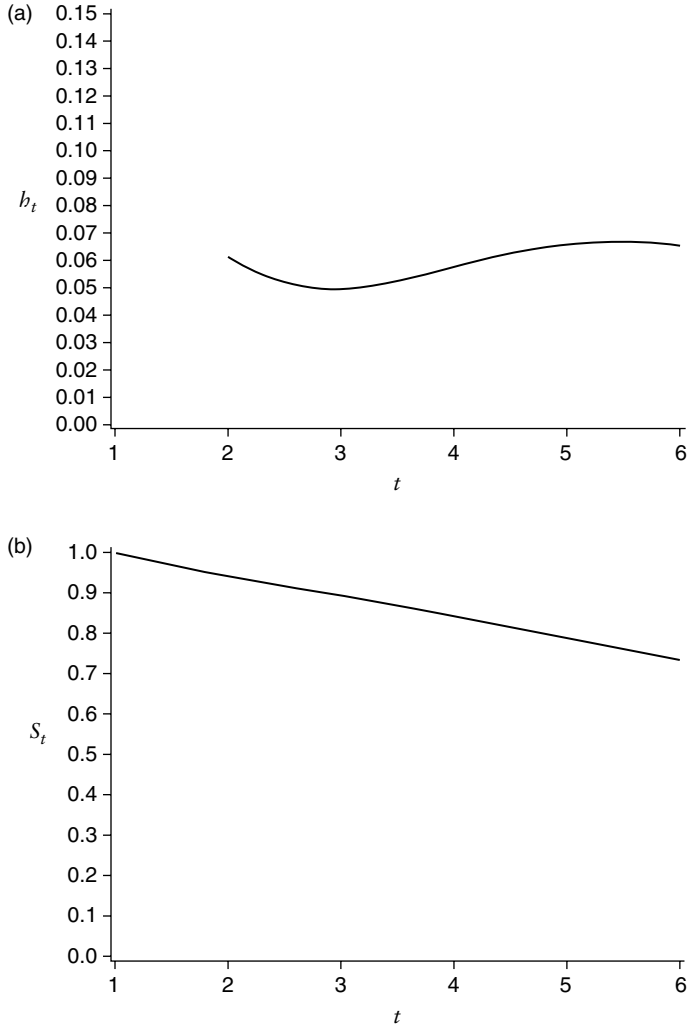

Figure 12.1. Plots of (a) Hazard and (b) Survivor Function for Diabetes in the Health and Aging Data Set.

who had not previously reported diabetes would report the disease. The survivor function is calculated according to Equation (12.2), which, for example, shows that the probability of survival at $t = 4$ was $S_4 = (1-h_4)(1-h_3)(1-h_2)(1-h_1) = (.939)(.931)(.942)(1) = .841$. The estimated chance that any randomly chosen respondent will "survive" until the fourth interview without being diagnosed with diabetes is about 84%.

Figure 12.1 presents smoothed plots of (a) the hazard function and (b) the survivor function. The hazard plot suggests that the rate of newly reported diagnosis is relatively flat over time, with a dip at the third wave, and a slight rise overall. The risk of (reported) diabetes, therefore, does not change very much over time. The survival curve exhibits a fairly linear decline, a consequence of the relatively stable hazard rate. Had the hazard rate

increased over time, then the survival curve would have shown a nonlinear (decelerating), rather than a linear, decline.

Discrete Time Survival Analysis

Discrete time survival analysis is suitable for questions about the timing of the occurrence of an event for which there are a reasonable number of time points or intervals available. Panel studies with structured data collections, such as monthly, annual, or biennial waves, represent a typical example. What constitutes a reasonable number of time points is really up to the analyst. As the number of time points increases, the necessary coding and the estimation process potentially become more unwieldy. As the number of time points increases, there is greater precision in the timing of the event and continuous methods such as Cox regression may be more convenient. When data are less structured and there are few cases with common event times, then discrete methods become more cumbersome, impractical, or impossible. The discrete time survival approach is flexible in how covariates may be incorporated, with a simple and straightforward way of including time-invariant and time-varying covariates.

Logistic Regression Approach

The discrete time survival approach, which was originally proposed by Cox (1972) who also proposed the widely used continuous-time survival model, was popularized in part because survival models can be estimated with logistic regression, a procedure which was much more widely available in computer software programs than other survival analysis procedures. The general logistic approach to discrete survival analysis can be conveniently adapted for estimation in any structural modeling package that handles binary measured variables. Though approaches to survival analysis for SEM had been suggested earlier (Pickles et al., 1994; Vermunt, 1997), Masyn's general approach to discrete time survival analysis (Masyn, 2003; Muthén & Masyn, 2005), which also incorporates latent class analysis, has been the most widely adopted. As this structural modeling approach is based on the logistic regression modeling approach, I will discuss the logistic approach to discrete survival analysis briefly first.

The logistic approach to discrete time survival analysis reconfigures the data into a separate record for each repeated measurement, referred to as a *person-period* or long format. The structural equation model approach, discussed shortly, does not require this reconfiguration, however. When data are reconfigured with a separate record for each time period, a special logistic regression model is used to represent the probability of an event occurring at each time point. The fact that the intercept for any logistic regression model relates to the proportion that the outcome is equal to 1 can be exploited to model the hazard at each wave when the data are configured into the person-period format. A series of dummy variables that indicate whether or not the event occurred at each time point can be used to substitute for a usual single intercept. A simple model with no covariates is stated as

$$\text{logit}\left(h_{ti}\right) = \ln\left(\frac{h_{ti}}{1-h_{ti}}\right) = \beta_2 D_{2i} + \beta_3 D_{3i} + \cdots \times \beta_t D_{ti} \tag{12.3}$$

The dummy variables in Equation (12.3), D_{2i}, D_{3i}, …, D_{ti}, are binary indicators for each time point, each equaling 1 for its respective time point and 0 for all other time points. Betas are used as they are throughout the text to refer to unstandardized regression coefficients. The dummy variables begin with D_{2i} in this notation, because the first opportunity

experiencing the event begins with the second time point under the presumption that the baseline measurement uses only cases to which the event has not occurred yet.

To estimate the discrete survival model with logistic regression, the person-period data set is analyzed. The binary event serves as the outcome in the model with the full set of time point dummy variables as the predictors. Notice that Equation (12.3) has no intercept. The usual intercept is omitted, because the mean information normally captured by the intercept is partitioned by the set of dummy variables instead. (The software must have an option for excluding the intercept from the model.) The data set includes only records for cases still in the risk pool at each time point, so there is no record included for a time point if the event has occurred at a prior time point for that case or that case has been censored from that point forward (e.g., dropped out of the study). The records can be omitted by coding the outcome or the dummy variables as missing for records where the event has already occurred or has been censored, because logistic procedures will omit any cases with missing data (i.e., listwise deletion).

Regression slopes for each of the dummy variables provide information about the hazard for each time point. Specifically, they represent the log-odds increment over one time point in the hazard. Transformation of the regression coefficients using the usual logit transformation,

$$h_{ti} = \frac{1}{1 + e^{-\beta_t}} \tag{12.4}$$

returns the hazard rate for a particular time point. This analysis approach is based on the logit link function, but the complementary log-log link is an available alternative for discrete data (Prentice & Gloekler, 1978; see Allison, 2010 for an introduction to this approach). A different transformation is required to obtain the hazard estimation, however, where $h_{ti} = 1 - \exp\left[-\exp\left(\beta_t\right)\right]$ is used instead of Equation (12.4). A straightforward extension of the discrete-time survival model is possible by adding any number of time-varying or time-invariant covariates.

SEM Approach

Understanding the logistic regression approach to discrete survival analysis is important, because the same rationale can be adapted for estimation with an SEM approach (Masyn, 2003; Muthén & Masyn, 2005). The basic unconditional discrete survival model, equivalent to that obtained with the logistic regression approach outlined above, can be estimated with a very simple structural equation model that uses maximum likelihood (ML) estimation for binary variables. A standard person-level data set with repeated measurements for each record is analyzed (i.e., data are not reconfigured into a person-period format). Binary indicators of the first occurrence of the event serve as indicators of a latent variable, and hazard values for the unconditional discrete survival model are obtained from the threshold estimates for each occasion-specific event indicator.

To translate the logistic model to a structural model, we need a set of binary repeated indicators, y_t, representing the first or only occurrence of the event. For each case, a $T-1$ set of y_{ti} indicators are coded 0 if the event has not occurred, 1 for the first occurrence of the event, and missing if the event has already occurred or the case has been censored. The probability for those in the risk pool can then be stated as

$$P\left(\mathcal{T}_i = t_i\right) = P\left(y_{ti} = 1\right)\prod_{a=2}^{t_i - 1} P\left(y_{ai} = 0\right)$$

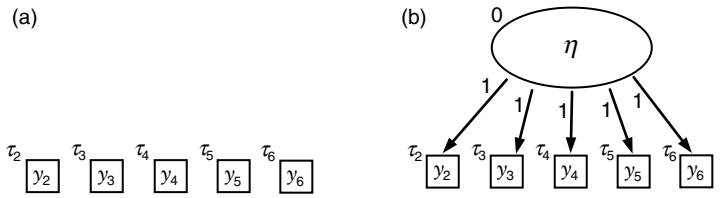

Figure 12.2 Specifications of the Unconditional Discrete Time Survival Model (a) Without Latent Classes or Continuous Factor and (b) with a Single Continuous Factor.

The equation says that the probability that the event occurs at occasion t (i.e., that the time of the event $\mathcal{T}$ is the same as t_i) is equal to the probability that y_{ti} equals 1 times the product of the probabilities the event has not occurred across all previous occasions, a, where $a < t$. Only time points $t = 2, \ldots T$ are included in the model because it is assumed that no events have occurred at the first time point, $t = 1$, and all cases are at risk for the event on the first occasion.

An individual likelihood of this function can then be given for those individuals at risk, where all values after the occurrence of the event or values that have been censored are excluded, resulting in a full ML for the sample that uses only the available data (Masyn, 2003; Muthén & Masyn, 2005). Under the condition that the excluded values are missing at random, the estimates will be equivalent to the logistic estimates for the discrete survival analysis under the assumption that censoring is noninformative. Estimation for both approaches is based on the available data for each case – only the variables for the observed values are included in the SEM approach and only the records for the observed values are included in the logistic approach. Because missing data are integral to the survival estimates, only estimates that make use of full information, such as binary ML or robust binary ML, should be used. WLSMV may not be the optimal approach because of the way missing data are handled in the presence of covariates, particularly if missing values may not meet the missing at random assumption (see Chapter 13 for more detail).

Specification. The unconditional discrete survival estimates can be obtained from two different structural modeling specifications. Figure 12.2 illustrates. In many applications, a discrete time survival model is estimated using a single latent class variable (discussed below), but this is not required to obtain the unconditional hazard estimates. Figure 12.2a shows the simplest model estimated for six time points with five binary event indicators, y_2–y_6. Figure 12.2b shows a single continuous factor with loadings set equal to 1 and factor variance set equal to 0. This specification is useful for implementing the proportional hazard assumption when covariates are included. Only five indicators are needed to model six time points if no events have occurred at baseline. In other words, the full sample is at risk for the event at the first time point, an approach that also implicitly assumes no left censoring. To be consistent with discrete-time survival analysis with logistic regression, binary ML estimation is assumed, so no measurement errors are depicted in the figures.

For all three specifications, threshold estimates associated with the $T - 1$ binary event variables can be transformed into hazard probabilities for each time point. Recall that under ML estimation for binary variables the threshold can be converted to the observed probability using the logit exponential transformation (see Chapter 1).

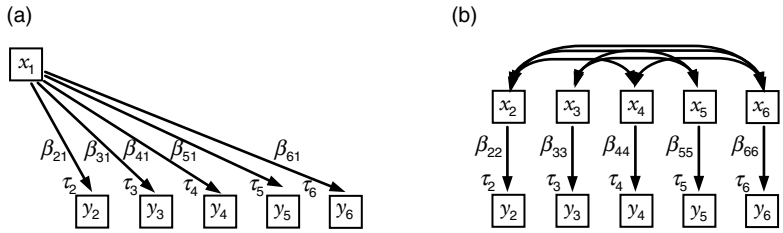

Figure 12.3 Discrete Time Survival Models with Covariates (a) Time-Invariant Covariate and (b) Time-Varying Covariate.

$$P(y=1)=\frac{1}{1+e^{\tau}} \tag{12.5}$$

In this equation, the regression coefficient from the logistic model that appears in Equation (12.4) is replaced by the threshold in the denominator on the right-hand side of the equation. This quantity is usually written as the negative of the parameter when stating the logistic transformation, but because the thresholds from the measurement model are reported as $-\tau$, the quantity $e^{-(-\tau)}$ simplifies to just e^{τ}.

Covariates. Two types of covariates may be included in the discrete survival model (depicted in Figure 12.3). Time-invariant covariates involve a single measurement of each case, whether the variable represents a stable characteristic or not. Most often the measurement of the time-invariant covariate will correspond with the baseline to model the effects on the risk of the event. High school GPA might be used to predict the time to graduation, for example. The hazard ratio can be computed by exponentiating the regression path estimate, e^{β}, and represents the odds of a unit increment in the covariate.[2] The discrete time approach to survival analysis also allows for the inclusion of time-varying covariates. Paths from each time-varying covariate involve synchronous effects, with each y_t regressed on each x_t, or lagged effects, with each y_t regressed on each x_{t-1}.

The goal of including a time-invariant covariate is often to explain the risk of the event whereas the goal of including a time-varying predictor is often to remove the effects of the covariate from the estimate of the hazard. As with latent growth curve models and other contexts, careful attention to the scaling of the covariate is advisable. Interpretation of the hazard depends on the zero-point of the covariate. For many covariates, a value of 0 provides a meaningless or undesirable point of interpretation for the hazard, and it will be useful to center the covariates around their means (see Chapter 7 for more about inclusion of time-invariant and time-varying covariates and covariate scaling). Although measured covariates are illustrated here, a distinct advantage of the SEM approach is the potential inclusion of latent variables. The estimated effects of latent variables with multiple indicators can mitigate attenuation due to measurement error, thereby increasing bivariate effects and improving the accuracy of statistical control when multiple covariates are included.

Proportional Hazards Assumption. Many survival analysis methods, including discrete time survival analysis, assume that the covariates have the same hazard function at each level. The *proportional hazard assumption* states that the hazard function considered across time applies equally well to each case where the ratio of one case's hazard relative

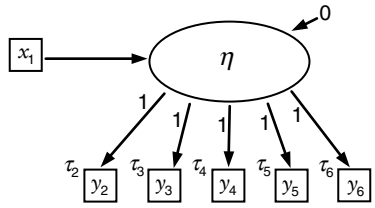

Figure 12.4 Proportional Hazard Constraint using a Single Latent Variable.

to any other case maintains the same proportion over time. Discrete-time survival analysis using logistic regression assumes proportional odds rather than proportionality of the raw hazard (Singer & Willett, 2003). The *proportional odds assumption* requires that the odds of the hazard must be parallel over time. More formally we could write $e^{h_{t(i=1)}} / e^{h_{t(i=2)}} = c$, where c is a constant value across time for any two cases (e.g., $i = 1$ and $i = 2$). The proportional hazards or proportional odds assumptions are commonly considered in the context of whether two or more groups have parallel hazard functions over time. Although the risk of one group (e.g., heart attacks for men) may be higher than another group (e.g., women), it is the ratio of risks in the two groups that must remain constant over time if proportionality holds. Even though the risk of a heart attack increases with age, the proportional odds assumption states that the ratio of the odds that a man will have a heart attack is proportionate to the odds that a woman will have a heart attack as they age.

The hazard function for the 0 group (e.g., men) is called the *baseline hazard*. "Baseline" in this setting refers to using one group as reference of comparison for the other groups and does not refer to the hazard at the beginning of the study. In a more general sense, the proportional hazard assumption suggests that the predictor has the same relation to the hazard across time.[3] In the continuous covariate case, the baseline hazard refers to the hazard function when the covariate is equal to 0; or, if there are multiple covariates, the baseline hazard is the hazard function when all covariates are equal to 0. Again, attention to how covariates are scaled is needed when interpreting the baseline hazard, and it may be desirable to center covariates as a default approach.

Proportionality assumptions can be explored graphically by plotting hazard values, threshold estimates (logit), or odds over time (for illustrations see Allison, 2010; Singer & Willett, 2003). The feasibility of the proportional odds assumption can be investigated by comparing a model in which the effect of the covariate is constrained to be equal across time (e.g., $\beta_{21} = \beta_{31} = \beta_{41} = \beta_{51} = \beta_{61}$ in Figure 12.3a) to a model in which the effect of the covariate is allowed to vary over time (Muthén & Masyn, 2005). Notice that the inclusion of a covariate (as in either depiction in Figure 12.3) estimates the baseline hazard, because logistic transformation of the thresholds gives the hazards when each covariate is equal to 0. The test is a comparison of whether the hazard function remains constant over time when the covariate is changed by one unit. For a binary covariate, the unit change is the same as the group difference.

When the covariate is time-invariant, a convenient and equivalent method of setting the proportionality constraint, illustrated in Figure 12.4, is to specify a single continuous latent variable (Muthén & Masyn, 2005; Voelkle & Sander, 2008). In this approach, the loadings are set equal to 1 and the latent variable variance is set equal to 0. The thresholds estimates are equal to the values obtained from the specifications shown in Figure 12.3a when longitudinal equality constraints are imposed on the paths and can be transformed to obtain the baseline hazard. Removing the constraint on the loadings

relaxes the proportionality assumption and can be compared to the model with the loadings set to 1. The single-continuous latent variable specification is especially convenient for inclusion of multiple time-invariant covariates.

Unobserved Heterogeneity. The problem of *unobserved heterogeneity* in survival analysis models concerns a kind of selection bias in which a there is a declining risk of the event later in the study, because cases with the greatest propensity for the event are eliminated from the risk pool early. If a product such as a washing machine is essentially durable and lasts for many years, but there are some models with defective parts, the premature failure of the defective models may slant interpretations about the failure rates of all washing machines. The defective machines are eliminated early from the risk pool and the hazard rate decreases at later time points because the remaining machines are less likely to fail. The problem is one of omitted variables because an unidentified variable exists that could account for the variation in the hazard function across cases if it were to be included in the model. The concern is that a declining hazard can be interpreted as representative of the general risk of the event for all cases when there is an important source of variation that, if it could be modeled, would lead to a substantially different interpretation of the probability of failure or of survival over time. Yet another, more general, way of viewing unobserved heterogeneity is as variance unaccounted for in the model (Andress, 1989; Heckman & Singer, 1984), which may be a result of covariate measurement error as well as omitted variables.

A mixture modeling framework is well-suited to exploring unknown sources of variance in the hazard function. Each of the models described thus far can be extended by estimating a latent class factor for unobserved groups (introduced in Chapter 10). Figure 12.5 uses a visual analogue to illustrate estimation of latent classes in three different variations of the approach. In the specification of these models, the latent class factor does not actually directly predict the variables in the model. Instead, separate parameter estimates are obtained for intercepts (log hazards) or other parameters for each of the unobserved groups. Latent classes, which categorize cases into a specified number of unknown discrete classes, can be used to identify subsets of cases in the data set that have distinct hazard functions. Figure 12.5a shows the specification with a single latent class factor, η^C, for modeling unobserved heterogeneity; Figure 12.5b shows the single latent class factor combined with a single continuous factor for modeling unobserved heterogeneity with proportional hazards assumption; and Figure 12.5c shows the single latent class factor combined with a single continuous factor and a time-invariant covariate.

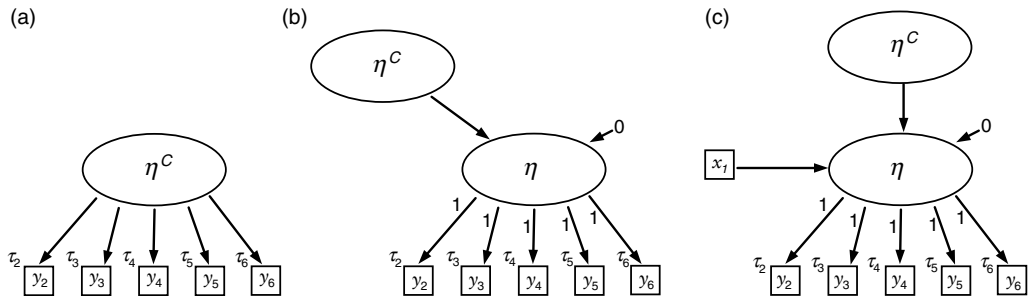

Figure 12.5 Inclusion of Latent Class Factors for Modeling Unobserved Heterogeneity (a) Basic Model, (b) with a Single Continuous Factor for Implementing the Proportionality Assumption, and (c) with a Covariate.

An advantage of the latent class approach is that the classes are derived from the data, thus potentially accounting for a mixture of hazard distributions (Masyn, 2003). It is data driven to the extent that the groups are unmeasured in the data, unlike multigroup analysis, which specifies known groups such as gender. It is confirmatory in the sense that the number of classes must be specified in the model and models with different numbers of classes can be statistically compared (Chapter 10 discusses and illustrates these model comparison tests). If the fit of the model with a single class is better than the fit of the model with two or more classes, then there is some supporting evidence that unobserved heterogeneity is not muddling the conclusions.

The utility of incorporating a latent class variable goes beyond simply identifying different hazard functions. Heterogeneity in the effects of time-varying or time-invariant covariates can also be modeled. If a certain washing machine feature, such as a steam cleaning cycle, is an important predictor of the machine's failure risk, the strength or even direction of its association with the hazard rate may differ across models. Specification of two or more classes allows differences among unobserved classes in the effects of the covariates. This general mixture modeling framework then affords considerable flexibility for the discrete survival model. Depending on the particular model specification, latent classes may be used to account for variation in the baseline hazard, the hazard function assuming or not assuming proportional hazards, the hazard for specific segments of timeline, missing data or dropout patterns, and the effects of covariates (for illustrations for several possible applications, see Muthén & Masyn, 2005; Voelkle & Sander, 2008). Predictors of latent classes themselves can also be added to the model in an attempt to explain estimated class membership.

Example 12.2. Discrete Time Survival Models

The health and aging data set was used to illustrate the SEM approach to discrete-time survival analysis. Indicators of self-reported diabetes diagnosis for Waves 2 through 6 were used. The coding for a sample of 20 respondents is illustrated in Table 12.2. Wave 1 responses are not shown, because all those reporting diabetes at this time point were eliminated, which results in a constant value. Respondents with intermittent missing data, in which one or more values at earlier time points were missing and subsequent time points were present, were also eliminated to simplify the example. Consider several patterns of codes in the table. There are individuals with no missing codes (e.g., $i = 2$) who do not drop from the study or report diabetes prior to Time 6 for which all codes are equal to 0. Individuals who do report a diagnosis prior to the last time point (e.g., $i = 18$) have missing codes, denoted by ".", for all time points after first reporting the event (indicated by a code equal to 1). Individuals who drop from the study have missing data for all time points starting with the time of the dropout ($i = 6$). These cases are censored cases.

An unconditional discrete time survival model was tested initially using robust ML for binary data with missing data estimation ($N = 5{,}508$). I used the simplest model specification depicted in Figure 12.2a, which estimated thresholds for five event indicators, but the specifications illustrated in Figures 12.2b, 12.5a (using a single class), or 12.5b (using a single class) would all produce identical results. Results yielded threshold estimates of 2.728, 2.956, 2.796, 2.657, and 2.661 for Waves 2 through 6, respectively. Inserting each of these values into the logistic transformation in Equation (12.5) yielded hazard values of .061, .049, .058, .066, and .065, indicating approximately 6% of cases who had not reported the disease earlier are at risk at any wave for reported diabetes. Plugging the hazard values into Equation (12.2) gives survival probabilities of .939, .892, .841, .786, and .734. Smoothed hazard and survivor functions are plotted

Table 12.2 Event Indicator Coding for a Diabetes Diagnosis in a Sample of 20 Respondents

i	*t* = 2	*t* = 3	*t* = 4	*t* = 5	*t* = 6
1	1	.	.	.	.
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	.	.
7	0	0	0	.	.
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
11	0	0	0	0	0
12	0	0	0	0	0
13	0	0	0	0	0
14	0	.	.	.	.
15	0	0	0	0	0
16	0	0	0	0	0
17	0	0	0	0	0
18	0	1	.	.	.
19	0	0	0	0	0
20	0	0	0	0	0

in Figure 12.6. The hazard function shows a dip at $t = 3$ but is otherwise fairly steady over time. The steady rate translates into a nearly linear decrease in the survivor function over time, suggesting that the survival probability declines linearly rather than at an accelerating rate.

To illustrate a model that includes a time-invariant covariate, a dichotomized version of the age variable was created, with ages 50–65 = 0 and older than 65 = 1. Although continuous age that has been centered would likely be preferable in nearly all instances, the dichotomized version helps illustrate a baseline hazard. The specification of the survival model followed Figure 12.3a where each diabetes indicator was predicted by the age category variable. The robust ML estimator was used. An initial version of the model constrained all paths from the age variable to the diabetes indicators at each time point to be equal. This imposes a proportional hazard constraint. No chi-square is available for this model, but the –2LL value was –4,987.541 and the sample size adjusted BIC (aBIC) was 10,007.700. The single estimate of the path to the indicators was .038, which was nonsignificant. Exponentiating this value gives the odds ratio for age $e^{.038} = 1.039$, indicating an approximately 4% greater risk of diabetes in the older age group. The threshold values for this model were 2.728, 2.956, 2.797, 2.658, and 2.662 which correspond to hazard values of .061, .049, .057, .065, and .065. These hazard values are the baseline hazard function for the 50–65 age group, because the younger group was coded as 0. Hazard values for the older age group could be obtained by reversing the coding of the two age categories. Observe that these baseline hazards are very similar to those obtained for the model without a covariate, suggesting that there was little difference between the two age groups in the hazard ratio across time. Indeed, a subsequent model allowing the paths from the age variable to vary over time had a very similar –2LL value (–4,986.440) and aBIC (10,027.243). A naïve likelihood ratio test can be used to compare the two –2LL values (Muthén & Masyn, 2005). With a difference of four parameters in the two models, the likelihood ratio chi-square would not be significant, $-2LL_{M0} - 2LL_{M1} = 1.101$, ns.

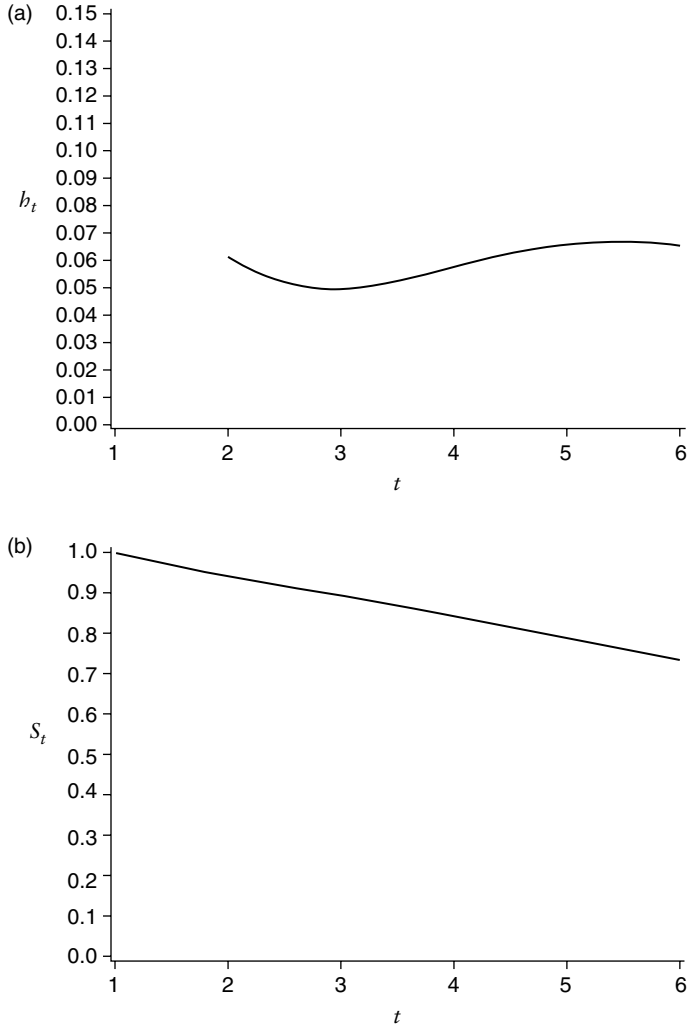

Figure 12.6 Diabetes (a) Hazard and (b) Survivor Functions Across Time.

The use of time-varying covariates was illustrated by including body mass index (BMI) measured at Waves 2 through 6 as synchronous predictors of each diabetes indicator. Each of the BMI values were centered to obtain estimates of the thresholds for diabetes for individuals who had an average BMI value. The paths were constrained to be equal across time points, but this constraint could be relaxed if desired. The intercept values from this model were 2.802, 3.022, 2.854, 2.704, and 2.701, corresponding to hazard values of .057, .046, .054, .063, and .063.

To investigate whether unobserved heterogeneity might impact the hazard function, a discrete time survival model of diabetes diagnosis that included (centered) continuous age as a time-invariant covariate was tested. The proportional hazard constraint approach was used, with the age variable predicting a single continuous latent variable. The continuous factor was specified with loadings set equal to 1 and its variance equal

to 0. Two latent classes were specified. Estimation problems were encountered requiring some parameters to be set equal across classes in order to reach convergence. The estimation difficulties most likely indicated that two or more classes were not viable. The tests comparing the fit of the two-class model to a single-class model were not significant, with the Lo–Mendell–Rubin adjusted likelihood ratio test (LMR) equal to 3.722, $p = .494$, and the Vuong–Lo–Mendell–Rubin likelihood ratio test (VLMR) equal to 3.794, $p = .488$, supporting the hypothesis that there was no additional benefit of a second class. The findings then suggest that there is unlikely to be substantial sources of unobserved heterogeneity.

Cox Regression

Cox regression, sometimes referred to as the semiparametric, partial likelihood, or just the proportional hazards model, is a survival analysis method used when the time of the event is measured on a continuous scale. Since Cox proposed the method in 1972, it has become the most widely used approach to survival analysis. One of the primary reasons for its popularity is that it does not require any assumption about a particular distribution, as do the so called "parametric" methods, such as the Weibull, Gompertz, or exponential failure models. Cox regression is not widely available in SEM programs, so I have placed more emphasis in this chapter on the discrete-time survival model, which can be implemented in more structural modeling programs. Nonetheless, there are some important statistical and application advantages of the Cox method that warrant at least a brief introduction to its application with structural models.

Continuous vs. Discrete

The objectives of Cox regression are the same as those for discrete time survival analysis, but the Cox approach is a more general model in that discrete times can also be used with Cox regression. The approach still focuses on the question of whether and when an event occurs, but has greater flexibility in handling timing of the event. Discrete time survival analysis is often not practical when there are many discrete time points, data are truly continuous, and a more accurate date exists for the event. All that is needed for Cox regression analysis is an indicator of whether or not the event occurred by the end of the observation period and a variable measuring the number of units of time until the event occurs, if ever. Thus, dummy variables for discrete intervals are not required for the analysis and any case may have a real number value for the time variable.

Researchers usually choose between the discrete time survival or the Cox regression method based on the exactness of the time variable and the number of time units available. When the number of time points is very large and varies considerably across cases, such as the number of days from when a driver receives a driver's license until he or she is in a car accident, discrete survival analysis can become impractical or impossible without grouping event times. Unequal time intervals are more difficult to manage with the discrete survival method but are of no difficulty with the Cox regression method. When data are sparse and there are few events for each time point, the logistic method may not produce reliable estimates and standard errors (Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). Analysis of finer-grained increments of time will always have greater precision than grouping continuous time into fewer cruder intervals (Thompson, 1977). As one example, had the date that the doctor diagnosed diabetes existed in the health and aging data set, use of the exact date of diagnosis (although not necessarily the onset of diabetes) would provide more statistical precision than using the wave of the interview. As long as the rank

order of the events does not change, there would be no effect of using a coarser metric of time, but grouping necessarily entails loss of some information about the relative order of the time that the event occurs. Where the available measure of time to event has few units, it may make little practical difference in the results whether the discrete or the Cox method is used. The basis of the Cox model is an underlying continuous measure of time, however, making it a more general framework for survival analysis.

The Cox method can certainly be applied to data where there are few time units (Allison, 1984), but the specifics of the estimation approach can matter. When there are fewer values for the time variable, there will tend to be more *ties* – instances in which multiple cases have the same time value. The precision of estimates of the Cox model can vary depending on how these ties are resolved in the estimation process, because partial likelihood estimates are based on the rank order of the units of time. One method of dealing with ties is the discrete method (Cox & Oakes, 1984), and, when this method is used, the Cox model gives the same results as discrete-time survival analysis with logistic regression. In this sense, the discrete method can be viewed as a special case of the more general Cox method. The most accurate estimation option for handling ties appears to be the *exact method* (Kalbfleisch & Prentice, 1980).[4]

Basic Concepts

Before discussing the implementation of the Cox method in structural equation models, let us first review some basic concepts behind the approach. The concept of the hazard and survival for the Cox model follow that of the discrete model, except that time is a continuous value and the event is conceived of as occurring within an interval in that time. Instead of Equation (12.1), which defines the probability of event occurrence at a particular time point, given that it has not occurred before, the continuous concept conceptualizes hazard as the rate or instantaneous risk that the event will occur within an interval. Thus, it is better to consider the hazard as a function of time, as designated by $h(t)$, in the continuous context, because we can more conveniently express it in terms of a probability distribution.

More formally, the hazard function asks, "given a certain interval width, how often does the event time $\mathcal{T}$ occur between two ends of an interval, t and Δt, if the event has not occurred before that interval?" For the following equations, the index i for individual cases is omitted to simplify the notation, but the probability for an individual case is implied. The hazard function for the continuous case, therefore, is stated as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \mathcal{T} < t + \Delta t \mid \mathcal{T} \geq t)}{\Delta t}$$

The hazard function, $h(t)$, is a representation of the risk or rate of the events occurrence within a certain interval. Δt is defined in terms of an increment in continuous time that can be considered in ever decreasing quantities to a lower limit of 0, which the $\lim_{\Delta t \to 0}$ notation on the right of the equation refers to. As with the discrete hazard, the rate is conditional on the event having not previously occurred, that $\mathcal{T} \geq t$. The probability of the event occurrence within the interval as shown in the numerator on the right takes into account the width of the interval Δt as shown in the denominator. The hazard is not strictly a probability, because it can go over 1, depending on the units of measurement of the increment in t.

The survivor function – the cumulative probability that the event has not occurred up to a given point in time – can be stated in a parallel fashion to the probability in Equation (12.2).

$$S(t) = P(\mathcal{T} > t)$$

As before, the survivor function signifies survival, or the probability of still being at risk for the event, and is the complement to the event having occurred prior to time t. If we consider the unconditional probability that an event will occur at any particular time point as

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \mathcal{T} < t + \Delta t)}{\Delta t}$$

with $f(t)$ representing the unconditional probability density function (pdf) for the event, then the hazard function can be restated as a conditional density function in relation to the cumulative density function (cdf) of survival.

$$h(t) = \frac{f(t)}{S(t)}$$

This equation is a more statistical form of the hazard definition, stating it as the risk of the event, taking into account that the event has not occurred up until that point in time.

The probability formulation is the foundation of the continuous time survival approach, which can estimate the hazard in a linearized regression model with information about whether and when the event occurred. In the equation for the Cox regression model, the hazard is a function of the baseline hazard, $h_0(t)$ (recall that this is the hazard probability when x equals 0). An important difference from other regression models is that the intercept for the equation is the natural log of the baseline hazard, $\beta_0 = \ln[h_0(t)]$. With this modification, the Cox regression model looks like the generalized regression model with the log of the hazard as the link function, here given for a single covariate.

$$\ln[h(t)] = \beta_0 + \beta_1 x_1 \tag{12.6}$$

We then find the raw hazard by translating back (i.e., taking the antilog), which gives us

$$h(t) = h_0(t) e^{\beta x} \tag{12.7}$$

Much like the logistic regression for discrete survival, the Cox regression model implies a nonlinear relation between time and the hazard. Following from this equation, it is simple to obtain the hazard ratio for a predictor by raising the exponent to the power of the regression coefficient, e^{β}.

Continuous-Time Survival Structural Equation Models

With special estimation features, a continuous-time survival regression can be incorporated into a larger structural equation model. A variable representing the number of units of time until the event or censor serves as an outcome, but there are no other special parameter statements needed. Figure 12.7 illustrates two possible models. Figure 12.7a is equivalent to a survival regression model with several covariates. Figure 12.7b is one example of incorporating the survival analysis into a more complex model in which a latent variable serves as a mediator (Roth & MacKinnon, 2012). There are many more structural models imaginable that involve a continuous-time survival outcome, including multiple group, latent class, growth curve, and time series models. As with discrete-time survival models, continuous-time survival models should be supplemented with descriptive statistics, such as life tables for grouped continuous data as well as graphical displays of hazard and survivor functions.

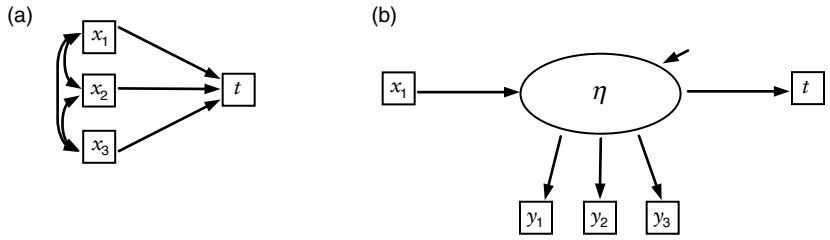

Figure 12.7 Structural Equation Model with Cox Regression.

Whereas estimation of discrete-time survival models only requires software capable of analyzing data with binary observed variables, special estimation features are required to estimate continuous-time survival analysis. For the continuous-time survival approach that I will focus on, an estimation approach to the partial likelihood for the log-hazard model from Equation (12.6) is needed. This can be incorporated into an ML-based approach using the EM algorithm (e.g., Asparouhov, 2014).[5]

The analysis requires a variable for the number of units of time that each case was observed, whether interview waves, months, or years, and a variable for whether or not the event has occurred during the study. The event variable is usually referred to as the *censor variable*. This is a binary variable representing whether the event has been recorded or censored because the event has not yet occurred when observation for that case ends.[6] The time variable takes on the value of the number of units of time (e.g., months) since the beginning of the period of observation (or other defined start point) that the event occurred or the number of units for which the case was observed. If the full period of observation for a pediatric dentistry study is 60 months and a patient has her first cavity at 31 months, the censor variable would be assigned a code of 0 indicating the event has occurred. The time variable would be assigned a value of 31. If a patient experiences no cavities over the full duration of study, the censor variable would be assigned code of 1 indicating the case was censored, and the time variable would be assigned a value of 60. If the patient dropped out of the study at 22 months without having experienced a cavity, the censor variable would be assigned a value of 1, indicating the case was censored, and the time variable would be assigned a value of 22, the last month of observation.

Example 12.3: Structural Equation Model with Cox Regression

Two simple continuous-time structural models were tested implementing the Cox regression approach. A new version of the social exchanges data set was used for these analyses. This data set contains information about deaths of the participants up to approximately nine years after the initial interview date. A censor variable was created to indicate whether the participant had died during the nine-year period (code of 0) or whether the participant was censored (code of 1). The time variable indicated the months until death or censor. The first model was a simple Cox regression (following Figure 12.7a) with unwanted advice, calculated as an average of the three unwanted advice questions used in earlier analyses. Results showed that unwanted advice was not significantly associated with the likelihood of death, $\beta = -.068$, SE = .067, ns. The hazard ratio, $e^{\beta} = .934$, suggested that an increase of one point on the unwanted advice scale was associated with a decrease in the hazard of approximately $(1/.934) \times 100 - 100 = 7.1\%$. This is the opposite of what would

normally be expected, but may suggest that, though advice was unappreciated, it had a small, nonsignificant beneficial impact on physical health (e.g., Lewis & Rook, 1999).

A second model was tested to illustrate both the use of Cox regression in a mediational model as well as incorporation of a latent variable in the model. The model resembles the model in Figure 12.7b with unwanted advice as a predictor of a latent variable for depression (four indicators), which in turn predicted the months until death. Results indicated that unwanted advice was associated with greater depression, $\beta = .147$, $\beta^* = .200$, $p < .001$, but that the depression factor was not significantly related to risk of death, $\beta = .056$, HR = 1.058.

Comments

There are several decisions to be made about how to specify a structural equation model for survival data. One major decision is whether to use a discrete or continuous approach. Discrete survival models are more obviously suited when there are few possible event time points, as is the case with panel studies in which the precise time of the event is unknown. Cox regression for continuous event times can be used in such circumstances and will even provide equivalent results depending on the method of handling ties. When more precise information about the event times is available and data are not time structured, there are increasing advantages to using a continuous-time survival method for reasons of statistical accuracy and convenience. Though time-varying covariates can be included in Cox regression models with survival analysis procedures used in conventional statistical software, there are no simple features for including time-varying covariates in the currently available structural modeling software. These considerations then suggest that the decision to use discrete or continuous survival analysis when there are few possible event times may hinge on the data structure, convenience, software availability, and the need for incorporating time-varying covariates. If the time to the event is more precisely measured and event data are not neatly time-structured, the choice of a continuous-time survival approach becomes increasingly obvious.

When covariates are included, careful attention should be paid to scaling of both time-varying and time-invariant covariates. The baseline hazard estimates the probability of the event when the covariate is equal to zero, a value that may not be desirable in the original scaling of the covariate. Consequently, centering covariates may be more reasonable as a default strategy than as a special consideration. Chapter 7 discusses scaling options for time-varying and time-invariant covariates in greater detail.

Finally, I have focused on Cox regression as the approach to continuous time survival in this chapter but parametric modeling is also possible, using Weibull, Gompertz, or exponential distributions. The Cox (or semiparametric) approach is more popular and makes fewer assumptions, whereas parametric approaches, can provide more precise statistical estimates when the distribution of the hazard is known or theorized. Results from the semiparametric and parametric approaches will often lead to the same conclusions, though the choice of the parametric model runs the risk of inaccuracy if the distribution specified is incorrect. In many social science applications, there is insufficient information or theoretical detail for assuming one hazard distribution over another, making it difficult to implement a parametric approach with any confidence.

To date, there are few existing published examples in which survival analysis is included in a structural equation model, because these approaches have only recently been developed. There are many hypotheses concerning the prediction of discrete events that should take into account right censoring, however, and the development of SEM approaches to

survival analysis makes possible many fascinating extensions that involve latent variables and longitudinal modeling techniques.

Extensions

This chapter is a general introduction to two popular survival analysis methods for SEM, the discrete-time survival analysis and the Cox regression method, but there are many possible extensions to these basic models. Muthén and Masyn (2005) discuss the incorporation of elements of growth curve models as well as several variations of latent class analysis for modeling unobserved heterogeneity with discrete-time survival SEM. Voelkle and Sander (2008) also illustrate the modeling of unobserved heterogeneity for discrete-time SEM using latent classes. Masyn (2003, 2009) extends the discrete-time survival approach to repeated events (e.g., repeat criminal offenses), and Malone and colleagues illustrate dual-process discrete-survival models that investigate the association of two event variables (Malone, Lamis, Masyn, & Northrup, 2010). Asparouhov and colleagues (2006) discuss extensions of continuous survival models for mixture and multilevel applications. It is also possible to estimate continuous-time survival with parametric models (e.g., Weibull) with several variations (Asparouhov, 2014). Moerbeek (2014) discusses sample size and effect size computations in the context of latent class models and long-term survivor models.

Recommended Readings

The material in this chapter is an elementary introduction to the most essential concepts of survival analysis, but many more details can be found elsewhere. There are several excellent, accessible book-length introductions to discrete and continuous regression survival analysis (Allison, 2010; Singer & Willett, 2003). For the more mathematically inclined, Cox's 1972 original paper is an extraordinarily concise proposition of the discrete and continuous regression concepts, and Kalbfleisch and Prentice (2002) is a comprehensive treatise of semiparametric and parametric survival models. Masyn's dissertation (2003) lays out many of the basic issues involved in the SEM approach to discrete-time survival analysis and the article by Muthén and Masyn (2005) more briefly introduces the latent class mixture modeling approach. Larsen (2004) gives an example of application of latent class analysis with continuous-time survival SEM and an example of the use of continuous latent variables with continuous-time survival analysis (Larsen, 2005).

Notes

1 I use notation that differs from that of other sources to be consistent with the rest of the text and to avoid confusion with notation used throughout. T_i is usually used to designate the event time, and j is usually used to designate a particular time point of the study (e.g., Allison, 1984; Singer & Willett, 1993).

2 Strictly speaking, for the logistic regression approach, e^{β} is the odds ratio. If the complementary log-log link is used, e^{β} is the *hazard ratio*, representing the increase in the hazard rate (e.g., Singer & Willett, 2003). The hazard value is the same as the relative risk.

3 For the logistic regression approach, a way of viewing the assumption is to say that there is not an interaction between time of the event and the covariate (e.g., Graham, Willett, & Singer, 2012).

4 The Efron (1977) and Breslow (1974) methods are other options that provide more statistical accuracy than the discrete method and are less computationally intensive than the exact method. The Efron approach has clear advantages over the Breslow approach, even though the Breslow approach is commonly implemented as a default in software programs. Where the exact method

option of handling ties is available, it is statistically preferable over the other methods (see Allison, 2010, for more details).

5 Currently, only Mplus software (Muthén & Muthén, 1998–2012) incorporates estimation features for continuous-time survival analysis into a general SEM program. Specification of parametric (Weibull) or semi-parametric (Cox regression) modeling of events is possible, but I limit discussion to the more popular Cox regression approach. For a Cox model, Mplus uses the equivalent of the Breslow method of ties (Asparouhov, Masyn, & Muthén, 2006).

6 Some software programs expect the censor variable to be coded as 0 for the event and 1 for censored, whereas other programs expect the censor variable to be coded as 0 for censored and 1 for the event. The documentation should be consulted and it is wise to perform a check (e.g., correlation analysis) to verify that the program treats the censor variable coding as expected.

References

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Newbury Park, NJ: Sage.

Allison, P. D. (2010). *Survival analysis using SAS: A actical guide* (2nd Ed.). Cary, NC: SAS institute.

Andress, H. J. (1989). Unobserved heterogeneity in rate models: Does it matter? A Monte-Carlo study of different rate models using generalized linear models. *Methodika*, *3*, 1–24.

Asparouhov, T. (2014). *Continuous-time survival analysis in Mplus: Technical appendix*. Retrieved from http://www.statmodel.com/download/Survival.pdf.

Asparouhov, T., Masyn, K., & Muthen, B. (2006, August). Continuous time survival in latent variable models. In *Proceedings of the Joint Statistical Meeting in Seattle* (pp. 180–187).

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89–99.

Collett, D. (2003). *Modelling survival data in medical research* (2nd ed.). Boca Raton, FL.: Chapman & Hall/CRC.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *34*, 187–220.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data* (Vol. 21). London: Chapman & Hall/CRC Press.

Farrington, C. P. (1996). Interval censored survival data: a generalized linear modelling approach. *Statistics in Medicine*, *15*, 283–292.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, *72*, 557–565.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, *83*, 414–425.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, *42*, 845–854.

Graham, S. E., Willett, J. B., & Singer, J. D. (2012). Using discrete-time survival analysis to study event occurrence. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 329–371). New York: Routledge.

Heckman, J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica: Journal of the Econometric Society*, *52*, 271–320.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*, New York: Wiley.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–481.

Kim, J. S. (2003). Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*, 489–502.

Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, *60*, 85–92.

Larsen, K. (2005). The Cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. *Biometrics*, *61*, 1049–1055.
Lewis, M. A., & Rook, K. S. (1999). Social control in personal relationships: Impact on health behaviors and psychological distress. *Health Psychology*, *18*, 63–71.
Malone, P. S., Lamis, D. A., Masyn, K. E., & Northrup, T. F. (2010). A dual-process discrete-time survival analysis model: Application to the gateway drug hypothesis. *Multivariate Behavioral Research*, *45*, 790–805.
Masyn, K. (2003). *Discrete-time survival mixture analysis for single and recurrent events using latent variables* (Unpublished doctoral dissertation). University of California, Los Angeles.
Masyn, K. E. (2009). Discrete-time survival factor mixture analysis for low-frequency recurrent event histories. *Research in Human Development*, *6*, 165–194.
Moerbeek, M. (2014). Sufficient sample sizes for discrete-time survival analysis mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 63–67.
Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics*, *30*, 27–58.
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*, 1373–1379.
Pickles, A., Neale, M., Simonoff, E., Rutter, M., Hewitt, J., Meyer, J., Crouchley, R., Silberg, J., & Eaves, L. (1994). A simple method for censored age-of-onset data subject to recall bias: mothers' reports of age of puberty in male twins. *Behavior Genetics*, *24*, 457–468.
Roth, D. L., & MacKinnon, D P. (2012). Mediation analysis with longitudinal data. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.), *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 181–216). New York: Routledge.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Singer, J. D., & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics*, *18*, 155–195.
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.
Sørensen, A. B. (1977). Estimating rates from retrospective questions. In D. Heise (Ed.), *Sociological Methodology*, 1977, (pp. 209–223) San Francisco: Jossey-Bass.
Tuma, N. B., & Hannan, M. T. (1978). Approaches to the censoring problem in analysis of event histories. In K. F. Schuessler (Ed.), *Sociological Methodology, 1979* (pp. 209–240). San Francisco, CA: Jossey-Bass.
Thompson Jr, W. A. (1977). On the treatment of grouped observations in life studies. *Biometics*, *33*, 463–470.
Vermunt, J. K. (1997). *Log-linear models for event histories*. Thousand Oakes, CA: Sage Publications.
Voelkle, M. C., & Sander, N. (2008). University dropout. *Journal of Individual Differences*, *29*, 134–147.

Keywords
survival analysis, Cox regression

13 Missing Data and Attrition

Missing Data Concepts

Missing data are nearly inevitable in applied research, and, with longitudinal studies, it is even more common because of attrition. The goal of this chapter is to review methods of estimating longitudinal models in the presence of missing data, with particular focus on dealing with problems associated with attrition. There are many introductions to missing data theory and modern methods of missing data analysis (Allison, 2002; Enders, 2010; Little & Rubin, 2002), and, except for a brief review of concepts, I mostly leave the explanation of these approaches and their benefits to other sources. In contrast, there are few sources that deal explicitly with missing data due to attrition and fewer still that deal with attrition in the context of longitudinal structural equation models. I begin with a review of missing data concepts and a discussion of attrition before turning to analysis approaches under several missing data conditions.

Before continuing, I would like to comment on the placement of the discussion of missing data within the book. If missing data are an integral problem to applied research, why wait until the end of the text to discuss it? There are several reasons why I have chosen not to discuss missing data to this point. One reason is that considering missing data issues add complexity to already complex topics, and addressing missing data throughout would easily distract from the fundamental concepts most essential for understanding a given modeling approach. Another reason for a delayed discussion of missing data is that more will be gained by the reader if the potential biases can be considered in the context of the objectives of a particular model, something I try to emphasize in this chapter. Relatedly, several of the missing data-related ideas entail an understanding of some of modeling concepts covered previously in the book, such as mixture modeling. Lastly, missing data is fairly easily handled through full information maximum likelihood (FIML) estimation in all major SEM software programs, where it is sometimes even the default analysis setting. This means that handling missing data under ideal conditions is a very trivial matter as far as implementation goes. Ultimately, as should be clear by the end of this chapter, the presence of missing data or attrition should not dramatically change the analyst's general modeling strategy for longitudinal structural models, barring that a few additional, simple modeling techniques are considered, some supplemental attrition analyses are conducted, and results are interpreted in light of the potential biases that may be due to attrition.

Missing Data Concepts

Missing Completely at Random

A general conceptual framework for understand missing data conditions has been widely employed to understand statistical biases that may arise in the presence of missing data

(Rubin, 1976). One way to understand this framework is in terms of a sampling problem. If a sample that is taken from a population is randomly selected, we do not expect any biases in the estimate of the population mean based on the sample average. If the sample is considered to be merely a subset of cases from a larger group of potentially available cases, the cases not selected can be considered to be missing. When the sample is random, the likelihood of inclusion in the sample, or likelihood of missing data, cannot be related to any other variables by definition. This is the concept of *missing completely at random* (MCAR) – whether a variable is missing or not is unrelated to *any* variables in the data set. Under this condition, we expect no biases in the mean estimate or the variance estimates of the variable that has missing values. Because, in practice, missingness is probably rarely completely random, there are potential biases that may arise resulting from greater likelihood of missingness for some types of cases. A randomly selected module in a survey, in which a subset of respondents are interviewed, is one exception for which the researcher can be reasonably certain that a variable will be MCAR.

If a variable of interest is not missing completely random, biases may exist in inferences from the observed data. Estimating the mean of one variable with missing data will be biased to the extent that individuals with higher or lower values are more likely to be included. Consider the classic example in which survey respondents with higher income are less likely to report their income. This phenomenon results in individuals included in the sample that have a lower average income than would be the case had everyone responded. If there is no income difference between those who did and did not respond to the income variable or any other variable in the data set, we would say that values are MCAR.

To formalize the definition a little bit, call the income question y^{mis}, which has some missing values, and another variable in the data set, such as age, y^{obs}, which has observed values for all respondents. An indicator variable m will be equal to 1 if the respondent is missing values on the income variable and will be equal to 0 if the respondent has a value on the income variable. Had respondents been randomly selected to answer the income question or not, the variable would meet the MCAR assumption. Otherwise stated, m would have an expected correlation with either y^{mis} or y^{obs} equal to 0. In terms of probabilities, we could say that the probability of missingness, m, given the values of y^{mis} or y^{obs} is simply equal to the overall probability of missingness, $P(m|y^{mis},y^{obs}) = P(m)$, because m is independent of y^{mis} and y^{obs}.

Missing data definitions are usually stated in terms of matrices in order to apply more generally to any number of variables with missing data $\mathbf{Y}^{mis}$ and any number of variables with complete or observed data $\mathbf{Y}^{obs}$. $\mathbf{M}$ is a matrix of binary missingness indicators with values equal to 1 if its correspondent variable in $\mathbf{Y}^{mis}$ is missing and values equal to 0 if not missing.[1] $\mathbf{Y}$ is a full matrix containing all $\mathbf{Y}^{mis}$ and all $\mathbf{Y}^{obs}$ variables. In matrix terms then, MCAR means that the probability of missingness $\mathbf{M}$ taking into account, or conditioned on, $\mathbf{Y}$, is simply equal to the probability of missingness.

$$P(\mathbf{M} \mid \mathbf{Y}) = P(\mathbf{M}) \tag{13.1}$$

Otherwise stated, the conditional distribution of $\mathbf{M}$ given $\mathbf{Y}$ is equal to the marginal distribution of $\mathbf{M}$.

Missing at Random

A second important category of missingness is *missing at random* (MAR). If a variable is MAR, missingness on that variable can be related to other variables in the data set, just

not to the values of the variable with missing data. For the income example, the variable would meet MAR if income was related to age, as long as income was not independently related to whether or not the respondent answered the income question. In terms of the notation, the missingness indicator, m, can be related to other variables y^{obs} but not the variable y^{mis} that has the missing values. The fundamental difference between MCAR and MAR is that MAR only requires that there not be a relationship between the probability of missingness and the complete values of the variable of interest, whereas MCAR requires no relationships of missingness to any variable. More formally, the MAR assumption is

$$P(\mathbf{M} \mid \mathbf{Y}) = P(\mathbf{M} \mid \mathbf{Y}^{obs}) \tag{13.2}$$

Because $\mathbf{Y}$ contains both $\mathbf{Y}^{obs}$ and $\mathbf{Y}^{mis}$, the formula implies that once $\mathbf{Y}^{mis}$ is taken into account $\mathbf{M}$ is only dependent on $\mathbf{Y}^{obs}$.[2] The converse of this is that missingness, $\mathbf{M}$, is not related to the values of the variable with the missing cases, $\mathbf{Y}^{mis}$. In fact, missingness may be related to the variable of interest overall, but as long as this relationship no longer exists once covariates are taken into account, the variable is still considered to be MAR. Thus, missingness on a variable may be very non-random for any other variables in the model as long as the non-randomness does not extend to the observed variable of interest. When a variable does not meet the MCAR or the MAR conditions, it is said to be *missing not at random* (MNAR or, in some sources, abbreviated as NMAR), to be nonignorable, or to have informative missingness.

Figure 13.1 presents a schematic adapted from Schafer and Graham (2002, p. 152) that summarizes the distinctions between the MCAR, MAR, and NMAR conditions, which are referred to more broadly as *missing data mechanisms*.[3] The variable z represents the true, but unknown, cause of missingness. The cause z is used here for conceptual convenience, but it is worth noting that the formal definitions given above are stated in terms of dependency between the variable of interest and missingness rather than the causes of missingness, as the actual causal process behind missingness does not need to be known. In the figure, y^{mis} is the variable of interest that has some missing data, y^{obs} is another variable with no missing data, m is the missingness indicator, and z is the true cause of the missingness. Double-headed arrows represent the association of each variable. The figure shows no connection between m and y^{obs} or y^{mis} for the MCAR condition, indicating independence of missingness from the variable of interest or other variables. There is no connection between m and y^{mis} for the MAR condition, because y^{obs}, which has a relationship to both m and y^{mis}, has been taken into account. There is a relationship between m and y^{mis}, however, in the MNAR condition even after y^{obs} has been taken into account.

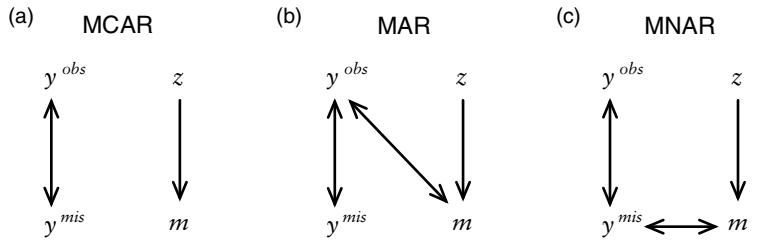

Figure 13.1(a–c) Analogue Representation of Missing Data Mechanisms. Adapted from Schafer and Graham (2002).

The term *ignorable* is used to refer to when values are MCAR or MAR. When missingness is ignorable, biases can be avoided by the use of appropriate missing data techniques. The term *nonignorable* is used when values are MNAR and biases cannot be completely avoided by use of appropriate missing data techniques. Ignorability and MNAR are usually used interchangeably, but ignorability includes a stipulation beyond MAR that the parameters that govern the missing data mechanism are equal (i.e., distribution of $\mathbf{Y}^{obs}$ and the conditional distribution of **M** given **Y**). If this stipulation is unmet and values are MAR yet not ignorable, then results will be unbiased though less efficient (Little & Rubin, 2002). For this reason, many authors treat MNAR and nonignorability as synonymous. In general, missing data mechanism distinctions are valuable for understanding how statistical analyses might be affected by missing data. The curse, however, is that they are all in the abstract, because we cannot know from the observed data whether missingness is related to the values of the variable of interest for any particular study.

Missing Data Analysis Approaches

These missing data distinctions are important, because the missing data mechanisms, have a bearing on whether certain strategies of handling missing data will result in biases or not. Strategies for handling missing data may include traditional methods, such as listwise deletion, pairwise deletion, or mean imputation, or modern methods, such as multiple imputation or FIML. Multiple imputation and FIML generally provide comparable results to one another (Graham, 2009). When values are MCAR, some traditional methods of handling missing data, such as listwise deletion, will give unbiased results, although modern methods will have greater efficiency (smaller standard errors). Under MAR, traditional methods will tend have biased estimates as well as less efficient estimates when compared with multiple imputation or FIML (e.g., Enders, 2001a; Enders & Bandalos, 2001).

FIML estimation for missing data is a commonly employed approach with structural equation models (see Enders, 2013 for a general overview of missing data estimation with SEM), so I will focus on this method.[4] A brief comment about FIML estimation is in order. The maximum likelihood (ML) estimation in the presence of missing data employed by most SEM software programs is usually referred to as "full information" maximum likelihood, because likelihood information is used from all cases, whether complete or not. The process is the result of summing the case-specific log likelihood values, so that data vectors do not need to be complete for each case. Thus, data from all cases in the sample are used in the estimation process and no cases are eliminated. Values are not imputed, but the parameter estimates are based on all of the available data for each case. The process is seamlessly implemented in SEM software either by default or with a simple optional specification. As standard FIML assumes multivariate normality, nonnormal, binary, and ordinal variables require modified versions of the estimation approach, but are also easily implemented where software features are available. Nonnormal and non-continuous data can be handled through robust adjustments or bootstrapping (Yuan & Bentler, 2000). Other sources provide a more thorough discussion of FIML estimation details (Arbuckle, 1996; Enders, 2001c, 2013).

Less work has been done on the performance of missing data estimation approaches for categorical variables. ML for binary or ordinal variables is a FIML approach, so should be expected to behave similarly to FIML for continuous variables under various missing data mechanisms. Robust estimates are also possible to improve standard errors under distributional assumption violations. Although more commonly used and widely available, WLSMV uses more of a hybrid approach to missing data, with some steps based on FIML and some based on pairwise deletion (Asparouhov & Muthén, 2010). All cases are used

in the analyses, but there are somewhat stricter assumptions regarding missing data when the outcome of interest or covariates are relied upon to meet MAR assumptions. WLSMV estimates are consistent and appear to be more efficient than listwise deletion based on initial simulation work. Because dependencies with covariates may not be fully accounted for with WLSMV, these authors recommend avoiding analyses with WLSMV when values are MNAR. Other techniques, such as Bayesian estimation and bootstrapping (e.g., Enders, 2002; Lee & Song, 2004), are possible solutions that may perform more like FIML under various missing data mechanisms (Asparouhov & Muthén 2010).

Patterns of Missing Data

There will nearly always be more than one variable with missing values in any data set. It can be informative to characterize the patterns of missing data across these variables (e.g., Little & Rubin, 2002), but the missing data patterns do not indicate which missing data mechanisms are relevant (MCAR, MAR, MNAR). Assumptions or reasons for missingness may differ across variables in the data set. Data may be missing for individual observed variables, often referred to as *univariate missingness* or item nonresponse. Univariate missingness can occur for many reasons, including poorly worded items or options, sensitive questions, or participant burden. For individual item nonresponse in surveys, the pattern of missingness often idiosyncratic and usually does not involve a large percentage of cases. In such instances, data may be MCAR or MAR or not large enough to seriously bias results. Univariate missingness can add up to a large percentage of the sample quickly when there are many variables involved, however. Another common pattern of missing data is missingness for an entire case at a particular observation. This pattern could be called *unit missingness* or multivariate missingness, where more than one variable is missing for a case in the same pattern. This may involve a section of a survey or the entire survey. Nonresponse to an entire wave is perhaps the most likely form of unit missingness in longitudinal research, and a special form of such unit missingness is attrition, in which values are missing at some wave after baseline and all subsequent waves. In the missing data literature, this is referred to as a *monotone missingness* pattern. If values are missing by design, such as use of survey models given to a subset of respondents or cohort sequential designs (Nesselroade & Baltes, 1979), variables are said to have *planned missingness*. This term might also be used to describe survey modules, cohort sequential designs, or longitudinal data that are not structured by scheduled waves of data collection. The latter circumstance would be exemplified by records-based data with observations for each case at occasions specific to each case (e.g., doctor's visits).

Attrition

The most common pattern of missing data in longitudinal studies will nearly always be an attrition pattern. Attrition is therefore a primary concern for researchers dealing with longitudinal data and requires careful attention when testing longitudinal structural equation models. There are few longitudinal studies that do not suffer from some level of attrition. There is no way to generalize about the cause of attrition in longitudinal studies, however. Participants no longer wish to continue, cannot be located, become too unhealthy, or die. There also is no standard rate of attrition that can be expected. In reviewing international longitudinal household surveys, Alderman and colleagues (Alderman, Behrman, Kohler, Maluccio, & Watkins, 2001) found that annual attrition rates varied between 1.5% and 28.2%, suggesting that attrition may vary considerably across studies. Attrition will vary depending on the number of waves, the

length of interval between waves, the population, participant burden, health, the mode of data collection, incentives, the study procedures, and retention efforts among other factors (Deeg, van Tilburg, Smit, & de Leeuw, 2002; Hunt & White, 1997; Sullivan, Rumptz, Campbell, Eby, & Davidson, 1996; Young, Powers, & Bell, 2006).

Attrition is a particularly problematic pattern of missing data for several reasons. Data are multivariate missing with values missing for an entire case at a certain wave and all subsequent waves. Over time the cumulative percentage of missingness on at least one wave can become fairly large. Dropouts are unlikely to occur for arbitrary reasons, and this means that missing values associated with attrition usually will not meet the MCAR assumption and may be more likely to violate the MAR assumption. The MAR assumption will not be met to the extent that dropout is associated with concurrent or future values of the variable with missing data (Kenward, Molenberghs, & Thijs, 2003). Studies about health, sociological factors, economic status, or psychological characteristics all will be vulnerable to dependency between their key variables of interest and probability of dropping from the study. For example, in health surveys, physical and mental health are common causes of dropout (Mirowsky & Reynolds, 2000), and, in education studies, poorer performing students are often more likely to drop out of college (e.g., Cabrera, Nora, & Castaneda, 1993).

Despite what may seem like several serious challenges that attrition poses, there are some advantages to attrition compared with other missing data patterns. Although there may not be a sole reason for attrition, there may be fewer reasons for missingness with attrition than there are for other patterns of missing data. With attrition, there is always more information available about potential variables related to missingness, because there is information about all cases at baseline. Many studies also include a brief follow-up tracking survey that provides some information about why participants drop out. The baseline data and any tracking data are potentially very valuable for mitigating biases due to attrition when they are included in the analyses.

Because we cannot know for sure whether missing values due to attrition meet the MAR assumption or not, there is no clear analysis strategy for any given longitudinal study. When sufficient information is available to account for the probability of attrition, however, FIML analyses can be used to obtain unbiased and efficient estimates. If a variable is MNAR because of attrition, then analysis strategies are more complicated and the prospects of mitigating bias are more tenuous. If there is reason to suspect the MAR assumption might be violated, there is less agreement about the best way to approach missing data analyses, but there is some consensus that analyses assuming MAR should be supplemented with sensitivity analyses designed to address values that are MNAR. I will discuss analysis strategies for handling missing data when missingness is due to attrition under both MAR and MNAR conditions. But first, let us consider a few basic approaches to exploring missing data patterns.

Exploring Missing Data and Missing Data Assumptions

Exploratory analyses should always be conducted to understand as much as possible about the patterns of missing data and how the observed values may differ from the missing values. Tests have been proposed to investigate whether a variable meets the MCAR criteria (e.g., Dixon, 1988; Kim & Bentler, 2002; Little, 1988; Muthén, Kaplan, & Hollis, 1987; see Enders, 2010, for an overview), but no such tests exist for exploring whether a variable meets the MAR criteria. Data that are MCAR are probably unlikely in many applied studies, unless by design, and attrition will often be related to some variables of interest. Tests of the MCAR assumption may not be of particular importance

themselves, because there is no harm and generally only advantages to applying modern missing data analysis methods, which do not require the stricter assumption of MCAR anyway (Enders, 2010). Without special circumstances in which population values are known (e.g., simulation studies, rare instances with real data), there is no way to determine whether a variable meets the MAR assumption that missingness on a variable is unrelated to the value of that variable (Schafer & Graham, 2002). Understanding which variables may be related to missingness, however, may be informative about potential biases worth noting and may be useful for identifying auxiliary variables (discussed later) that should be included in the model.

Univariate and Bivariate Approaches. One type of descriptive information valuable for characterizing missingness is *observed data coverage*. Coverage is the proportion of cases present for each variable in the data set and the proportion of cases present when considered pairwise. That is, what proportion out of the total sample size has complete data for any two variables considered together? When coverage is low on many variables, FIML estimation may have greater difficulty converging (Enders & Bandalos, 2001). Coverage statistics also provide univariate missing proportions for each variable and pairwise coverage proportions say something about the pattern of co-occurrence of missingness. For example, it is valuable to know if there are refusals to a set of questions on a particular topic, such as income-related questions. Such patterns of co-occurrence may suggest reasons for missingness and thus provide information about whether values of one variable are related to the likelihood that the variable is missing.

Several mean and variance comparison tests can be conducted to investigate whether variables in the data set differ for cases that are missing or not missing on another variable (Goodman & Blum, 1996). As a simple example, the values of y_1 can be compared for cases that are missing and not missing at Time 2. A binary variable, m, is constructed that indicates missing or not missing and serves as a grouping variable and can be used to conduct t-tests, for instance. Correlations between y_1 and m would serve the same purpose. The end of the study or other points of dropout can be used as the basis of grouping as well. Traditional t-tests or correlations are generally sufficient for these analyses, but the same goals can be accomplished with multigroup structural modeling and can be extended to include tests of equality of variances and covariances across missing and not missing groups (Kim & Bentler, 2002; Muthén et al., 1987). An informative strategy is to examine the correlations among key variables for those who drop out and those who do not drop out of the study. The indicator of missingness can be expanded to represent a reasonable number of patterns of missing data across multiple variables, as long as there is sufficient sample size in each group. A multivariate test by Little (1988) was developed for such a purpose, and multigroup SEM approaches can be readily applied in the multivariate case (Kim & Bentler, 2002). Although tests of the equality of correlations can be used (e.g, Cohen, Cohen, West, & Aiken, 2003), this is perhaps more conveniently handled through equal covariance tests in multigroup SEM via likelihood ratio tests with constraints on correlations in the attrition and non-attrition groups (Raykov, Lichtenberg, & Paulson, 2012). For any of these approaches, significant differences would suggest violation of the MCAR assumption. On the other hand, nonsignificant tests would be consistent with MCAR but not prove that it holds.

Exploring Correlates of Attrition. Although the MCAR comparison tests are often discussed with cross-sectional data in mind (cf. Muthén et al., 1987), they may be particularly useful for exploring differences in baseline or other complete data among individuals who drop out of a longitudinal study. And for longitudinal data, we nearly always have

complete data on the variable of interest at baseline. Differences in baseline values of health, for example, can be compared for those who drop out of the study and those who complete the study. More specifically, if m_2 is an indicator of missingness for y_2, we can test for differences on y_1 for missing ($m_2 = 0$) and nonmissing ($m_2 = 1$), providing a method for exploring its relation to missingness (Hedeker & Gibbons, 2006). The attrition indicator approach can also be modified to move beyond simply comparing dropout and non-dropout cases. Baseline values could be compared among attrition, intermittent missing, and complete groups, for instance. These ideas can be further extended to examining the probability of dropping out in relationship to changes in a variable of interest over time (Hedeker & Gibbons, 2006).

Remember that such analyses are not a direct test of the MCAR assumption, however, because it is the unobserved values of the variable at some later time point that are assumed to be independent of the likelihood of missingness at that time point. Values on the baseline measurement, y_1, are merely proxies for values on the follow-up measurement, y_2. If health is the cause of attrition, it is not health at baseline per se that causes the participant to dropout at follow-up. In theory, if the follow-up measurement was collected immediately after baseline, there could be a test of the relationship between the variable with complete values and missingness. The interval between waves makes the baseline value an imperfect representation of the values of the variable that is partially missing (Demirtas & Schafer, 2003; Feng, Cong, & Silverstein, 2012). Nonetheless, the earlier value of health may be more highly related to missingness at follow-up and the reason for missingness than any other variable, assuming the status of health at follow-up is the cause of attrition. Graham (2009) is less sanguine about the benefits of this type of analysis, stating "Knowledge will be gained from this practice, but not as much as researchers might think" (p. 569). Still, explorations of missing data in the longitudinal context have a distinct advantage over explorations in the cross-sectional context for understanding missing data mechanisms. I return to this point later when discussing MAR analysis strategies and auxiliary variables.

Attrition is not a missing data mechanism, but it is a particular missing data pattern that by tradition we suspect will bias results due to a type of self-selection. The type of biases that may exist because attrition violates missing data assumptions should be considered in the context of the type of longitudinal model. Bias in the mean estimates that stem from attrition will be particularly relevant to analyses that are connected to mean change or individual level change, such as ANOVA models, growth curve models, or latent difference score models. Results from these models also will be vulnerable to biases if the probability of missingness changes over time and is related to changes in the variable of interest over time (Hedeker & Gibbons, 2006). If health declines over time and the probability of missingness increases over time, the pattern of change in the variable is confounded with the pattern of missingness in the variable. Such a relationship may or may not be evident from bivariate explorations of the relationship between a variables earlier values and the likelihood of missingness. Other analyses, such as simplex models and cross-lagged panel models, will be more subject to bias due to attrition where missingness affects variances of individual variables and covariances among variables. Results from these models will be vulnerable to bias if changes in variances or covariances occur over time and these changes are related to changes in the probability of missingness.

Example 13.1: Missing Data Patterns in the Health and Aging Data Set

To illustrate one example of missing data for a longitudinal study, Table 13.1 presents BMI data taken from a second version of the health and aging data set (data sets are described at the beginning of the book Example Data Sets section at the beginning of the book).

Table 13.1 BMI Scores for a Sample of 20 Respondents in the Health and Aging Data Set

	BMI					
i	*Time 1*	*Time 2*	*Time 3*	*Time 4*	*Time 5*	*Time 6*
1	28.4058	29.8190	28.8468	28.8818	27.7172	.
2	24.6955	25.8119	23.9076	26.6079	26.7622	.
3	25.3338	25.7753	25.5106	25.8284	24.2496	25.3991
4	25.1134	29.6900	28.7574	28.3042	32.1537	29.1830
5	26.3591	27.3821	.	26.6782	.	.
6	30.0993	34.6796	36.7037	36.5473	33.6946	34.2607
7	32.6196	30.0074	30.3184	32.7086	.	.
8	14.7502	15.3656	18.5409	20.2100	.	.
9	17.9452	17.2409	16.4012	.	.	18.7759
10	26.7671	26.5133	29.9508	28.4260	28.2994	25.3191
11	14.9028	22.0361	21.3248	.	21.7940	.
12	31.8711	32.1656	30.9940	35.0346	29.1991	32.2162
13	14.9358	20.2679	19.2545	19.8230	20.7863	18.5615
14	21.8173	22.4075	21.6594	19.7735	21.5537	22.9930
15	26.9279	30.0469	30.1750	.	28.8625	.
16	32.8624	32.3462	34.1222	33.6520	34.7506	33.3949
17	28.5383	.	30.5656	30.4143	32.8563	29.4174
18	28.1739	24.6000	25.2089	25.8076	27.7432	22.1453
19	26.2152	27.7619	28.3247	27.7411	27.8343	26.1363
20	31.5193	.	.	.	.	.

Table 13.2 Missing Data Patterns for BMI in the Health and Aging Data Set

	f	%
Non-dropouts	5,919	68.79
Complete	4,217	49.01
Intermittent	1,702	19.78
Dropouts	2,686	31.21
Complete	1,968	22.87
Intermittent	718	8.34
Total observations	8,605	100.00

The first version of the health and aging data set included only cases with complete data. The data in Table 13.1 are from a second version generated to include missing data in a pattern closely resembling the data from the original study. Syntax and data sets used in the examples are available at the website for the book. The table contains a sample of 20 cases for BMI scores at each of the six waves. A period "." represents missing data at that time point. A variety of patterns of missing data are exemplified. Some cases have complete data ($i = 14$), some drop out after the first wave ($i = 20$), some have with intermittent missing data ($i = 17$), and some are missing the variable only for at the last wave ($i = 1$).

Table 13.2 summarizes several important patterns for the full data set. I define the term dropout in this example as those who illustrate an attrition or monotone missing pattern, defined as any cases with missing data at all remaining time points subsequent to having present data at any time point. Notice that under this definition those who are only missing data at the last time point are considered to have dropped out, even though there is no way to know whether an individual would have continued had there been additional

waves of data collection. Dropouts can be further categorized into those with complete data up until they dropout (e.g., $i = 7$ in Table 13.1) and those who have some intermittent data prior to dropping out (e.g., $i = 5$ in Table 13.1). Out of 8,605 cases, 2,686 (31.21%) dropped out of the study some time during the 12 years after the first wave. Of the remaining cases (5,919 or 68.79%), 4,217 (49.01%) had no missing values, and 1,702 (19.78%) had some (intermittent) missing values. Remaining cases had complete data for all waves of the study.[5] As these findings suggest, the most common pattern of missing data in this study was attrition. Of the 4,388 cases with some missing data on this variable, 2,686 (61.21%) had missing data because they dropped out.

Many more analyses will likely be needed to explore the pattern and correlates of attrition in most applications, but I illustrate a few here to give an idea of the types of analyses that will likely be useful for better understanding patterns and correlates of attrition. To start, I used a likelihood ratio test to compare mean BMI scores at Time 1 for those respondents missing data at Time 2 ($N_1 = 1{,}161$) to those not missing data at Time 1 ($N_2 = 7{,}444$). This analysis does not require FIML, because complete data are available at baseline and for the missingness indicator. Note that this analysis is the same as a *t*-test allowing for heterogeneous variances. A model allowing the means in the two groups to differ is just identified, so no fit information is available. The baseline mean on BMI was 27.160 for the missing group and 27.273 in the nonmissing group.

A subsequent model constraining the baseline BMI means to be equal had $df = 1$ and chi-square equal to .481, which was nonsignificant. Because the first model was just identified, the fit of the second model provides information about whether the two means differed in the missing and nonmissing groups. Thus, results indicated that those who dropped out at Time 2 were not significantly different from those who had data at Time 2. Although these results are consistent with MCAR, it is not a direct test of MCAR because the two-year interval between Time 1 and Time 2 interviews make baseline BMI scores an imperfect proxy for values of BMI at Time 2.

Using a similar multiple-group comparison, I also compared the variances of the baseline BMI scores in the two missing data groups. The model allowing different variance estimates in the two groups is the same as the unconstrained means model tested above. The BMI variance was 26.643 in the missing group and 25.952 in the nonmissing group. Although the dropout group had a slightly higher variance than the non-dropout group, the fit of the constrained model, $\chi^2(1) = .350$, ns, indicated that the variances did not differ significantly. As with the mean comparison, the variance comparison is consistent with MCAR but is not a test of the assumption. With this data set, the sample size was large enough to detect significance for fairly small effect sizes, so a nonsignificant result is hard to discount for having insufficient statistical power. In other studies, it will be important to consider effect size estimates and post hoc power tests to gauge the magnitude of effect and whether a test had sufficient power to detect differences.

As an example of a test of whether covariates are related to missingness, I computed a simple correlation between self-rated health and missingness on BMI (i.e., dropout). Those with missing data on BMI later in the study were in poorer health, because baseline self-rated health was associated with missingness at Time 2, $r = -.065$, $p < .001$. It may also be informative to determine whether the association between BMI and self-rated health differs between those who dropped out of the study and those who were retained. If BMI scores at Time 2 were related to missingness at Time 2, the overall health of the respondent might at least partially account for this unknown relationship. To explore the potential differences in this relationship, a multiple-group strategy was used to compare the covariance between self-rated health and BMI at baseline for those who stayed in the study and those who later dropped out for the study. The model that compared the baseline

covariance between BMI and self-rated health between those with and without Time 2 data was just identified.[6] The covariance in the missing group was –.886, $p < .001$, and the covariance in the nonmissing group was –1.172, $p < .001$, both suggesting that those with higher self-rated health had lower BMI scores. The missing group had a slightly lower covariance. The model constraining the covariances to be equal across the two missing groups, was not significant, $\chi^2(1) = 2.089$, ns, however, indicating no difference between the missing and nonmissing groups. Note that a *z*-test for two independent correlations also could be used to explore this hypothesis. Had baseline BMI scores been related to missingness, a further analysis step might be to conduct a multiple logistic regression to examine whether the relation between baseline BMI and missingness remained when baseline self-rated health was included in the model (Goodman & Blum, 1996). These tests indicate that although self-rated health appears to be related to dropout, there is no evidence that baseline BMI is related to later probability of dropping out or that the relationship between self-rated health differed for those who stayed in the study versus those who dropped out of the study.

These analyses do not provide direct tests of the MCAR assumption but do provide valuable information about the patterns of missing data and the potential correlates of attrition. Correlations of baseline values with missingness at later time points can provide more or less valuable information about the MCAR assumption. Their worth in informing us about the MCAR assumption depends partially on the quality of the baseline value as a proxy (i.e., its stability). In this particular case, the Time 1 measurement of BMI was highly correlated with the Time 2 BMI, $r = .908$, $p < .001$. It therefore seems very unlikely that the relationship of Time 1 BMI to missingness at Time 2 would be much different from the relation of Time 2 BMI to missingness at Time 2 if we were to have access to data for all cases. Many variables will not be so stable, but, in this case, Time 1 BMI scores would seem to be an excellent proxy for Time 2 BMI scores and the analyses especially informative about the MCAR assumption.

Comments

Presentation of longitudinal modeling results should always be accompanied by some reporting on missing data patterns, causes of attrition, and likely biases. Graham (2009) makes several valuable suggestions for reporting on attrition, most notably that specific numbers should be provided and that likely reasons for attrition and the nature of the possible biases due to attrition should be discussed. Just as importantly, researchers will better understand their findings and their area of study if they identify the variables from their data set that may be related to attrition and consider how their results may differ because of attrition, even if the results cannot be discussed at length in published reports. Collection of data that can inform reasons for missingness is a natural solution to the usual guesswork involved in understanding attrition (e.g., Morrison et al., 1997). Reasons for refusal, brief tracking surveys, and obtaining data from external sources (e.g., death records, motor vehicle records) require extra effort during the data collection phase, but will provide invaluable insights into the status of missing data assumptions and limitations of the inferences from the analyses. Both theory and exploratory analyses should guide inferences about likely biases, a topic that should be addressed in any Discussion section. Biases due to attrition will rarely be entirely avoidable and should be accepted as a nearly universal limitation of longitudinal studies. Scientific knowledge will only be enhanced by attempting to characterize and understand such attrition biases.

When differences are significant they may not be of practical importance. Effects size indices to assess the size of mean or variance differences (Nye & Drasgow, 2011) can aid

the determination of whether significant differences are large enough to warrant concern. These determinations, of course, will need to be made in the context of particular research hypotheses and theory. When important differences are determined, the researcher then needs to consider the variables available that may help account for the relation of a variable's value to missingness so that they can be incorporated into the specific models of interest.

Analyses when Missingness is at Least MAR

When a variable meets the MCAR assumption, there are a number of options for handling missing data that will produce unbiased parameters, but modern missing data methods, such as FIML, will be more efficient than commonly employed conventional methods, such as listwise deletion. When a variable does not meet the MCAR assumption but does meet the MAR assumption, FIML will be unbiased and more efficient than listwise deletion. A number of studies have suggested that substantial advantages of modern methods over conventional methods do not accrue until as many as 15% to 35% of cases have missing values (Enders, 2001a; Kromrey & Hines, 1994; Newman, 2003). The variation among these studies is likely to be due to the parameter estimated (e.g., mean, variance, covariance) and the strength of the relationship between variables in the data set and missingness. Moreover, modern missing data techniques appear to do well even when there is a small total sample size (e.g., 50 cases), and the percentage of cases with missing values is as high as 50% (Graham & Schafer, 1999). It therefore makes sense as a general strategy to use FIML for all models with missing data unless there are difficulties with estimation or other extenuating circumstances that would make the approach a less reasonable alternative.

Auxiliary Variables

The MAR assumption is often easier to meet than is often assumed initially, because it states that there can be a relationship between missingness on a variable and its values, just not after other variables are taken into account. The assumption concerns the relationship between values of variable with missing data and missingness conditional on other variables (refer to Figure 13.1b). If the variable y^{obs}, which is often called an *auxiliary variable*, is related to y^{mis} and to m but the association between y^{mis} and m can be eliminated, then the MAR assumption is met (Meng, 1994; Rubin, 1996). Auxiliary variables do not have to be the cause of missingness. They may just be a variable related to the cause of missingness and to the variable with missing cases. As a confounding variable in the relationship between m and y^{mis}, inclusion of the auxiliary variable in the model has the effect of accounting for the relationship between the values of the variable with missing data and missingness. Consequently, including auxiliary variables in a model reduces parameter bias that might otherwise occur because of the missing data. The choice of auxiliary variables will be specific to the model tested. The researcher will want to include auxiliary variables that are not already in the model but are related to values of the variable with missing data and to missingness.

Much of the current thinking about the strategy of including auxiliary variables is based on simulation work by Collins and colleagues (Collins, Schafer, & Kam, 2001). This study generated a data set that included a variable which had values related to probability of missingness on the variable yet also included variables that were confounded with this relationship. Their results show that substantial benefits of including auxiliary variables in the model could be gained when the auxiliary variable is both strongly related to the

variable with missing data and to missingness (a correlation of .4 or larger) and when the percentage of missing cases is high (50% or higher). Under these conditions, the authors found that there was negligible bias in mean, variance, and covariance estimates if the auxiliary variable was included. Omitting the auxiliary variable resulted in biases in these parameters, with bias increasing as the relation of the variable with missing data and missingness increased. Estimates also had efficient standard error estimates, with nominal confidence interval coverage if the auxiliary variable was included in the model. When the variable was already MAR, including auxiliary variables only related to values on the variable with missing data but not to missingness improved standard errors and statistical tests. Thus, it appears that there will generally be some benefit to including auxiliary variables even when they are not strongly related to missingness. When 25% or fewer cases were missing, the benefits of including auxiliary variables was not substantial, so, although there may be no harm in including auxiliary variables when the number of missing cases is minimal, it may not be critical to include them when the number of missing cases is not large. This finding is consistent with other simulation studies suggesting that, under MAR, differences in biases in parameter estimates across missing data methods are minimal (although with more important differences in efficiency) if the proportion of missing cases is small (e.g., Arbuckle, 1996; Enders & Bandalos, 2001).

Auxiliary variables can be used in structural equation models in three ways: (a) through a two-stage process using another software package to obtain an FIML estimated covariance matrix for input into an SEM program; (b) as a correlate of all measured variables in the model; or (c) or as an extra dependent variable. In the *two-stage approach* (Savalei & Bentler, 2009; Yuan & Bentler, 2000), a full information variance–covariance matrix is estimated for a set of auxiliary and model variables that contain missing data, typically employing an expectation maximization (EM) algorithm. The two-stage process is less convenient compared with the other two approaches that integrate the auxiliary variables directly into the model in a single stage. Figure 13.2a illustrates the second method, the *saturated correlates approach,* and Figure 13.2b illustrates the third method, the *extra dependent variable approach*. Graham (2003) shows that the three methods provide comparable results in terms of parameter estimates and standard errors. The two-stage approach is rarely used in practice, so I will not consider it further (but see Savalei & Bentler, 2009 for discussion of its advantages). The extra dependent variable approach has some drawbacks in terms of evaluating model fit, because the chi-square, degrees of freedom, and RMSEA differ from the other methods.

There are some special considerations for fit indices when auxiliary variables are included. For the saturated correlates approach, the chi-square values and RMSEA indices do not require any adjustments. The model adds no extra restrictions, so the degrees of freedom and chi-square are not affected. The relative fit indices, such as the Tucker–Lewis index and the comparative fit index, should use an independence model that includes the correlates in order to reflect only the substantive portions of the model. Caution is required in using the SRMR, which will be downwardly biased because the correlates will oversaturate the model. Enders (2013) recommends testing the model without the auxiliary variables to obtain the SRMR value if desired. Some software programs have provided special automated features for the saturated correlates approach, including the special computation of fit indices, and this provides an additional advantage in convenience for this approach.

For longitudinal studies, baseline or prior measurements are obvious candidates for auxiliary variables, because they will tend to have substantial association with the later measurements of the same variable (Graham, 2009). They may also be strongly related to

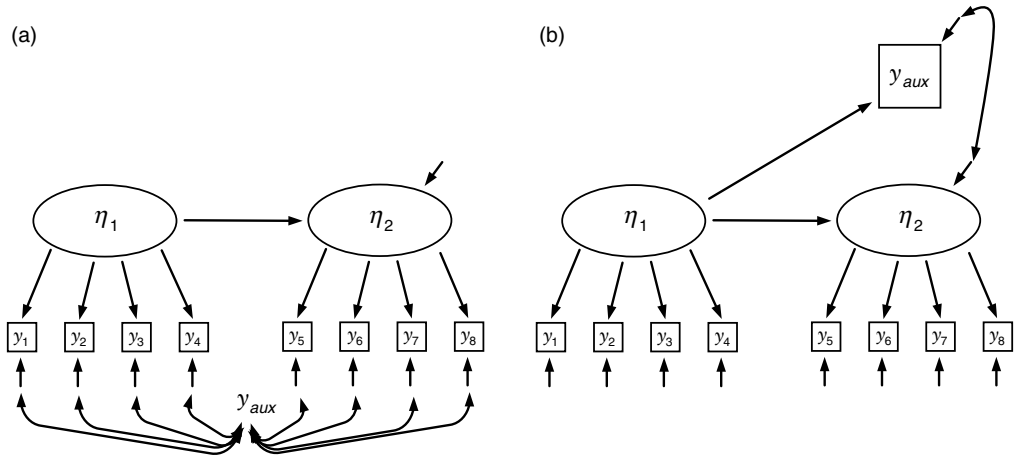

Figure 13.2 Inclusion of Auxiliary Variables (a) Saturated Correlates and (b) Extra Dependent Variable.

probability of missingness and reasons for attrition in many settings. For example, longitudinal studies in which health is a key variable of interest also frequently have health as an important reason for attrition or intermittent missing data. If earlier measurements are already included in the model, however, as is the case with many longitudinal models, the more complete values of such auxiliary variables are already taken into account in the model and cannot be employed as auxiliary variables.[7] And if baseline measures or measurements of the variable at a prior time are not included in the model for some reason, they should be considered as auxiliary variables. For most models with earlier time points included in the model, an informative exploratory step may be to test a modified model including later measurements of the outcome and other central variables in the model with and without prior measures of the outcome employed as auxiliary variables in order to investigate how means, variances, and covariances might be changed by their inclusion.

Example 13.2: Three Longitudinal Models Using Auxiliary Variables

Three models were tested to illustrate the inclusion of auxiliary variables with FIML using the saturated correlates approach. The use of auxiliary variables might increase the plausibility of the MAR assumption if missingness is related to values of the variables of interest and auxiliary variables are strongly related to these values. The auxiliary analysis approach is expected to reduce biases due to attrition and other missing data patterns as well as increase precision of standard errors compared to a listwise deletion approach. The first model tested followed the repeated measures ANOVA model in Example 3.5 using the contrast coding approach to test for mean changes in depression. Subscale scores for depression served as indicators of a general depression factor in the health and aging data set. Two time points were used to investigate whether the factor mean for depression changed significantly over the two-year period. Factor variances and means were identified with the effects coding approach (Little, Slegers, & Card, 2006) set with complex constraints. Longitudinal invariance of loadings and intercepts was assumed by constraining these parameters to be equal. Auxiliary variables were included using the saturated correlates approach (Figure 13.2a).

Two variables, baseline self-rated health and age, were used as auxiliary variables because of their possible association with depression and missingness. The motive for including these variables in this example was primarily didactic, and, in practice, the choice of auxiliary variables would be based on prior analyses or theory. The full data set had 8,605 cases available for the analysis, including variables with complete and missing values. The fit of the model did not reach generally accepted standards of fit for all indices, $\chi^2(12) = 1{,}082.275$, $p < .001$, CFI = .925, SRMR = .049. (The less than optimal fit of the model appears to be due to the measurement invariance constraints, but I will ignore this issue for these examples.) The mean of the difference factor was significant, $\alpha_1 = -.016$, SE = .006, $p < .05$, and indicated that depression was slightly but significantly lower two years after baseline. Given the intercept estimate, $\alpha_0 = .332$, SE = .005, which represents the mean for the depression at baseline, it is a simple matter to compute the factor mean at Time 2, with .332 + (–.016) = .316.

As a comparison, a model was tested using FIML without auxiliary variables. This model had a nearly identical chi-square, 1,082.544, with the same degrees of freedom and alternative fit index values. The parameter estimates and their standard errors were also unchanged, $a_1 = -.016$, SE = .006, and $\alpha_0 = .332$, SE = .005. The inclusion of the auxiliary variables in this case made no appreciable difference, either because of the inclusion of the baseline value of depression mitigated any effects of the auxiliary variables or because substantial differences in efficiency could not be discerned with the large sample size. A model using listwise deletion ($N = 7{,}286$; 15.33% missing) produced similar results in terms of fit. Parameter estimates differed slightly, $\alpha_0 = .329$, SE = .006, $\alpha_1 = -.014$, SE = .006. Although the intercept factor mean obtained in the listwise deletion model deviated from that obtained with the FIML model by only approximately 1%, the difference factor mean was overestimated by approximately 12% in the listwise model compared with the FIML model. Larger differences in bias would likely occur with higher rates of missing data and efficiency differences would be expected with a smaller overall sample size and higher proportion of missing data.

A second model was a cross-lagged panel model examining the reciprocal relationship between depression and self-rated health over the full course of the study. FIML estimation ($N = 8{,}605$) was used with two auxiliary variables, baseline BMI and age, that were expected to be associated with the variables in the model and missingness. This model did not have adequate fit according to the CFI, $\chi^2(20) = 1{,}490.174$, $p <$ .001. CFI = .914, SRMR = .051. Depression significantly predicted self-rated health over the 12-year period, $\beta = -.316$, SE = .034, $\beta^* = -.127$, $p < .001$, and self-rated health significantly predicted depression over the 12-year period, $\beta = -.047$, SE = .006, $\beta^* = -.121$, $p < .001$. A model tested without auxiliary variables had a nearly identical chi-square fit, parameter estimates, and standard errors.

A listwise deletion version of the cross-lagged model ($N = 5{,}237$; 39.1% missing) was tested for comparison. Based on the alternative fit indices, this model had a similar fit to the previous FIML models, $\chi^2(20) = 1{,}174.196$, CFI = .910, SRMR = .050, although the chi-square value was smaller due to the reduced sample size. The cross-lagged paths for depression predicting self-rated health, $\beta = -.308$, SE = .037, $\beta^* = -.127$, $p < .001$, and for self-rated health predicting depression were similar to those obtained with the FIML models, $\beta = -.048$, SE = .007, $\beta^* = -.119$, $p < .001$. The parameter estimates differed from the FIML models by only approximately 1–2%. The standard errors were approximately 9–16% larger, however, which was a more substantial bias. The difference in standard errors might be expected to be even larger had there been a smaller overall sample size.

A third model tested was a growth curve model of BMI across the six waves (12 years) of the health and aging data set. This model ($N = 8{,}605$) used FIML missing data estimation

with age and diabetes diagnosis at baseline serving as auxiliary variables. Because the impact of auxiliary variables is model specific, different auxiliary variables (e.g., diabetes) may be relevant and the same auxiliary models (e.g., age) may play a different role in this model compared with the prior models. The model fit the data well, $\chi^2(16) = 488.160$, CFI = .992, RMSEA = .059. The average intercept was 27.266 and the average slope estimate was $\alpha_1 = .117$, SE = .008, $\alpha_1^* = .222$, $p < .001$, suggesting a significant increase in BMI of about one-tenth of a point per two-year period. Variances of intercept and slope were both significant, 24.388, $p < .001$, and .278, $p < .001$, respectively, suggesting individual variability in the baseline values and growth.

Retesting this model with FIML estimation but no auxiliary variables gave similar results in terms of fit and parameter estimates, $\chi^2(16) = 485.520$, CFI = .992, RMSEA = .058, $\alpha_0 = 27.264$. $\alpha_1 = .120$, SE = .008, $\alpha_1^* = .228$, $p < .001$, ψ_{00}, = 24.388, $p < .001$, and $\psi_{11} = .278$, $p < .001$. The similarity of results may be partly because more complete BMI data from earlier time points were included in the model, and these values are highly related to later BMI values with more missing data. Earlier BMI values thus function as strong auxiliary variables in the model; other auxiliary variables may have little additional effect. Exploratory analyses also indicated that baseline BMI was not related to subsequent dropout, so attrition biases may be minimal. A listwise deletion version of the model ($N = 4{,}217$; 51.0% missing) had a substantially lower chi-square due to the smaller sample size and fit the data well, $\chi^2(16) = 384.253$, CFI = .991, RMSEA = .074. The parameter estimates also were similar, $\alpha_0 = 27.264$, $\alpha_1 = .123$, SE = .010, $\alpha_1^* = .236$, $p < .001$, $\psi_{00} = 24.128$, $p < .001$, and $\psi_{11} = .271$, $p < .001$. Although the slope estimate differed from the FIML model by only approximately 5%, the standard error for the slope estimate was about 25% higher in the listwise deletion model. The difference in efficiency would likely be even more noticeable with a smaller overall sample size.

Comments

A variable with missing values is considered MNAR only if there are no model or auxiliary variables that account for the relationship between missingness and the variable with missing data. To the extent that the association between values of the variable and missingness can be eliminated, the variable meets the MAR condition. But it may be infrequent that such associations can be fully eliminated, and under such circumstances we may look to FIML estimation with auxiliary variables to reduce biases that might occur, albeit not eliminate them altogether.[8] This notion suggests violation of the MAR could be viewed as a matter of degree (Collins et al., 2001; Graham, 2009). As Graham, referring to the MAR assumption for missing data estimators, succinctly states, "Rather than focusing on whether the MI/ML [multiple imputation/maximum likelihood] assumptions are violated, we should answer the question of whether the violation is big enough to matter to any practical extent" (Graham, 2009, p. 567). In other words, we may increase the plausibility of the MAR assumption when auxiliary variables are included even though we may not meet it entirely.

Given that we cannot know whether a variable meets the MAR assumption, it seems that the most reasonable course is to employ missing data analyses that are based on MAR and include auxiliary variables even though we do not know whether appropriate assumptions are met. We may not take this course of action with complete certainty that all biases will be eliminated, but we could proceed with the certainty that the best approaches for mitigating any biases have been used. Naturally, such analyses need to be supplemented with some detailed information about the nature of the attrition pattern (e.g., rates, correlates), discussion of likely causes, and speculation about potential biases. Contrast this

approach with the common practice of analyzing only those who complete the study, using listwise deletion, and omitting any detail about the nature of the attrition. All simulation work to date unequivocally suggests that complete case analysis will only result in maximizing biases that can occur due to attrition. This state of affairs is often difficult for researchers to digest, myself included initially, but it is clear that the better practice is to employ missing data estimation rather than eliminate cases.

Analyses when Values are not Missing at Random

Although inclusion of auxiliary variables may help to reduce potential biases, there is always the possibility that the relationship of the variable of interest with missingness will not be accounted for by auxiliary variables. Several approaches have been developed specifically to be used if the researcher wishes to test models under the assumption that values are MNAR rather than MAR. There is no one preferred approach, and all MNAR approaches require some limiting assumptions, the status of which is also unknown for any given data set. There is a fair amount of consensus that MNAR analysis approaches should be used as sensitivity analyses only (Verbeke, Molenberghs, Thijs, Lesaffre, & Kenward, 2001; Thijs, 2002), because they may produce substantially different results depending on the approach and the assumptions made. If the model assumptions are correct, there is likely no harm in conducting MNAR analyses, but, if the model assumptions are not correct, the MNAR analyses may perform even more poorly than MAR analyses (Demirtas & Schafer, 2003). Although there has been considerable development work done in this area, there are many variations on approaches. Because of limited space, however, I will provide an introduction to only four of the approaches – selection, pattern mixture, shared parameter, and symmetric pattern models.

Selection Models

Basic Concepts. The selection model, proposed by Heckman (1976), combines a substantive model with a model that predicts missingness. When these two models are linked together, the substantive model can be improved by the incorporation of estimates from the missingness model. The selection model is based on a decomposition of the joint probability of the variable of interest and missingness. One can think of the joint probability as representing the extent to which MAR has been violated. Heckman's conceptualization entails that the degree to which values are not missing at random is a function of the conditional probability of missingness given values of **Y** and the marginal distribution of **Y**.

$$f(\mathbf{M},\mathbf{Y}) = P(\mathbf{M} \mid \mathbf{Y}) f(\mathbf{Y}) \tag{13.3}$$

Notice that the conditional probability for missingness $P(\mathbf{M}|\mathbf{Y})$ is the same as the left-hand side of Equations (13.1) and (13.2), the probability equations for MCAR and MAR. The expressions $f(\mathbf{M},\mathbf{Y})$ and $f(\mathbf{Y})$ refer to the density function for the distribution of continuous variables.[9] One can think of any density function as a continuous counterpart to the probability that a binary variable equals a certain value. So, this decomposition shows how estimates from the substantive model, such as a growth curve model, might correct for the MNAR condition by incorporating how variables from the model predict missingness.

The rationale of the selection model formulation is that dependency between **M** and **Y** is the degree that values are MNAR. If we account for the dependency, then $P(\mathbf{M}|\mathbf{Y})$ is zero and the data meet the MAR condition. Following this rationale, Heckman proposed

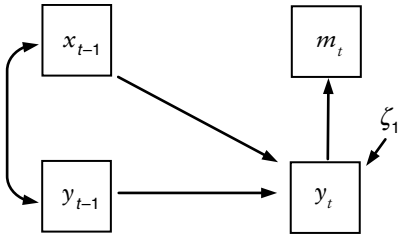

Figure 13.3 Selection Model for a Lagged Regression.

a two-component analysis strategy that predicts the binary outcome of missingness *m* in one model and the continuous outcome of *y* in another model. The two equations can then be combined using predicted probabilities and an appropriate transformation, a process related to the propensity score strategy used for analysis of non-equivalent control group studies.

Instead of this older, two-step process, a structural model based on the selection approach can be estimated with the outcome variable that contains missing values and a missingness indicator included in the model simultaneously. There is more than one possible approach. Most selection models specify a direct path from the outcome variable to the missing data indicator, but a correlation between disturbances might also be included (e.g., Muthén & Jöreskog, 1983). Figure 13.3 depicts one possible selection model in the context of a simple lagged regression model (see Chapter 4). In this depiction, a predictor variable x_{t-1} (e.g, self-rated health) assessed at baseline predicts an outcome at a later time point y_t (e.g., depression) controlling for the baseline assessment of the outcome, y_{t-1}. The missingness indicator m_t is a binary variable representing whether or not the outcome, y_{t-1}, is missing at time *t*. The outcome variable contains some missing data and the indicator variable indicates whether data is present on the outcome variable for each case. Note that the direct outcome and correlated disturbance approaches bear a more than passing resemblance to the extra dependent variable and saturated correlates approaches to auxiliary variables. Because a binary outcome and missing data are involved, an estimation approach is needed that will handle missing data and binary variables. Full ML estimation with binary variables is preferable to WLSMV when data are thought to be MNAR (Asparouhov & Muthén, 2010). This general model could be the basis for several variations or extensions involving latent variables, more time points, more predictors, or more outcomes in the context of repeated measures ANOVA models, and cross-lagged panel models, latent difference models, or growth curve models.

Growth Curve Models. One popular selection model is an application to latent growth curve models. Wu and Carroll (1988) proposed selection modeling with the standard growth curve model by using intercept and slope factors from a separate analysis step to predict missingness indicators at each time point. Figure 13.4a illustrates an example of the Wu and Carroll approach in which dropout indicators at each wave are predicted by intercept and slope factors. In the figure, m_2, m_3, m_4, m_5, and m_6 represent binary dropout indicators for each wave after the first. The dropout indicators are assigned values for each wave or are missing if the case is dropped from the study as a prior time point (sometimes referred to as "survival" or "event history" indicators). For the selection growth model, the predictor variables are the intercept and slope factors, so the model is really accounting for the joint distribution of α_0 and α_1 with **M** instead of **Y**. The presence of missing data and binary dependent variables also suggests the need for full ML estimation. This

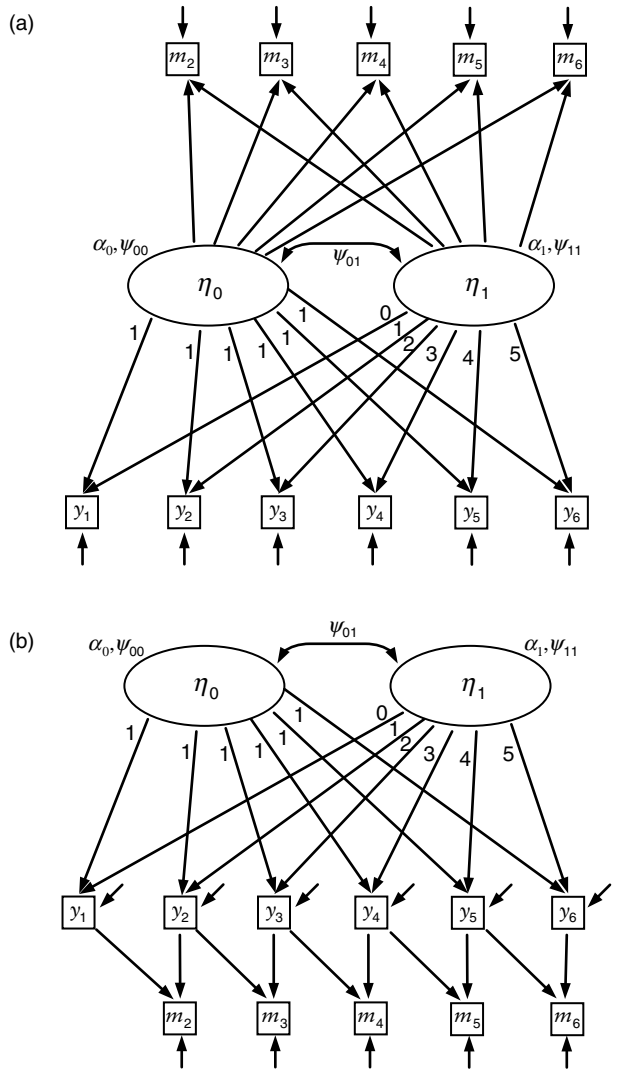

Figure 13.4 Selection Model Specifications for a Latent Growth Curve Model (a) Wu–Carroll Parameter-Dependent Approach and (b) Hedeker–Gibbons Outcome-Dependent Approach.

specification assumes missingness is dependent on intercept (e.g., baseline values) or the slope (i.e. rate of change) random effects.

A variation on the model is shown in Figure 13.4b discussed by Diggle and Kenward (1994). Diggle and Keward proposed lagged effects on each indicator, with m_t regressed on y_t and y_{t-1}, to take into account a possible effect on missingness from the prior time point to model complete data. They liken prediction of the missingness indicator by the concurrent values of the variable to modeling the MAR mechanism. If missingness m_t depends on y_t, then MAR is not met, but if m_t no longer depends on y_t once $y_{t-1}$is taken into account, then the variable meets the MAR assumption (see also Kenward, 1998 for more discussion of this logic). This rationale for including current and prior measurements together is that the MNAR dependency is at least partially modeled, but,

as the proponents of this approach are well aware, there is no guarantee that missing data mechanisms have been fully modeled or that biases have been eliminated (Demirtas & Schafer, 2003). Additional predictors can also be included in an attempt to further account for the dependency.

The Diggle–Kenward approach models the dependency of missingness on observations at each occasion rather than on the intercept and slope parameters, which is the approach of Wu and Carroll (1988). The Diggle–Kenward approach could be said to be *outcome dependent* and the Wu–Carroll approach could be said to be *parameter dependent* (Dantan, Proust-Lima, Letenneur, & Jacqmin-Gadda, 2008).[10] The choice of the outcome-dependent model over the parameter-dependent model must be made on the basis of the assumptions about the nature of the MNAR mechanism – if missingness is dependent on growth parameters or on occasion-specific observed variables – provided both models are not considered. The models thus have different assumptions about the particular variables that account for dependency of complete data values on missingness.

Indicator Coding. The binary missingness indicators can be coded differently, depending on the desired assumptions of the researcher (Enders, 2011). One method of coding the indicators is to code 0 for all present data points up until the point of dropout, and then all subsequent time points are given a code of 1. This could be referred to as permanent dropout and assumes a perfect monotone pattern of missing data. For this approach, all intermittent missing data for a case is coded as if observed and, only when all remaining waves are missing, is a code of 1 used. For example, case 5 in Table 13.1 would receive the codes 0 0 0 0 1 1. An alternative approach is to code all time points prior to dropout as 0 and all time points after the first code of 1 as missing. Case 5 would receive missingness indicator codes of 0 0 0 0 1 ., with the last time point given a missing value for the indicator. This last approach can be described as a survival analysis code. Because both approaches ignore intermittent missing data, the intermittent missing values are assumed to be MAR, with outcome values for these data points estimated with FIML but not included in the selection model as part of the missing data indicator. An alternative coding scheme is a multinomial approach that distinguishes between intermittent missing data points and dropout, where 0 is assigned if the values are intermittently missing, 1 is assigned if dropout, and 2 is assigned if the value is observed. Case 5 in Table 13.1 would then be coded as 2 2 0 2 1 1. This scheme treats the intermittent missing values as MNAR, thus including them as part of the selection model.

Example 13.3: Selection Models

An MNAR selection model examined the effects of self-rated health and depression measured at baseline on self-rated health measured 12 years later (Time 6). The model has the same form as the model shown in Figure 13.3, except that depression was estimated as a latent variable using the three subscale scores as indicators. A single dropout indicator, representing those who dropped out of the study after baseline, was used as an outcome predicted by the self-rated health measured at Time 6. Data were complete for baseline measures, but self-rated health at Time 6 had missing values. During the interval, 2,168 cases (25.2%) dropped from the study and were missing at Time 6. The robust ML estimator for missing data ($N = 8{,}605$) was used to accommodate the binary dropout indicator.

No statistical test of overall model fit was available for this model. The sample size-adjusted Bayesian Information Criterion (aBIC), which could be used to compare to alternative models, was 95,164.556. Depression significantly predicted self-rated health over the 12-year period, $\beta = -.288$, SE = .033, $\beta^* = -.115$, $p < .001$. The cross-lagged effect

was of a similar magnitude to the effect obtained in Example 13.2 using auxiliary variables. The dropout indicator was significantly predicted by Time 6 self-rated health, $\beta = -.372$, SE = .039, $\beta^* = -.225$, $p < .001$, suggesting dependency between missingness and self-rated health. The negative path indicates that those with better health were significantly less likely to drop out of the study. The sensitivity analysis thus suggests some evidence of dependency between self-rated health and missingness, given the model, but the results for the cross-lagged path were not dramatically different from the MAR analysis.

Next, an MNAR selection growth curve model of BMI was tested using the health and aging data ($N = 8{,}605$). The analysis is parallel to the MAR growth curve analysis in Example 13.2, and, thus, is one sensitivity analysis that might be conducted to explore the effects of violation of MAR. The model specification follows that depicted in Figure 13.4b with dropout indicators at each wave as outcomes predicted by the observed variables (Diggle & Kenward, 1994). Robust ML estimation for binary variables was used to obtain logistic estimates of the effects on dropout indicators. Survival coding of the dropout indicators was used with missing values assigned to the indicators at all waves after dropout. There were five dropout indicators for those who dropped out of the study beginning at Wave 2. Paths from BMI observed variables to dropout indictors were included for each synchronous wave and each previous wave (e.g., dropout at Time 2 predicted by BMI at Time 1 and Time 2). Equality constraints were placed on the lagged paths and the synchronous paths across time.

The percentage of dropouts ranged from 4.3% at Time 2 to 13.3% at Time 6. There was no chi-square or standard fit indices for this model, but the aBIC was 213,153.773, which could be used as a comparison to alternative models. The average intercept estimate was 27.264 and the average slope estimate was .119, $p < .001$. The standardized estimate for the average slope was .226. The lagged effect of BMI on dropout indicators was nonsignificant, $\beta_{t,t-1} = -.018$, OR = .982, as was the concurrent effect of BMI on dropout, $\beta_{t,t} = -.003$, ns, OR = .997. The concurrent selection effect suggests trivial dependency on missingness, though they should not be considered a test of MAR. The results are also very close to those obtained in Example 13.2, suggesting that, to the extent that this analysis captures any violation of MAR, there would be little impact of attrition on conclusions.

Comments

The selection model has considerable flexibility in how indicator variables are coded, allowing representation of intermittent and dropout patterns to address different assumptions about missingness. There is also flexibility in how indicator variables are incorporated into the model so that dependency of observed variables or parameters can be examined. Selection models are notoriously sensitive to distributional assumptions and the model specifications, however, and can result in biased estimates relative to MAR. The selection model approach is only one of several MNAR approaches, and, unfortunately, different approaches can produce large differences in results.

Pattern Mixture Models

Basic concepts. Pattern mixture models are an alternative approach to analysis of data that are MNAR (Little, 1994). The approach contrasts with the selection modeling approach, because the implied direction of the path between missingness and the observed variables or model parameters travels in the opposite direction, with missingness indicators predicting observed variables or model parameters. The latent growth curve model, where growth factors or observed variables are modeled as a function of missingness, is the most

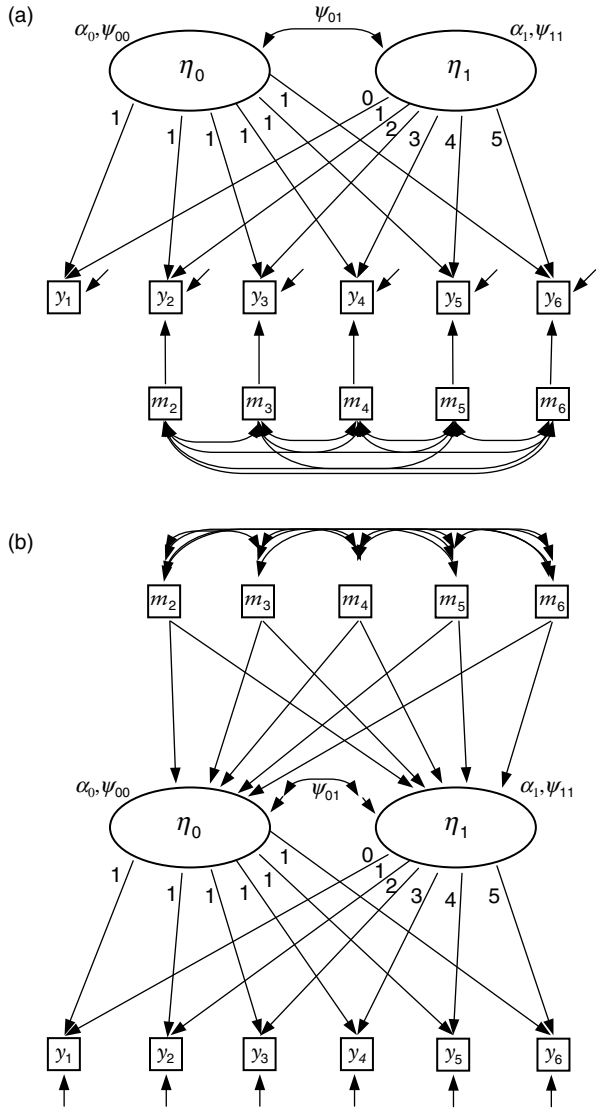

Figure 13.5 Pattern Mixture Model Specifications for a Latent Growth Curve (a) Pattern Mixture Model with Missingness Indicators Predicting Each Observed Variable and (b) Pattern Mixture Model with Missingness Indicators Predicting Growth Parameters.

commonly discussed application of the pattern mixture approach, but other models are possible. Figures 13.5a and 13.b illustrate two variants of such a model.

The rationale for the models stems from a decomposition of the joint probability distribution for **Y** and **M** that differs from the selection model.

$$f(\mathbf{Y}, \mathbf{M}) = f(\mathbf{Y} \mid \mathbf{M}) P(\mathbf{M}) \tag{13.4}$$

The equation resembles Equation (10.3) given for mixture models, which also entails the product of a conditional probability and a marginal probability.[11] The pattern mixture approach gets its name because values can be modeled as mixture distribution of missing

data pattern groups. One way to think about Equation (13.4) is as a distribution of **Y** divided according to two or more missingness patterns **M**.

Equation (13.4) for the pattern mixture factorization contrasts with Equation (13.3) for the selection factorization in that the right-hand side includes the conditional probability of **Y** given missingness multiplied by the marginal probability of missingness. Both equations are valid factorizations of the joint probability of **Y** and **M**, representing the degree to which the values are MNAR. As before, **M** can represent any number of missing data patterns, but it is easier to conceptualize **M** as a dropout pattern matrix for our purposes. The pattern mixture model is sometimes considered to be a more appropriate factorization of the joint probability, because the conditional probability **Y** is divided in terms of the probability that a case is missing and observed, an underlying conceptualization that mirrors the causal process in which **Y** scores are a consequence of missingness rather than a cause of missingness. Following this logic leads to a modeling of missing data in which observed variables or parameters are predicted by a missingness indicator. This is an alternative approach to the grouping method for modeling pattern mixture model. As we might infer from the fact that the selection and pattern mixture models imply a different direction of the causal arrow, the two approaches also involve different assumptions results and may not produce the same results.

Indicator Variable Approach. There are two major specifications approaches for pattern mixture models. One major approach uses missing data or dropout indicators for each wave except the first wave, which is usually assumed complete. In contrast to the illustrations in Figures 13.4a and 13.4b of the selection model, the pattern mixture model regresses the observed *y* values or growth factors on the missing data indicators *m*. Figure 13.5a shows the form with *y* values as the outcome and Figure 13.5b shows the form with growth factors as the outcome. The first model form follows the approach for the multilevel regression approach to growth curves proposed by Hedeker and Gibbons (1997). The second of these models parallels the selection approach of Wu and Carroll (1988) in that the model parameters rather than observed values are dependent on missingness. Dropout indicators are created much like they are in selection models, although, because they function as exogenous variables in the model, there are differences. Survival coding cannot be used, because cases may be eliminated when all cases are missing values for one of the indicators. Indicators with multiple categories representing several patterns also are not allowed, because they are nominal groups. Capturing multiple pattern categories then requires a set of dummy coded variables. With missingness as predictors, however, other forms of dependency can be modeled. Hedeker and Gibbons advocate for testing interactions between dropout indicators and substantive variables in the model to address changes in dependency over time. The choice between modeling observed variables or parameters depends on the assumptions about the potential process by which missingness might bias the model. Little (1995) distinguishes between conditions when dropout may be outcome dependent, in which underlying values at a certain point in time are related to probability of dropout, and conditions when dropout may be parameter dependent, in which changes over time are related to probability of dropout.

Known Class Approach. The other major specification approach estimates the model with several known classes. This approach has evolved out of the standard multigroup model approach originally proposed by Muthén and colleagues (1987). Known latent class models are a special case of latent class models where the groups are defined a priori. Estimates for different missing pattern classes are combined using proportional weighting, as a mixture of different estimates. The mean of the full sample, for example, can be estimated as

a weighted function of the two groups, the complete (observed) and incomplete (missing) groups. If the probability of membership in the incomplete group is π_{mis} and the probability of membership in the complete group is π_{obs}, then the full sample mean is a weighted composite of the two groups, $\hat{\alpha} = \pi_{obs}\alpha^{obs} + \pi_{mis}\alpha^{mis}$. The underlying assumption is that α^{mis} and α^{obs} are from different distributions under MNAR. If the distributions are the same, we have MCAR, or, if we assume independence conditioned on other variables, MAR. Variances or other parameters can be conceptualized similarly.

As a simple example, a rudimentary longitudinal pattern mixture model, such as a simple regression equation where we assume only two patterns of missing data, those who complete the study and those who drop out of the study. We will call the measurement of the later time point that includes missing data y_t^{mis} and the measurement at the prior time point that has complete data y_{t-1}^{obs}. The equation for the mean of the second time point is

$$\bar{y}_t^{mis} = \beta_0^{obs} + \beta_1^{obs}\bar{y}_{t-1}^{mis}$$

The equation tells us that the estimated mean of the missing variable depends on the mean of the variable at the first time point and estimated relationship between the two variables. (Notice that this equation is a simple algebraic manipulation of the intercept formula, $\beta_0 = \bar{y} - \beta_1\bar{x}$.)

The problem with pattern mixture models is that parameters cannot be estimated for missing data. We have no information about the mean, variance, or covariances involving y_t^{mis}, so we cannot estimate the intercept or slope without some identifying assumption. One identifying assumption might be to estimate the mean of y_2^{mis} using the mean from an earlier time point that has complete data, such as y_1^{obs}. This is known as *complete case identification* (Little, 1993), because the estimate of the mean for the variable with missing data is based only on the mean of the complete cases. Molenberghs, Michiels, Kenward, and Diggle (1998) show that this restriction is the same as assuming the MCAR. Thus, the complete data identification is not useful if the purpose of the analysis is to model an MNAR mechanism.

This simplified example can be extended to additional time points, additional missing data patterns, and other types of longitudinal models, such as latent growth curve models. As additional time points are added, missing data patterns might include only completers and dropouts, or several patterns such as cases with intermittent missingness, those with complete data, and dropouts. The more missing data patterns that are involved, the more parameters for a given model will be unidentified. For example, in a growth model with three time points, a linear effect would not be identified for cases missing two time points and a quadratic effect could not be estimated for individuals missing one time point. Distinguishing between intermittent and attrition patterns yields additional pattern groups. A priori grouping is one way to address these complexities.

With several time points, complete case identification would use the conditional distribution of completers for the conditional distribution of other patterns with insufficient data. *Neighboring case identification* uses the conditional distribution from most similar missing data pattern for the unidentified group. Considering only attrition patterns, if there were three waves and missing patterns were divided into completers, dropouts after the first wave, and dropouts after the second wave, the neighboring case identification approach would use the conditional distribution from the group that dropped out after first wave for the conditional distribution of the group that dropped out after the second wave. *Available case identification* uses a weighted average of the distributional values from all other patterns. For four waves and only attrition patterns, estimates for

the dropouts after the third wave would be based on a weighted average of all the conditional distributions from the completers, dropouts after the first wave, and dropouts after the second wave. Molenberghs and colleagues (1998) show that the pattern mixture model using available case identification is equivalent to assuming values are MAR. In other words, missingness at later time points may still be dependent on later values under this identifying restriction, a condition that has been referred to as *future time dependence* (Kenward et al., 2003). Kenward and colleagues propose an analysis strategy that is not future dependent, given one dropout pattern (e.g., dropouts start with the third wave). Missingness indicators for all prior time points are included in the model and conditional distributions for all other dropout patterns are used for the unidentified estimates.

The pattern mixture approach to growth curves modeling estimates weighted composites of any parameters from the model, such as intercept and slope means or variances from a growth curve model. When slope means are expected to differ between dropouts and complete cases, for example, an MNAR adjusted composite would be proportionally weighted by the two groups, $\hat{\alpha}_1 = \pi_d \alpha_1^d + \pi_{nd} \alpha_1^{nd}$, where $\hat{\alpha}_1$ is the mixed population estimate of the slope factor mean, π_d and π_{nd} are the known proportions of dropouts and non-dropouts (completers), respectively. When there are too few time points for a parameter to be identified, such as for a group with only two time points available for a model with a quadratic slope, then estimates from complete cases, a neighboring pattern group, or a weighted average of other pattern groups (available case) would be substituted.

This type of mixture model estimation (McLachlan & Peel, 2000) is used to obtain a common estimate using posterior weighting across groups, which contrasts with the traditional multigroup approach that estimates parameters within groups. Either parameter estimates must be combined with programming or manual computations (see Enders, 2011, pp. 309–312, for details) or a single latent class can be used to combine the known class parameter estimates (Kim, Mun, & Smith, 2014). For manual estimates, careful attention to standard error computation is needed, as the obtained estimates are not appropriately derived automatically in most programs. Multivariate delta method standard errors, which entail obtaining first derivatives of the estimates and parameter covariance matrices, are recommended (Hedeker & Gibbons, 1997; Hogan & Laird, 1997).[12]

The known class specification is flexible, because any of the structural modeling parameters, such as measurement intercepts, loadings, latent means, and latent variances, can be constrained equal across groups for sensitivity analyses using different assumptions. In contrast to the older multiple group approach, the known class method allows for incomplete covariance matrices within each of the known groups. Chi-square difference tests are valid to compare parameters across groups (Kim et al., 2014), providing a potentially useful method for exploring variation in longitudinal measurement invariance across missing data patterns. The approach also allows for specification of different pattern identification strategies through complex model constraints.

The two major approaches to specification of the pattern mixture model, the indicator approach of Hedeker and Gibbons (1997) and the known class approach, should usually produce comparable results, although the indicator approach makes some implicit assumptions about some model parameters across pattern groups. The two approaches might be considered loosely analogous to the standard multiple-indicator multiple-cause (MIMIC) and multigroup modeling approaches to group comparisons. Both have the same general aims yet have potentially different assumptions and different advantages in terms of specification flexibility (see Chapter 3 for more discussion of the relationships between MIMIC and multigroup models).

Example 13.4: Pattern Mixture Latent Growth Curve Model

Three versions of pattern mixture models for latent growth curves were used to illustrate MNAR sensitivity analyses. As in previous models, linear growth of BMI over six waves was investigated. The first model was an outcome-dependent model as presented in Figure 13.5a and the second model was a parameter-dependent model as presented in Figure 13.5b. Coding of the dropout indicators for these models was the permanent dropout method in which the indicator receives a code of 1 for all waves subsequent to the initial wave of dropout. The predictive paths were estimated with equality constraints across time. These models are based on standard FIML as no special estimation is need for binary exogenous variables. The third model (not shown in a figure) was a known class pattern mixture model in which a single combined estimate was obtained through weighted averaging of the class-specific estimates using a single latent class variable. Six known classes were used to represent completers and five possible points of dropout (Waves 2 through 6). This grouping ignores intermittent missing data patterns (i.e., a monotone pattern is assumed) and thus assumes MAR for these patterns. Fewer known classes also could be used in any number of desired groupings (e.g., completers, dropouts, intermittent).

The results from the three models are summarized in Table 13.3. As is clear from the table, the three specifications produced very similar results. And in each case, results were similar to the selection model and MAR analyses. The outcome-dependent analysis results may not be valid, because there were estimation problems (nonpositive definite matrix warning). The estimation issues appeared to stem from a very weak relationship between the dropout indicators and the observed variables, $\beta = -.015$, SE = 1.012, $\beta^* = -.001$, ns. When these coefficients were constrained to zero, the results did not change and the convergence issue was resolved.

The parameter-dependent model required longitudinal equality constraints on the path coefficients between the dropout indicators and the slope parameter in order to achieve convergence. In contrast to the results from the outcome-dependent model, the effects of the dropout indicators in the outcome-dependent model did significantly predict the slope factor, $\beta = -.050$, SE = .021, $p < .05$, average $\beta^* = -.022$, and significantly predicted the intercept factor for three out of the five indicators (unstandardized coefficients ranged from –.250 to –.937 and standardized coefficients ranged from –.016 to –.036). There thus appears to be little dependency between dropout and observed variables but more substantial dependency between dropout and growth parameters, suggesting that attrition may be related to changes in BMI over time more than occasion-specific observed values.

Comments

Pattern mixture models represent an alternative approach to selection models and have considerable flexibility for modeling MNAR with longitudinal models. Application of the pattern mixture modeling approach has almost exclusively been with growth curve models, however. Every simulation I have encountered, for example, has used growth curve models. It is worth underscoring that biases or efficiencies of pattern mixture models (as well as other MNAR approaches) will be sensitive to the type of model specified. The latent growth curve model, for instance, is based on differences over time, and the average intercept and slope parameter estimates may be affected by mean dependency on missingness but less sensitive or not sensitive at all to variance dependency on missingness. Variances of slopes and intercepts would be expected to be biased by dependency of variances on missingness. Where the covariance elements differ across patterns, we may expect differences in the intercept-slope covariance, the effects of covariates, or any

Table 13.3 Results from three Specifications of Pattern Mixture Growth Curve Models of BMI

	Outcome-dependent				*Parameter-dependent*				*Known class*			
Parameter	*Estimate*	*SE*	*Standardized estimate*	*Sig.*	*Estimate*	*SE*	*Standardized estimate*	*Sig.*	*Estimate*	*SE*	*Standardized Estimate*	*Sig.*
α_0	27.264	.055	5.521	<.001	27.433	.066	5.555	<.001	27.264	.055	5.521	<.001
α_1	.120	.008	.228	<.001	.127	.009	.240	<.001	.120	.008	.228	<.001
ψ_{00}	24.388	.395	1.000	<.001	24.310	.394	.997	<.001	24.388	.401	1.000	<.001
ψ_{11}	.278	.008	1.000	<.001	.278	.008	.998	<.001	.278	.008	1.000	<.001
ψ_{01}	–.328	.041	–.126	<.001	–.331	.041	–.127	<.001	–.378	.041	–.126	<.001
χ^2	624.379[a]			<.001	587.896			<.001		NA		
df	45				40					NA		
CFI	.990				.991					NA		
RMSEA	.039				.040					NA		

[a] Solution had a nonpositive definite matrix.
NA = not available.

autoregressive components. The point is that biases will be a function of the specific model tested, and attention must be paid to the specific structural model when attempting to account for potential biases of attrition.

Shared Parameter Models

Shared parameter models (Albert & Follman, 2009; Beunckens, Molenberghs, Verbeke, & Mallinckrodt, 2008; Roy, 2003) represent yet another approach to estimation of data that are MNAR. The shared parameter model differs from the other two approaches to MNAR analysis, because the variables of interest or the model parameters (e.g., slope and intercepts) as well as the missingness indicators are jointly modeled as function of an unobserved random variable, either a latent class variable or a continuous latent variable. In other words, the arrow points from the random variable as a common cause of both the observed variables (or parameters) and the missingness indicators. Let us concentrate first on the latent class approach and then discuss other related approaches. A review of Chapter 10, which introduced latent class analysis, may be in order.

The latent class MNAR analysis is similar to a pattern mixture model in that parameters are estimated as a function of groups of missing data patterns, but, whereas pattern mixture models use known groups for missing data patterns, shared parameter models use an estimation of missing data pattern groupings given the observed variables and model parameters. One cited advantage of the latent class model in this context is that many missing data patterns can be grouped on an empirical basis (Dantan et al., 2008). Analysts using pattern mixture models are usually faced with many missing data patterns, with some having too few cases to include as a separate group, so fewer groups are created on a theoretical or empirical basis.

The shared parameter latent class model is based on the factorization of the joint distribution of **Y** and **M** where the dependency is on a latent class factor, η^C (Roy, 2003).

$$f(\mathbf{Y},\mathbf{M}) = \sum_{c=1}^{C} f\left(\mathbf{Y} \mid \eta^C\right) P\left(\mathbf{M} \mid \eta^C\right) P\left(\eta^C\right) \tag{13.5}$$

On the right-hand side of Equation (13.5) each factor is conditional on the latent class variable, indicating that the distribution of **Y** and the probability of missingness can vary across classes weighted by the probability of class membership $P\left(\eta^C\right)$. As this factorization suggests, any structural model that includes *y* and missingness indicators, *m*, can be viewed as a shared parameter model if grouped by a latent class variable. Figure 13.6 shows one possible model in which a latent class variable is used to classify both growth parameters and missingness indicators (Muthén, Asparouhov, Hunter, & Leuchter, 2011). Such a model allows the dependency between the outcome and missingness to be empirically grouped. Notice that the factorization has the basic structure of the pattern mixture model in Equation (13.4) in that joint probability is a function of the distribution of **Y** and probability of **M**.

The latent class MNAR analysis differs from the pattern mixture analysis in another fundamental way. Although both are methods of grouping cases, the known classes of the pattern mixture model constitute a fixed effect, whereas the unknown classes of the latent class shared parameter model constitute a random effect. More generally, the mathematical rationale of the shared parameter model is a factorization of the joint distribution of **Y**, **M**, and a vector representing one or more random parameters, θ (Albert & Follman, 2009).

$$f(\mathbf{Y},\mathbf{M},\theta) = f(\mathbf{Y} \mid \mathbf{M},\theta) P(\mathbf{M} \mid \theta) f(\theta) \tag{13.6}$$

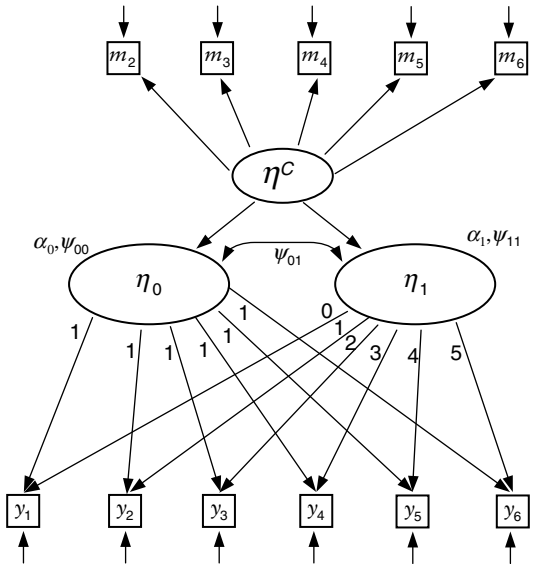

Figure 13.6 Example of a Shared Parameter Model with a Latent Class Factor.

The joint distribution of the observed variables, missingness patterns, and the parameters is factorized into conditional distributions of **Y** given missingness patterns and the parameters, the conditional probability of missingness given the parameters, and the marginal distribution of the parameters. Like the latent class version of the factorization given in Equation (13.5) the factors on the right-hand side of Equation (13.6) are conditional on the random parameter.

It is instructive to consider how the shared parameter model contrasts with the selection model and the pattern mixture model. Conceptually the shared parameter model differs from the other models, because both missingness indicators, m, and observed variables, y, are modeled together with a random parameter of interest, θ (e.g., slope factor). The selection model, in contrast, specifies parameters as the cause of missingness indicators with no direct relationship between y and m. Figure 13.7, which serves as a general guide for distinguishing among the major MNAR approaches, illustrates this distinction.[13] The pattern mixture model differs from the selection model and the shared parameter, because it does not model a direct relationship between the parameter and the observed y. The diagram is a useful overview, but it is a simplification. Some selection models (Figure 13.4a) are more closely related to the shared parameter model because y is modeled as a function of m and θ. In the Wu and Carroll (1988) selection growth model, for example, observed variables and missingness indicators are a function of the random growth factors. This is why such a selection model also is said to be a shared parameter model (Albert & Follmann, 2009). Albert and Follman discuss the connection between these models in greater detail. The pattern mixture approach differs more substantially from the shared parameter model, however. Even though one specification of the pattern mixture growth curve model (see Figure 13.5b) has a direct relationship between the missingness indicator and the parameters, the relationship between the missingness indicator and the observed variable is indirect.

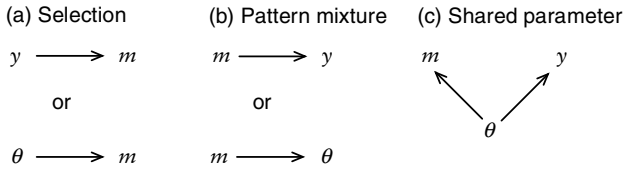

Figure 13.7 Summary Diagram of MNAR Modeling Concepts (Adapted from Diggle, Heagerty, Liang, & Zeger, 2002, figure 13.3, p. 303, by Permission of Oxford University Press).

Example 13.5: Latent Class Shared Parameter Model

A latent class shared parameter model was specified to illustrate one approach to MNAR sensitivity analyses. This analysis followed the general method outlined by Roy (2003). Because the number of classes is unknown, prior experience, theory, or some empirical guesswork is needed to arrive at a desired number of classes. The Lo–Mendell–Rubin adjusted likelihood ratio test (LMR) and the Vuong–Lo–Mendell–Rubin likelihood ratio test (VLMR) were used to compare to fit of an alternative model with one fewer classes to aid in selecting the optimal number of classes. An initial model specified three latent classes. The sample size-adjusted Bayesian Information Criterion (aBIC) was 194,130.221, which can be used to compare to models with a different number of classes. The estimated proportions of the three classes were .849, and .123, .027, which could be used for manual computation of a weighted average of any parameter across classes. The interpretation of these classes would need to be determined using theory, plots, or information from external variables. The combined parameter estimates (obtained from a separate run) produced growth parameter estimates that were highly similar to the prior MNAR tests, $\alpha_0 = 27.264$, $\alpha_1 = .120$, $p < .001$, $\psi_{00} = 24.388$, $p < .001$, $\psi_{11} = .278$, $p < .001$, $\psi_{01} = -.328$, $p < .001$. The LMR (9.069, $p = .1027$) and VLMR (9.402, $p = .092$) statistics did not quite reach significance levels, suggesting that a two-class model would not fit significantly better. A four-class model was tested as a comparison, and the aBIC was 194,137.144, slightly worse than the prior model. The LMR (10.343, $p = .091$) and the VLMR (10.723, $p = .082$) tests were not quite significant, suggesting that the fit of the three-class model was not much better than a four-class model. The estimated class proportions, .292, .134, .021, and .553, indicated one low-frequency class, suggesting little additional benefit of a fourth class. Overall, the results leave ambiguous the number of classes and their interpretation.

Symmetric Pattern Models

Symmetric pattern models are a lesser known alternative method MNAR analysis. Although originally proposed for handling nonignorable data in Item Response Theory (IRT) analyses (Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999), they fit neatly into the shared parameter framework, because they model missingness and observed outcome variables as a function of latent variables. Figure 13.8a is a simple representation. A cursory comparison to the shared parameter diagram in Figure 13.7 suggests that the symmetric pattern model concerns the joint distribution of **Y** and **M** which are independent, conditional on the latent variable. In this sense, the model follows the shared parameter factorization in Equation (13.6).

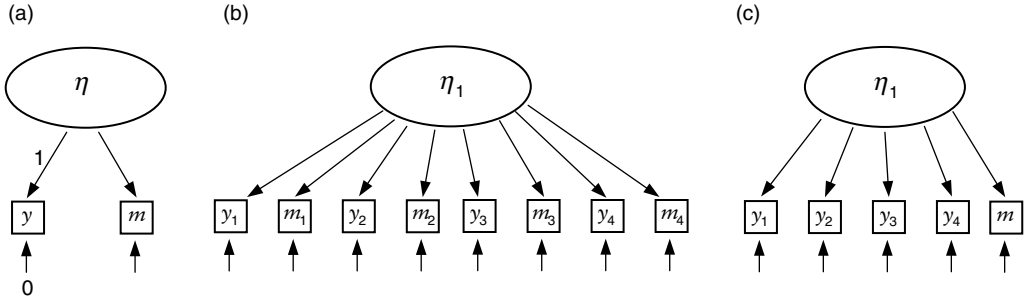

Figure 13.8 Simple representation of a symmetric pattern model (a) single *y* and single *m*, (b) multiple *y* and multiple *m*, and (c) multiple *y* and single *m*.

There are several potential specifications that might be used, including single indicators for *y* and *m* for each factor (Figure 13.8a), multiple indicators of both *y* and *m* for each factor (Figure 13.8b), or multiple *y* indicators and a single *m* indicator for each factor (Figure 13.8c). The latter specification could be used to represent unit missingness in which *m* is coded as missing when all y_j for each time point are missing. As with other MNAR model approaches, missingness indicators might be coded in several ways to capture different assumptions about dependency over time (e.g., intermittent, monotone, survival). A latent variable with only two indicators is not identified, of course, so constraints are necessary. It would not make sense to constrain the loadings for *y* and *m* to be equal to one another in this case, and the *m* loading should be free to represent the dependency of the latent factor to missingness. Therefore, setting the loading for *y* equal to 1 and constraining the measurement residual variance for *y* to be equal 0 would provide one logical variance scaling for each factor and would follow the single-indicator latent variable specification. As with the quasi-simplex model, identification may be achievable with freely estimated loadings for *y* with equality constraints, if a sufficient number of time points are available and two of the measurement residuals are set equal to 0. An estimation approach, such as full information ML, that allows both continuous and categorical indicators is required. Each *y* would not have to be continuous, however, and the extension of the approach to binary, ordinal, or count variables would be a simple matter.

Although not an approach that is fully distinguishable from the shared parameter model, the symmetric pattern model opens up a variety of longitudinal modeling options, including longitudinal measurement invariance and MNAR estimate for multiple indicator latent variables for virtually any longitudinal model. Falcaro, Pendleton, and Pickles (2013) discuss several such models. Figures 13.9 and 13.10 illustrate two possible longitudinal applications proposed by Falcaro and colleagues. Although not required, each model in the figure omits the missingness indicator from the first wave under the presumption there is complete data. Notice that the latent growth curve specification in Figure 13.10 differs from that of the selection or pattern mixture specifications, because the missingness indicator is related to a common factor at each occasion rather than directly to the observed *y* variables or the growth parameters. Dependency between missingness and growth parameters cannot be modeled in this growth curve specification without additional paths estimated.

Because each factor must be defined by observed variables and one or more missing data indicators, the fit of the measurement portion of the model is likely to suffer unless the association between missingness and other indicators is strong. Missing data indicators may load poorly on the factor and empirical identification issues may arise, leading to additional model constraints in order to identify the model. This will be the case, in

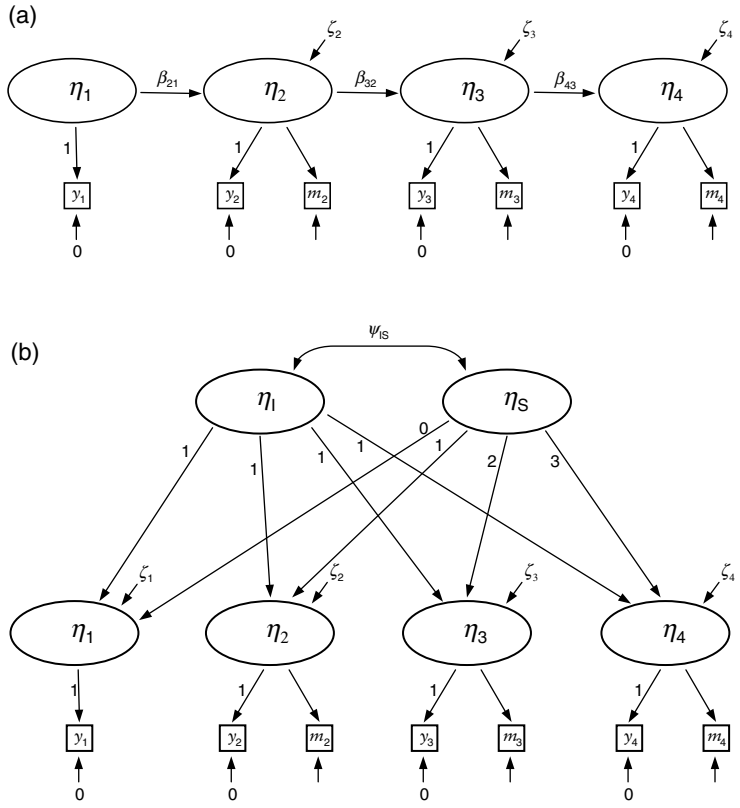

Figure 13.9 Symmetric Parameter Simplex Model.

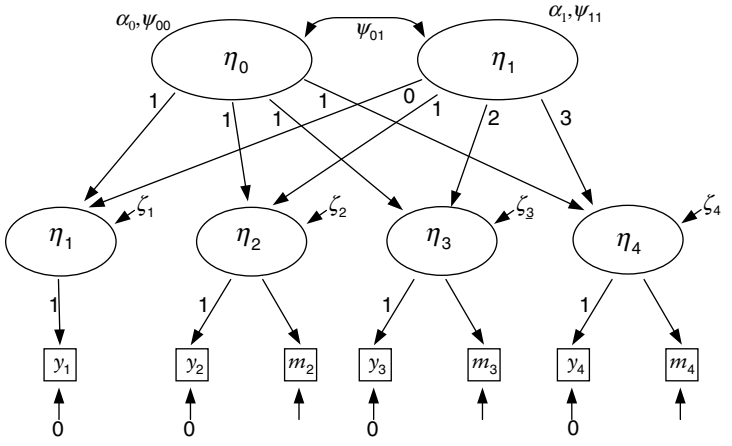

Figure 13.10 Symmetric Parameter Latent Growth Curve Model.

particular, if there is only one or only a few substantive indicators for each factor. Some possible constraints that would resolve identification issues might include equality on loadings for missingness indicators equal across waves, equality constraints on multiple substantive indicators of a factor for a given occasion, or other longitudinal equality constraints on parameters of interest.

Example 13.6: Symmetric Pattern Models

Application of the symmetric pattern model was illustrated with a simplex model and a growth curve model. The simplex model was a variant of the perfect simplex model for self-rated health in Example 5.5 that used six waves of the single self-rated health item from the health and aging data set ($N = 8{,}605$). Specifications were as shown in Figure 13.9, but longitudinal equality constraints were placed on loadings for missing data indicators and autoregressive paths to identify the model. In addition, autoregressive coefficients and disturbances were constrained to be equal over time. Robust ML estimation was used to accommodate the binary missingness indicators. The model had a large Pearson chi-square value, $\chi^2(26) = 13{,}708.194$, $p < .001$, aBIC =172,786.502. The fit of the model is not central to the goal of the MNAR sensitivity analyses, however. Missing indicators did load significantly on the factor, with standardized loadings ranging from .692 to .726, and this suggested a potentially important level of dependency. The autoregressive estimates were significant, with the standardized values suggesting a very stable construct over time, ranging from .996 to .997. Comparison to path estimates from a similar MAR model with autoregressive constraints suggested that missingness dependency may have a fairly substantial biasing impact on the autoregressive estimates, as the standardized estimates from the MAR model were approximately .7.

A second model demonstrated the specification of a symmetric pattern latent growth curve model, structured as illustrated in Figure 13.10. As a comparison with the selection, pattern mixture, and shared parameter models, a linear growth model of BMI over six waves was tested. Equality constraints on first-order residual variances, $\text{Var}(\zeta_t)$, were required to identify the model (i.e., the homogenous variance assumption). The resulting model also had an extremely poor fit, $\chi^2(58) = 20{,}334.018$, $p < .001$. Missingness indicators loaded significantly on each occasion-specific factor, although the standardized values were relatively small (between .159 and .172). Parameter estimates for the growth model were similar to those obtained from MNAR and MAR analyses earlier, $\alpha_0 = 27.316$, $\alpha_1 = .127$, $p < .001$, $\psi_{00} = 24.258$, $p < .001$, $\psi_{11} = .307$, $p < .001$, $\psi_{01} = -.391$, $p < .001$, suggesting no change in conclusions from the results.

Extensions

The preceding discussion provides only an introduction to the possible MNAR analyses. An SEM framework that includes latent class factors makes possible the combination of any number of the elements of selection models, pattern mixture models, and shared parameter models. Muthén and colleagues (2011) provide several examples of how elements of the major MNAR approaches can be integrated to model missingness dependency with latent growth curve models. As one example, they argue that the basic latent class pattern mixture growth model of Roy (2003) may obscure differing trajectories when missing data patterns are also classified, and they propose a modification to disentangle the two by separate latent classification of trajectories and dropouts. Beunckens and colleagues (2008) propose a hybrid that combines parameter-dependent selection and latent class shared parameter modeling of growth curves. Gottfredson, Bauer, and Baldwin (2014) present an alternative hybridization of the selection model and latent class shared parameter approaches. Yuan and Little (2009) discuss a more saturated hybrid of the shared parameter model that considers direct dependence between observed values and missingness in addition to the shared parameter.

I did not consider more complex missingness dependency, such as nonlinear dependencies, multiple patterns, time-to-dropout survival models, and interactions, and these are important issues to take into account. The structural models for MNAR variables may

also be adapted for use with binary, ordinal, or count observed variables for SEM software programs with features for these types of variables. Within the SEM literature, missing data estimation and MNAR modeling, such as pattern mixture models, are rarely illustrated with models other than growth curve models (cf. Falcaro et al., 2013), but MNAR sensitivity analyses are certainly possible for the other longitudinal models discussed in this text.

Attrition also may be modeled as MNAR using methods related to either discrete (e.g., Yuan & Little, 2009) or continuous-time survival analysis (e.g., Hogan & Laird, 1997). SEM approaches to survival analysis are introduced in Chapter 12 and could be used to build upon some of the MNAR models described in this chapter. One coding approach to occasion-specific missing data indicators, for example, is closely related to discrete-time survival analysis (Enders, 2011; Muthén et al., 2011). These approaches are designed in part to take into account the propensity for earlier drop out versus later drop-out, more explicitly modeling the dependency between time of dropout and values on the observed outcome.

Comments

The primary focus of the chapter has been on handling missing data due to attrition and I have given much less attention to intermittent missing data. The lack of emphasis is not intended to suggest that researchers should ignore intermittent missing data patterns. Part of the reason for less emphasis on the intermittent pattern is that a smaller proportion of missing data in longitudinal studies is due to intermittent data than is due to attrition, so its biasing effects may be of secondary concern compared with attrition. Part of the reason is also that general discussions about missing data, amply represented elsewhere, are applicable to the intermittent and cross-sectional cases but not always informative about what to do with attrition. The reasons for intermittent missing data are likely to be different from the reasons for attrition, even though intermittent missing values may also not always meet MAR assumptions. Olsen and Schafer (2001) suggest a dual process of modeling intermittently missing data and dropouts. Muthén and colleagues (2011) show how such a model can be specified with a structural modeling approach.

Taking a step back, attrition deserves special focus when discussing missing data because of its presumed risk of violations of MAR. The good news is that incorporation of auxiliary variables with otherwise standard FIML analysis presents excellent prospects for reducing or eliminating some of the potentially biasing effects of attrition. The bad news is that we can never be certain whether we have eliminated all sources of the violations of MAR. For that reason, MNAR sensitivity analyses may be a reasonable common course for longitudinal models. But they must be considered sensitivity analyses rather than a superior alternative to MAR analyses. In Kenward's (1998) terms, "The MNAR analysis tells us about inadequacies of the original model rather than the adequacy of the MNAR model" (pp. 2731–2732).

MNAR analyses present a dizzying array of approaches, variations on specifications, and implied assumptions, making consideration of MNAR analyses a daunting task. Because results may vary by any of these choices and because the true missing data processes are unknown, it is impossible to know which modeling strategy is best or most trustworthy. Comparisons of MCAR analyses to MAR analyses will not shed any light on whether or not MAR assumptions have been met. If, however, the results of different sensitivity analyses do not suggest any changes in conclusions, it may be reasonable to conclude that there are few biases due to attrition that would change the results. If results differ, researchers are left only with reporting the range of possible conclusions and attempting to apply theoretical reasoning or external data to interpreting the results and anticipating potential biases (Muthén et al., 2011).

Recommended Readings

There are several accessible book length introductions to missing data analysis (Allison, 2002; Enders, 2010, McKnight, McKnight, Sidani, & Figueredo, 2007). More technical treatments can be found in Little and Rubin (2002) and in Schafer (1997). Several articles with broad overviews of missing data are useful as well, particularly for their more general perspectives (Graham, 2009; Rubin, 1976; Schafer & Graham, 2002). Perusal of some of the key simulation studies that demonstrate the unbiased and efficient estimation of modern missing data analysis under MAR is worthwhile for insights into their advantages, optimal conditions, and limitations (Enders, 2001a; Enders & Bandalos, 2001; Collins, et al. 2001). More on estimation of missing data specifically for structural equation models can be found in Allison (2003), Arbuckle (1996), and Enders (2001c, 2013). General discussions of missing data analysis with emphasis on longitudinal studies are rare but very valuable (Feng et al 2012; Graham, 2009; Hedecker & Gibbons, 2006, chapter 14).

Notes

1 The missingness indicator is more commonly denoted by **R** in the literature, where a value of 1 indicates that there is a response and a value of 0 indicates the response is missing. Without loss of the applicability of any of the equations, I follow Little and Rubin's (2002) use of **M** instead, in part to avoid confusion that might arise by referring to **R** as indicating missingness.

2 Equations (13.1) and (13.2) are simplified. A more complete equation (e.g., Little & Rubin, 2002, p. 12) includes an additional element, ϕ, representing probability parameters that govern the relationship between missingness and variables in the model, $P(\mathbf{M} \mid \mathbf{Y}, \phi) = P(\mathbf{M} \mid \mathbf{Y}^{obs}, \phi)$. As the more complete equation suggests, **M** is conditioned on ϕ as well as elements of **Y**, so the form of the relationship of missingness to values should also be taken into account.

3 The term "mechanism" can be confusing and is not used consistently throughout the literature. Most often the mechanism refers to whether a variable falls under MCAR, MAR, or MNAR, but it may also refer to the form of the relationship between missingness and the variable of interest (e.g., linear, curvilinear) or the probability distribution of missingness, not to mention the occasional more casual use of the common meaning that implies causal processes.

4 Multiple imputation is possible with SEM, but it is not often employed. Mplus, for example, has special features that help with multiple imputation estimation. A two-stage process is also possible in which covariances and means are estimated using a separate software package that employs FIML estimates using an expectation maximization (EM) algorithm (see Graham, 2003) and then these values are used as input into SEM software. The two-stage approach is less convenient and is infrequently used because nearly all major SEM software programs allow for direct FIML estimation.

5 This data set is comprised of simulated values that closely match the Health and Retirement Study beginning in 1994, which is the second wave of the study. The original data set that served as the model for these data in this example started with respondents who had complete data in 1994, so a portion of the original sample that did not include data at the second time point were eliminated. It is therefore likely that the dropout rate presented here is underestimated as a result, because those who completed both the first and second waves should be less likely to drop out later in the study. This example is also overly simplistic because I examine only one variable, BMI. Although not many, there are some cases missing that are not missing for an entire wave, so the designation of attrition may be somewhat inaccurate when considering other variables in the data set.

6 I did not constrain the variances of the variables in the two groups to be equal because the prior test indicated they did not differ significantly, but the covariance could differ in other examples simply because the variances of one or both of the variables differed across groups.

7 I use the phrase "more complete" here because the auxiliary variable itself need not necessarily have complete data. Data may be missing on the auxiliary variable, and even when the mechanism of missingness for the auxiliary variable is MNAR, there appears to be no bias in the estimation of the model parameters of interest (Enders, 2008).

8 I have neglected the possibility that associations among the variable of interest, missingness, and auxiliary variables are not linear, but nonlinear relations may also exist (Collins et al., 2001). If these relations are assumed to be linear in the modeling process but are not, biases will not be eliminated.

9 As before, this formulation is a simplification of the more formal expression. Each of the elements within the parentheses usually includes a vector of unknown parameters representing probability parameters that govern the relationship between missingness and variables in the model, typically symbolized by ϕ, ψ, ξ, θ, or ω, designated separately for each component. A convenient way to think about the simplified expression is that there is an underlying assumption that the model is correctly specified in each factor.

10 The parameter-dependent pattern mixture model is also sometimes referred to as a "shared parameter" model because missingness is modeled as function of a random effect (e.g., Enders, 2011). I return to this topic later under the section Shared Parameter Models.

11 The pattern mixture equations are also simplifications, because the more formal statement includes the parameter vector, ϕ.

12 The delta method for standard errors is used with the robust or sandwich estimation available in many software programs (Huber, 1967; White, 1982) when robust maximum likelihood or WLSMV estimation is requested. The multivariate delta method standard errors will be obtained when using the mixture modeling with single latent class method.

13 Diggle and colleagues (Diggle, Heagerty, Liang, & Zeger, 2002, p. 303) also provide a useful diagram as an overview of the MNAR modeling approaches. Although Figure 13.7 was inspired by their diagram concept, it takes into account the wider array of MNAR analyses that have been proposed, especially within the SEM framework.

References

Albert, P. S., & Follmann, D. A. (2009). Shared-parameter models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 433–452). Boca Raton, FL: Chapman & Hall/CRC.

Alderman, H., Behrman, J. R., Kohler, H. P., Maluccio, J. A., & Watkins, S. C. (2001). Attrition in longitudinal household survey data. *Demographic Research*, *5*, 79–124.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545–557.

Arbuckle, J. L. (1996) Full information estimation in the presence of incomplete data. In G. A. Marcoulides and R. E. Schumacker (Eds.), Advanced structural equation modeling: Issues and techniques. Mahwah, NJ: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. (2008). *Auxiliary variables predicting missing data.* Unpublished technical report. Retrieved from http://statmodel2.com/download/AuxM2.pdf.

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data.* Unpublished technical report, retrieved from http://www.statmodel2.com/download/BayesAdvantages18.pdf.

Beunckens, C., Molenberghs, G., Verbeke, G., & Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics*, *64*, 96–105.

Cabrera, A. F., Nora, A., & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education*, *64*, 123–139.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. New York: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Dantan, E., Proust-Lima, C., Letenneur, L., & Jacqmin-Gadda, H. (2008). Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts. *The International Journal of Biostatistics*, *4*, 1–26.

Deeg, D. J., van Tilburg, T., Smit, J. H., & de Leeuw, E. D. (2002). Attrition in the Longitudinal Aging Study Amsterdam: The effect of differential inclusion in side studies. *Journal of Clinical Epidemiology*, *55*, 319–328.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, *22*, 2553–2575.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, *43*, 49–93.

Diggle, P., Heagerty, P., Liang, K. Y., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford, UK: Oxford University Press.

Dixon, W. J. (1988). *BMDP statistical software manual: to accompany the 1988 software release*. Los Angeles: University of California Press.

Enders, C. K. (2001a). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*, 713–740.

Enders, C. K. (2001b). The impact of non-normality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*, 352–370.

Enders, C. K. (2001c). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128–141.

Enders, C. K. (2002). Applying the Bollen-Stine bootstrap for goodness-of-fit measures to structural equation models with missing data. *Multivariate Behavioral Research*, *37*, 359–377.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling*, *15*, 434–448.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological Methods*, *16*, 1–16.

Enders, C.K. (2013). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller, Structural equation modeling: A Second course (2nd ed., pp. 493–519). Charlotte, NC: Information Age Publishing.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457.

Falcaro, M., Pendleton, N., & Pickles, A. (2013). Analysing censored longitudinal data with non-ignorable missing values: depression in older age. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*, 415–430.

Feng, D., Cong, Z., & Silverstein, M. (2012). Missing data and attrition. In J. T. Newsom, R. N. Jones, & S. M. Hofer (Eds.). *Longitudinal data analysis: A practical guide for researchers in aging, health, and social science* (pp. 71–96). New York: Routledge.

Goodman, J. S., & Blum, T. C. (1996). Assessing the non-random sampling effects of subject attrition in longitudinal research. *Journal of Management*, *22*, 627–652.

Gottfredson, N. C., Bauer, D. J., & Baldwin, S. A. (2014). Modeling change in the presence of nonrandomly missing data: Evaluating a shared parameter mixture model. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 196–209.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Graham, J.W. (2003) Adding missing-data-relevant variables to FIML-based structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, *10*, 80–100.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 120–137.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York: Wiley.

Hogan, J. W., & Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, *16*, 239–258.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 221–233). Berkeley: University of California Press.

Hunt, J. R., & White, E. (1997). Retaining and tracking cohort study members. *Epidemiologic Reviews*, *20*, 57–70.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in Medicine*, *17*, 2723–2732.

Kenward, M.G., Molenberghs, G. & Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika*, *90*, 53–71.

Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, *67*, 609–623.

Kim, S. Y., Mun, E. Y., & Smith, S. (2014). Using mixture models with known class membership to address incomplete covariance structures in multiple-group growth models. *British Journal of Mathematical and Statistical Psychology*, *67*, 94–116.

Kromrey, J. D., & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Educational and Psychological Measurement*, *54*, 573–593.

Lee, S. Y., & Song, X. Y. (2004). Bayesian model comparison of nonlinear latent variable models with missing continuous and ordinal categorical data. *British Journal of Mathematical and Statistical Psychology, 57*, 131–150.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202.

Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125–134.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, *81*, 471–483.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112–1121.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, *10*, 538–573.

Mirowsky, J., & Reynolds, J. R. (2000). Age, depression, and attrition in the National Survey of Families and Households. *Sociological Methods & Research, 28*, 476–504.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431–462.

Molenberghs, G., Michiels, B., Kenward, M. G. & Diggle, P. J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, *52*, 153–161.

Morrison, T. C., Wahlgren, D. R., Hovell, M. F., Zakarian, J., Burkham-Kreitner, S., Hofstetter, C. R., … Jones, J. A. (1997). Tracking and follow-up of 16,915 adolescents: minimizing attrition bias. *Controlled Clinical Trials*, *18*, 383–396.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *163*, 445–459.

Murray, G. D., & Findlay, J. G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, *7*, 941–946.

Muthén, B., Asparouhov, T., Hunter, A. M., & Leuchter, A. F. (2011). Growth modeling with nonignorable dropout: alternative analyses of the STAR* D antidepressant trial. *Psychological methods*, *16*, 17–33.

Muthén, B., & Jöreskog, K. G. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review*, *7*, 139–174.
Nesselroade, J. R., & Baltes, P. B. (1979). *Longitudinal research in the study of behavior and development*. New York: Academic Press.
Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*, 328–362.
Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*, 966–980.
Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, *96*, 730–745.
O'muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: a latent variable approach to item non-response in attitude scales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*, 177–194.
Raykov, T., Lichtenberg, P. A., & Paulson, D. (2012). Examining the missing completely at random mechanism in incomplete data sets: A multiple testing approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 399–408.
Ribisl, K. M., Watlon, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. A., & Bootsmiller, B. (1996). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning*, *19*, 1–25.
Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, *59*, 829–836.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489.
Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 477–497.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Boca Ratan, FL: Chapman & Hall/CRC Press.
Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*, 147–177.
Sullivan, C. M., Rumptz, M. H., Campbell, R., Eby, K. K., & Davidson, W. S. (1996). Retaining participants in longitudinal community research: A comprehensive protocol. *The Journal of Applied Behavioral Science*, *32*, 262–276.
Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, *3*, 245–265.
Tsonaka, R., Verbeke, G., & Lesaffre, E. (2009). A semi-parametric shared-parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, *65*, 81–87.
Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., & Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics*, *57*, 7–14.
White, H. L., Jr. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25.
Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, *44*, 175–188.
Young, A. F., Powers, J. R., & Bell, S. L. (2006). Attrition in longitudinal studies: who do you lose? *Australian and New Zealand Journal of Public Health*, *30*, 353–361.
Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200.
Yuan, Y., & Little, R. J. (2009). Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, *65*, 478–486.

Keywords
missing data, attrition, full information maximum likelihood, missing completely at random, missing at random, missing not at random, nonignorable

Appendix A Notation

The structural modeling notation system I use for this book follows a traditional system that uses Greek symbols for each of the parameters based on a matrix organizing principle, usually referred as LISREL notation, short for "LInear Structural RELations" (Frisch & Waugh, 1933). The general SEM model and notation system is perhaps more accurately called the JKW model, after the authors credited with synthesizing and expanding decades of prior work on path analysis and factor analysis into a highly generalizable structure equation framework (Jöreskog, 1973; Keesling, 1972; Wiley, 1973). The LISREL term has become primarily associated with the software developed by Jöreskog and Sörbom (1974), but the notation system has become widely applied regardless of the SEM software package used.[1] I use LISREL notation throughout the book for one very important reason: a large majority of statistical articles about SEM use this notation. Many introductory textbooks now avoid LISREL notation in order to increase accessibility, which is indeed an objective I sympathize with. For those who wish to learn more about SEM after an initial introduction, however, unfamiliarity with the LISREL notation system leaves readers with what I believe to be a serious literacy gap.

Although LISREL notation is tied to matrix algebra, it is really not necessary to know matrix algebra to read and understand this book. Matrix algebra, a kind of shorthand system that can be used for manipulating many simultaneous equations, is convenient for describing the linear regression equations used in SEM. Learning the Greek symbols associated with the LISREL notation is a separate matter from understanding matrix algebra and is, at least, an initial step. I do encourage the reader to learn matrix algebra to increase the understanding of this topic and to better understand some of the mathematical underpinnings of SEM. There are a fairly limited number of definitions and simple algebra rules that can be absorbed with a small investment in effort. I do not provide an introduction to matrix algebra with this text, because there are many excellent introductions (e.g., Bollen, 1989; Hayduk, 1987; Mulaik, 2009; Namboodiri, 1984).

"All-y" LISREL Notation

Most of the formulas in this book use an abbreviated version of the full LISREL notation that is commonly used by authors and easier to learn. The full LISREL notation system (described later in this appendix) distinguishes between exogenous variables and endogenous variables. Exogenous variable are not caused by other variables in the model and endogenous variables are those caused by other variables in the model.

Measurement Model Parameters

Table A.1 is a summary of all of the Greek symbols used in the LISREL model. Each latent variable is designated by η ("eta"). In this text, I will index latent variables with subscript

Table A.1 All-y LISREL Notation

Individual parameter	*English spelling*	*Parameter matrix*	*Description*
λ	Lambda	Λ	Factor loadings
ψ	Psi	Ψ	Variances and covariances of latent variables
β	Beta	**B**	Causal paths
θ	Theta	Θ	Measurement residual variances
ε	Epsilon	ε	Measurement residuals, variances are elements in the theta matrix, $\text{Var}(\varepsilon)=\theta$
η	Eta	η	Latent variables
ζ	Zeta	ζ	Structural disturbances
α	Alpha	α	Latent variable means
ν	Nu	ν	Measurement intercepts

k, up to a total of K latent variables in the model. Observed variables will be indexed with j, and there are J total observed variables in model. Loadings, λ_{jk} ("lambda"), represent a regression of a measured variable y_j on factor η_k, using subscript jk to indicate the jth observed variable is predicted by the kth latent variable. The "effect" always precedes the "cause" in the order of subscripts for loadings (and structural paths). The intercept in this regression is ν ("nu") with subscript j. We can then write an equation for a simple regression that represents the relation of the observed variable to the factor.

$$y_j = \nu_j + \lambda_{jk}\eta_k + \varepsilon_j$$

We could add an index i representing an individual case in the data set for the observed variable y_{ji}, the latent variable η_{ki}, and the measurement residual (or error term) ε_{ij}, but I omit this in most instances to simplify the notation as much as possible.

The individual parameters are organized into matrices or vectors, matrices with one column (or, if transposed, a row). Capital letters (bolded in this text) are used to represent each matrix. Loadings are organized into a matrix with J rows and K columns, said to be of dimension $J \times K$. Each entry in the loading matrix represents the intersection of an observed variable and a factor. For example, a model with two latent variables with three indicators loading on each factor would be a 6×2 $\mathbf{\Lambda}$ matrix, with rows corresponding to observed variables y_1, y_2, y_3, y_4, y_5, and y_6, and columns corresponding to latent variables η_1 and η_2,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}$$

The 0s show that the indicator loading is set equal to 0, or, in other words, does not load on that factor.

The measurement residuals are organized into a square matrix $\mathbf{\Theta}$ ("theta") with the diagonal elements representing the variances $\text{Var}(\varepsilon_j)=\theta_{jj}$, and the off-diagonal elements representing covariances among measurement residuals (e.g., $\text{Cov}(\varepsilon_1,\varepsilon_2)=\theta_{12}$). The

two-factor example mentioned above might have a 6×6 $\boldsymbol{\Theta}$ matrix that looks like the following, if each of the measurement residuals were freely estimated and one covariance between y_1 and y_2 was estimated.

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} & 0 & 0 & 0 & 0 \\ \theta_{21} & \theta_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix}$$

The observed variables and the latent variables are vectors (single-column matrix) with J and K rows, respectively. The factor variances and covariances have the symbol ψ ("psi") and appear in the $\boldsymbol{\Psi}$ matrix, which is square with dimension $K \times K$. The diagonal elements are the variances and the off-diagonal elements are the covariances. If the two-factor example estimated both factor variances and the covariance, the $\boldsymbol{\Psi}$ matrix would be

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}$$

The measurement model states the covariance matrix in terms of these matrices:

$$\boldsymbol{\Sigma}(\theta) = \boldsymbol{\Lambda}\boldsymbol{\psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

The prime symbol $'$ indicates the Λ matrix is transposed (rows and columns are switched). Factor means, α_k ("alpha") and measurement intercepts, ν_j, are not included in the measurement equation above, but they can be added to the model. Each is a vector of the same name in the matrix system, $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$, respectively.

Structural Model Parameters

The structural portion of the model involves paths between latent variables, represented by β ("beta"). Although β is used to represent a standardized coefficient in some regression texts, it represents a unstandardized coefficient here. I will use β^* for a standardized coefficient instead. The order of the subscripts is such that the dependent, or "effect", variable precedes the predictor or "cause" variable. For example, a path for η_2 predicted by η_1 would be labeled β_{21}. The path coefficients are organized into the $\mathbf{B}$ matrix, with the dimensions $K \times K$. Naturally, many of the elements will be 0 in practice, because usually only one direction can be estimated in practice. Disturbances (residuals, errors) in the structural model are represented by ζ ("zeta"). Disturbances may appear as a vector of individual parameters ζ. Because dependent variables have only conditional variances, the variances of the disturbances are diagonal elements in the Ψ matrix, where $\text{Var}(\zeta) = \psi$. Covariances of disturbances, $\text{Cov}(\zeta,\zeta)$, are off-diagonal elements in the Ψ matrix.

The formal LISREL notation system assumes only structural relations among latent variables not between observed variables or between latent variables and observed variables. Each observed variable must be an indicator of a latent variable, even if there is only one indicator per latent variable (identified by setting the loading equal to 1 and the measurement residual equal to 0). Most SEM software packages, however, allow structural paths between measured variables and measured and latent variables. As a consequence, many articles and texts allow structural paths directly between observed variables and

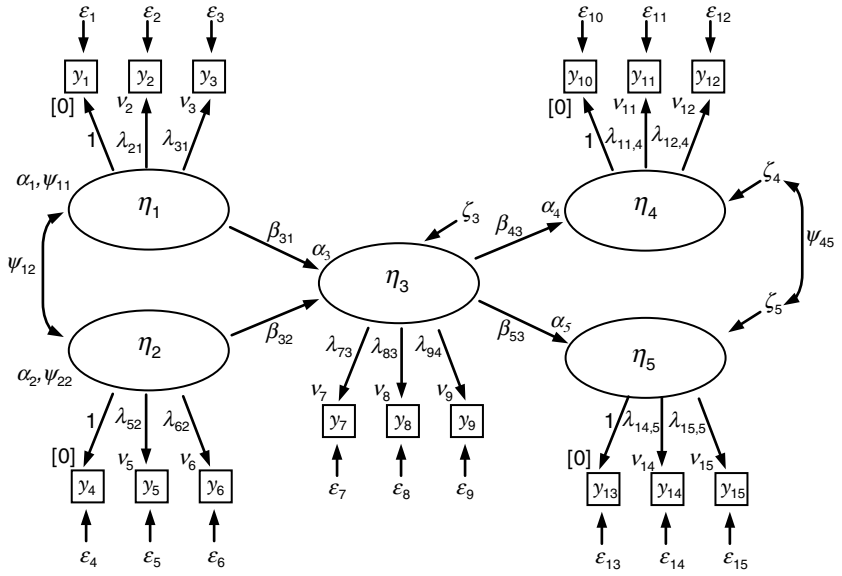

Figure A.1 All-y LISREL Notation.

latent variables. This convenience has no impact on the underlying mathematics, however. I therefore use *x* and *y* in equations when there are structural relations among them (predictive paths or correlations) and show them within squares with direct relations to latent or other observed variables in figures.

Path Diagrams

Figure A.1 summarizes the notation in the depiction of one possible model. Notice that when variable numbers have two digits, a comma is used to separate subscript number pairs (e.g., $\lambda_{15,4}$). I follow most of the usual path diagram conventions for structural models. One exception is that I do not represent means and intercepts as triangles in the diagrams as in the RAM diagram approach (McArdle & McDonald, 1984), primarily to simplify the diagrams of some of the rather complex models in some chapters. Instead, when means or intercepts are estimated in the model, I depict this by placing the symbol next to the ellipse (latent means and intercepts) or rectangle (measurement intercept).[2] For variances, ψ_{kk} appears next to the circle or rectangle in a similar fashion.

When a parameter is to be set to a specific value, such as 0 or 1, the number appears in the diagram in the location in which the parameter normally appears. Mean or intercept values are in square brackets to distinguish them. Figure A.1 illustrates the use of specific values, where the first loading is set equal to 1 and the measurement intercept is set equal to 0, shown as [0]. These are commonly used values for the referent or marker method of identifying the factor variance and mean.

Full Matrix Notation

The "all-y" notation is commonly used by authors, but its use is not universal. The original and more formal LISREL system involves separate matrices for exogenous and

Table A.2 Full LISREL Notation

Exogenous parameter	*English spelling*	*Exogenous matrix*	*Description*	*Endogenous parameter*	*English spelling*	*Endogenous matrix*	*Description*
λ_x	Lambda-x	Λ_x	Factor loadings for loadings on exogenous latent variables	λ_y	Lambda-x	Λ_y	Factor loadings, λ_x for loadings on exogenous latent variables, y for indicators on endogenous latent variables
ϕ	Phi	Φ	Variances and covariances of exogenous latent variables, Var(ξ) and Cov(ξ,ξ)	ψ	Psy	Ψ	Disturbance variances and covariances among disturbances, Var(ζ) and Cov(ζ,ζ)
γ	Gamma	Γ	Causal paths, endogenous predicted by exogenous	β	Beta	B	Causal paths, endogenous predicted by endogenous
θ_δ	Theta-delta	Θ_δ	Measurement residual variances for x variables	θ_ε	Theta-epsilon	Θ_ε	Measurement residual variances for y variables
δ	Delta	δ	Measurement residual, variances are elements of theta-delta matrix, Var(δ)=θ_δ	ε	Epsilon	ε	Measurement residuals, variances are elements in the theta matrix, Var(ε)=θ
ξ	Ksi	ξ	Exogenous latent variables	η	Eta	η	Endogenous latent variables
				ζ	Zeta	ζ	Structural disturbances
κ	Kappa	κ	Exogenous latent variable mean	α	Alpha	α	Endogenous latent variable mean
ν_x	Nu-x	ν_x	Measurement intercepts for x variables	ν_y	Nu-y	ν_y	Measurement intercepts for y variables

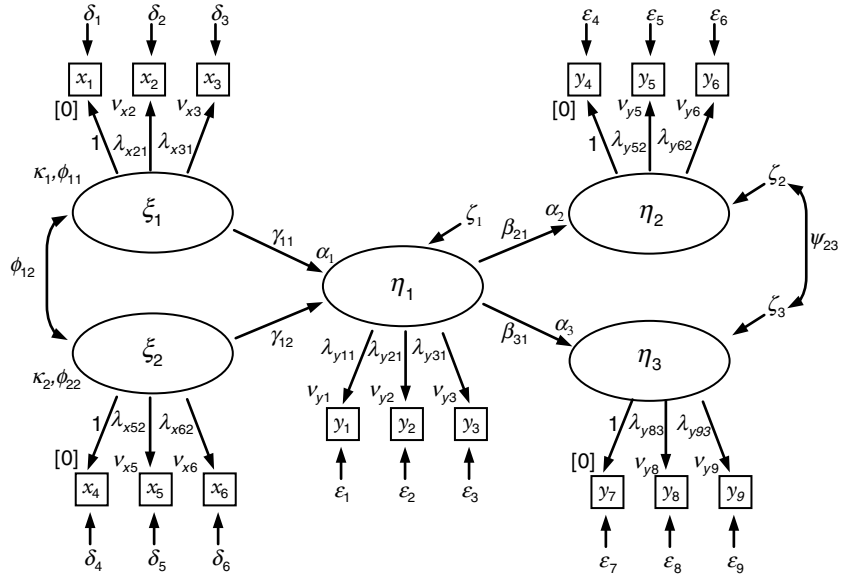

Figure A.2 Full LISREL Notation.

endogenous parameters. *Exogenous* variables are those not predicted by other variables in the model and *endogenous* variables are those predicted by other variables. This distinction usually can be dropped without loss of generality, but the full notation is needed for clarity in some instances. To simplify as much as possible, I use the "all-y" system whenever possible.

Table A.2 summarizes all the symbols and matrices used for the full LISREL notation. Each parameter has a separate notation for the parameter and accompanying matrix depending on the role of the variable as exogenous or endogenous in the model. Even when observed variables, x, are used as indicators of exogenous variables, and are therefore predicted by another variable, they are still considered exogenous in the model under the notation system. The symbols used for the endogenous parameters, sometimes with added y subscript, are the same as in the "all-y" system, but exogenous parameters and matrices use the following symbols: ϕ and $\mathbf{\Phi}$ for latent variable variances and covariances, λ_x and $\mathbf{\Lambda}_x$ for loadings, θ_δ and $\mathbf{\Theta}_\delta$ for measurement residual variances and covariances, κ and $\mathbf{K}$ for latent variable means, and ν_x and $\boldsymbol{\nu}_x$ for measurement intercepts. Figure A.2 depicts the full LISREL notation version of the same model shown in Figure A.1.

Other Notation Details Specific to this Text

My preference is to avoid subscripts wherever possible. When the indexing is obvious or not necessary, I omit subscripts. For instance, I refer to an observed variable as y instead of y_i, omitting the subscript because it can be assumed that a variable varies across individual cases in the data set unless otherwise indicated. Admittedly, there is imprecision in doing this and potential confusion, but I believe the benefits of simplicity outweigh the costs.

The following subscripts are used whenever indexing is needed: i for individual case, j for the jth observed variable, k for the kth factor, and t for the time point. For multiple

groups, *g* is used to designate a group-specific value. Although this does not ever occur in the text, a full example would be to refer to an observed variable as y_{tijkg}, for an observed score *y* at time point *t*, for case *i,* on observed variable *j,* loading on factor *k,* in group *g*. Where any of these are understood or not necessary, they will be omitted. For example, where there is only one observed variable or one factor and the context is clear, I will omit the *j* subscript for a particular observed variable, and/or the *k* subscript referring to a particular factor. If the data are cross-sectional, I will omit the *t* subscript. Likewise, for most formulas, I will omit subscripts from Greek LISREL matrix symbols when the all-y format is used or when the reference is easily understood from the context (e.g., Λ_y will be simply Λ, and λ_y will be simply λ). To refer to a parameter being held constant across time points, such as a survey question that has been repeated over time, I will use a subscript in parentheses, as in $_{(j)}$ or $_{(1)}$.

For any references to regression analysis or logistic regression analysis, I use β instead of *b* to refer to unstandardized coefficients. I do this in part to reduce the number of different symbols used overall but also to emphasize the equivalence of regression coefficients and path coefficients from a structural equation model. Primes,′, are used to distinguish between estimates obtained with referent and factor identification (mainly in Chapter 1), and should not be confused with the transpose operation that is used in matrix algebra (which appears in a few places in the text as well, but only next to a matrix). To denote an indicator that is not equal to *j*, the notation j° is used occasionally.

Notes

1 The LISREL notation system is neither universal nor necessary. There are many minor variations and several major alternative notation systems. Most notably, the Bentler–Weeks system (Bentler & Weeks, 1980), which is associated with EQS software, is another matrix-based notation alternative.

2 I borrowed this convention from my colleague, Rich Jones.

References

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika, 45*, 289–308.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, *1*, 387–401.

Hayduk, L. A. (1987). *Structural equations modeling with lisrel*. Baltimore, MD: Johns Hopkins University Press.

Jöreskog, K. G. (1973). Analysis of covariance structures. *Multivariate Analysis*, *3*, 263–285.

Jöreskog, K.G. & Sörbom, D. (1974). *LISREL III* [Computer software]. Chicago, IL: Scientific Software International, Inc.

Keesling, J. W. (1972). *Maximum likelihood approaches to causal analysis* (PhD. thesis). University of Chicago.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37*, 234–251.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. New York: CRC Press.

Namboodiri, N. (1984). *Matrix algebra: An introduction*. Beverly Hills, CA: Sage.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 69–83). New York: Seminar.

Appendix B A Primer on the Calculus of Change

Many readers may not have a background in calculus, so I offer this brief primer for a slightly different perspective on describing change over time. The focus here is on one of the major topics in calculus, derivatives, which can be used for describing continuous curves. Although it is not necessary to know anything about calculus to learn from this text or to have a solid understanding of the analysis of change, some additional insights about continuous change over time can be gained by a small investment in learning some simple concepts associated with derivatives.

Definition of a Derivative

The basic building blocks for the ideas behind derivatives can first be expressed in terms of a slope. One way to conceptualize a regression slope is as a ratio of change in y over the change in x. Remember "rise over run" as one definition of the slope? Derivatives are based on the notion of continuous change in y relative to the continuous change in x. The linear slope can be described in terms of the first derivative, where the slope is a ratio of the change in y between two points relative to the change in x between two points, $\beta_1 = (y_2 - y_1)/(x_2 - x_1) = \Delta y / \Delta x$. The derivative, however, estimates this ratio for infinitesimally small increments. The infinitesimally small increments are often described in terms of a limit with notation $\lim_{x \to 0} f(x)$. An equation, such as that for a line where y is a function of x, is denoted $f(x)$.

The process of calculating the derivative from an equation involves reducing the exponent and pre-multiplying the term by the exponent. If we take a simple equation with one term written as $f(x) = cx^n$, where x is a variable, c is any coefficient value, and n is an exponent, then the steps to obtain the derivative are summarized by the following notation:

$$f'(x) = ncx^{n-1}$$

The exponent in the original equation is reduced by one unit, $n - 1$, and the coefficient c is multipled by n, the original exponent value. This process is done for each term in an equation. As a slightly more complicated example, consider the quadratic equation $y = 4 + 2x + 3x^2$. According to the rule for constants, 4 is eliminated (a constant cannot change). For the remaining terms, the derivative is calculated by subtracting 1 from each exponent and multiplying its term by the old exponent. Thus, the second term is modified by subtracting 1 from the exponent implied in the original term, so x^1 is modified to be $x^{1-1} = x^0 = 1$. The original term $2x$ is modified to $2(1) = 2$. The third term in the equation is modified by multiplying the coefficient 3 by the exponent 2 and then reducing the

exponent for the variable to 1, yielding the term $6x^1$. We can then drop that exponent, because x^1 is the same as x, resulting in the derivative $f'(x) = 2 + 6x$

A more statistical application may be more helpful. By taking the derivative of the linear regression line, $\hat{y} = \beta_0 + \beta_1 x_1$, we find the slope, β_1. Because β_0 is a constant, it is removed according the derivative rule. β_1 is the coefficient for x, and when x is modified, $x^{1-1} = x^0 = 1$, this term becomes simply $\beta_1(1) = \beta_1$. Although it is obvious to us that β_1 is the slope in this case, given what we know about straight lines, the derivative computation can be used for much more complex functions for curves, yielding important information from the equation.

Functions of Time

If we restate the derivative concept in terms of change over time, then the slope represents the change in y per unit of time, t, and the derivative would be notated as

$$\lim_{\Delta t \to 0} \frac{\Delta y}{\Delta t} = \frac{dy}{dt} = f'(t)$$

which states that the first derivative $f'(t)$ of some function $f(t)$ gives the slope or rate of change in y per change in t at infinitesimally small increments. The function $f(t)$ can be any curve as long as it is continuous and smooth enough to be defined. For a straight line, the slope, of course, gives the change in y per unit change in t. For a curved line, the first derivative gives information about rate of change at any given point along the curve. This is sometimes described as the instantaneous rate of change. In physics, the derivative gives the instantaneous velocity. At any point along the road, how fast are you traveling?

For the quadratic equation of change in the form $f(t) = y = \beta_0 + \beta_1 t + \beta_2 t^2$, the derivative is $f'(t) = \beta_1 t^{1-1} + 2\beta_2 t^{2-1} = \beta_1 + 2\beta_2 t$. Let us use a more concrete example. Say we have estimated a nonlinear predicted line for some outcome variable (e.g., body mass index), and we obtain the following values: $\beta_0 = 5$, $\beta_1 = 2$, and $\beta_2 = 3$. Our function with these values thus looks like $f(t) = 5 + 2t + 3t^2$. The first derivative is then calculated to be

$$f'(t) = 0 + 2t^{1-1=0} + 2 \cdot 3t^{2-1=1}$$
$$f'(t) = 2 + 6t$$

The derivative is useful in this case because it gives the slope of the tangent line. In general, a tangent line approaches the curve at only one point at an infinitesimally small distance. It is considered to be "parallel" to the curve at that point. In the case of a quadratic equation, the tangent line tells us what the linear rate of change is at some value of t, or sometimes said to be the instantaneous rate of change. Figure B.1 is a plot of the quadratic equation and a tangent line.

If we were to draw tangent lines along different points of a curve, our tangent line would have a different slope depending on where we chose our point of tangent. So the slope varies, having the value of the derivative function for any selected value of t, which I will call t_a. If we select $t_a = 4$, then our tangent line slope will be $2 + 6(4) = 26$. We then have the slope, but we do not yet know what the intercept is, so the line cannot be drawn without some additional information. In this case, the point-slope equation will give us the information we need to plot the tangent line provided we have the slope and the coordinates of a point on the original curve. The point-slope formula, $y = \beta_{1t}(t - t_a) + y_a$ can be used to obtain values of y for a set of values of t inserted into the equation if we have the coordinates for just one of the points on the original curve (t_a, y_a) and the slope of the line, β_1.

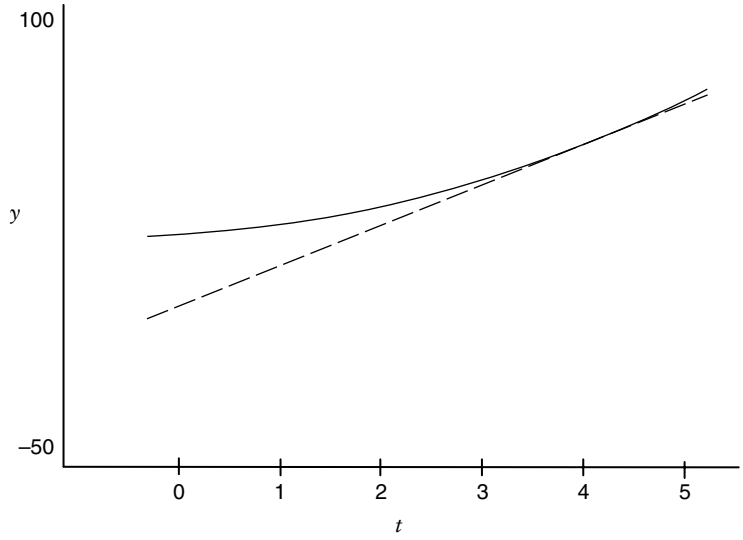

Figure B.1 Quadratic Curve (Solid Line) with Tangent (Dashed Line).

To find the tangent line at a certain point along the curve, there are several steps necessary:

(1) Insert the desired value of t, called t_a, into the first derivative equation, $f'(t)$, to find the slope of the tangent line at a certain value on the x-axis.
(2) Find the value of y_a at t_a using the original quadratic equation. This provides the necessary vertical adjustment.
(3) Insert both x_a and y_a into the point-slope formula, $y = \beta_{1t}(t - t_a) + y_a$ for values of t to plot the tangent line.

If the desired point for the tangent line is 4 (i.e., $t_a = 4$), then we first need to find the slope of the tangent line.

$$\begin{aligned}\beta_{1t} &= 2 + 6t_a \\ &= 2 + 6(4) \\ &= 26\end{aligned}$$

Then, calculate the y_a coordinate

$$\begin{aligned}y_0 &= 5 + 2t_a + 3t_a^2 \\ &= 5 + 2(4) + 3(4)^2 \\ &= 61\end{aligned}$$

Finally, compute a set of y values for the tangent line using a set of t values (e.g., $t = 0, 1, 2, 3, 4, 5$) with the point-slope formula, $y = 26(t - 4) + 61$ so that the tangent can be plotted. This set of values is used to plot the tangent line as shown in Figure B.1.

Notice that if we enter the value 0 for t_a, then it is evident that the slope of the tangent line is equal to the constant from the derivative equation, $2 + 6t = 2 + 6(0) = 2$. In other words, if the point of contact of the tangent line is at the first time point, the slope

of the tangent line will be equal to the constant from the derivative equation. This makes sense because the constant in the derivative equation was the linear coefficient in the quadratic equation. This suggests that the linear slope in the quadratic equation gives the instantaneous rate of change when $t = 0$. The tangent line can provide other useful information about a curve. When the slope of the tangent line is equal to 0 (the slope is flat), then it meets the curve at the minimum or maximum point, depending on whether the quadratic curve is positive (minimum) or negative (positive). So, the first derivative can be set equal to 0. Solving then gives the value of t for the lowest point occurring on the curve, $2 + 6t = 0$, $t = -2 / 6 = -.333$.

Second Derivative

The second derivative is the derivative of the first derivative. In terms of discrete points, it is the difference between differences, $\left(\Delta y_{2-1} / \Delta t_{2-1}\right) - \left(\Delta y_{3-2} / \Delta t_{3-2}\right)$. This is again a simplified form, because the second derivative is more precisely stated in terms of limits

$$\lim_{\Delta t \to 0} \frac{\Delta y_{3-2} / \Delta t_{3-2}}{\Delta y_{2-1} / \Delta t_{2-1}}$$

In more conceptual terms, we are finding the differences in the rate of change, or, in other words, the rate of change of the rate of change. If describing motion, the first derivative gives the instantaneous velocity, and the second derivative gives the instantaneous acceleration. At any point along the road, how much is your speed increasing? In terms of change over time, the second derivative gives information about the acceleration or deceleration of change. For a trajectory of weight gain over time, acceleration represents the degree to which someone's rate of weight gain has slowed down or sped up.

The notation for the second derivative is not universal and may be expressed in any of several ways.

$$\frac{d^2 y}{dt^2} = \frac{df'(t)}{dt} = f''(t)$$

For our quadratic curve example, $y = 5 + 2t + 3t^2$, the second derivative can be found by taking the derivative of the first derivative.

$$f'(t) = 2 + 6t$$
$$f''(t) = 6$$

Glossary

Chapter 1

Measurement residual. Error term from the measurement model, ε_j. Unaccounted for variance after the factor variance is taken into account, $\text{Var}(\varepsilon_j) = \theta_{jj}$.

Factor variance identification. Scaling constraint imposed to define the latent variable variance by setting the variance of the factor equal to 1.

Referent loading identification. Scaling constraint imposed to define the latent variable by setting one factor loading equal to 1.

Measurement bias. Systematic variance associated with a social group that increases or decreases the mean of a measured variable.

Measurement error. Random error associated with a measured variable. Lack of reliability.

Systematic variation. Any non-random variation unique to a latent variable indicator (measured variable). Same as "unique variance" or "unique factors."

Mean structures. Structural equation model estimates of factor means and measurement intercepts.

Referent intercept identification. Constraint imposed to define the latent variable mean by setting the measurement intercept to 0.

Factor mean identification. Constraint imposed to define the latent variable mean by setting the factor mean to 0.

Effects coding identification. Constraint on loadings or measurement intercepts such that the average equals 1 or 0, respectively. Avoids setting the factor variance or factor mean to an arbitrary value or to a value tied to just one indicator.

Link function. Mathematical function used with discrete variables that connects the linear regression model to the hypothetical, unobserved continuous variable y^*.

Logit. Link function used with logistic regression. The natural log of the observed probabilities, $\ln[p/(1-p)]$.

Cumulative distribution function (cdf). Cumulative form of a statistical probability distribution that can be used to derive the probability of occurrence equal to or less than a certain value.

Probit regression. Regression model for binary or ordinal variables that assumes the residual variance distribution is normal.

Delta parameterization. Identification method used to define the y^* distribution with the measurement residual variance constrained but the y^* residual variance freely estimated.

Theta parameterization. Identification method used to define the y^* distribution with the y^* residual variance constrained but the measurement residual variance freely estimated.

Chapter 2

Measurement invariance. General term referring to the equality of measurement model parameters over time, including factor variance, loadings, factor mean, measurement intercepts, or measurement residuals.

Configural invariance. Identical factor structure over time, including the same indicators for each factor.

Weak factorial invariance. Loadings invariant over time, with all other measurement parameters free to vary.

Strong factorial invariance. Loadings and measurement intercepts invariant over time, with all other measurement parameters free to vary.

Strict factorial invariance. Loadings, measurement intercepts, and measurement residual variances invariant over time, with all other measurement parameters free to vary.

Structural factorial invariance. All measurement parameters invariant: loadings, measurement intercepts, measurement residuals, factor variance, and factor means.

Partial measurement invariance. Measurement parameters for some indicators but not others are invariant. May refer to any parameter, including loadings, measurement intercepts, or measurement residuals.

Standardization problem. Refers to how the referent chosen for identifying the factor variance or factor mean may obscure partial invariance tests, because factor variance and loading estimates are interdependent.

Single occasion identification approach. Scaling approach to factor variance of factor mean by setting the variance and mean on one time point but allowing it to be freely estimated at other time points. Only feasible when used in conjunction with loading and intercept equality constraints.

Factor-ratio tests. Approach to partial invariance tests used to identify the correct source of invariance. Entails likelihood ratio tests of the equivalence of the ratios of pairs of loadings or pairs intercepts.

Chapter 3

Time-structured data. Data collected at regular intervals consistent across cases.

Single variable difference score model. Test of comparison of two means with a pre-computed difference score.

Equality constraint model. Mean structure model with correlated factors used in conjunction with nested test to compare means.

Contrast coding model. Model with intercept and factor used to compare two or more means.

Latent difference score model. Model with predictive paths equal to 1 and latent variable that reflect individual differences. Can be used to compare two or more means.

Marginal homogeneity hypothesis. Test of marginal or overall proportions, such as the comparison of the number of yes responses at Time 1 and Time 2.

Axial symmetry hypothesis. Compares proportions of same responses over time when there are more than two categories. For example, yes–yes, no–no, and don't know–don't know.

Chapter 4

Absolute level stability. Lack of change in an average or level of an individual score. A score is stable to the extent that the value is the same on each occasion.

Relative level stability. Stability defined in terms of the correlation of scores over time. High correlations indicate greater stability, regardless of the absolute level of the scores.

Autoregression. Regression equation predicting measurement at a later time point from an earlier time point, where y_t is regressed on y_{t-1}.

Autoregressive effect. Path coefficient (slope) from the autoregression.

Stability path. Path coefficient (slope) from the autoregression.

Inertia. Path coefficient (slope) from the autoregression.

Regression toward the mean. Tendency for scores to vary stochastically over time when the autocorrelation is not perfect. Implies that extreme scores will move toward the mean and scores near the mean will move toward the extremes over time.

Entropic decay. An alternative term for regression toward the mean. Opposite of stability over time. Borrowed from the physics concept referring to the tendency toward greater disorganization.

Yule's Q. A standardized measure of stability used with binary data representing the proportion of cases that have the same values over time.

Lagged regression. Estimates the effect of a hypothesized cause measured at an earlier time point on a hypothesized effect measured at a later time point.

Cross-lagged effect. Path from a predictor variable measured at an earlier time point, usually including the simultaneously measured outcome variable as a covariate.

Lord's paradox. Refers to conflicting conclusions that can be implied by the unconditional difference score model and the conditional cross-lagged model.

Chapter 5

Panel study. Study with repeated measurements. Typically a time structured design with the same regular intervals for all cases and involving no experimental or quasi-experimental manipulation.

Granger causality. Investigating causal directionality – whether x causes y or y causes x.

Cross-lagged panel model. Structural equation or path model that includes autoregression and cross-lagged paths for two variables over two or more time points.

Non-recursive effects. Bidirecational paths, usually referring to cross-sectional relationships. The term "recursive" refers to directional, so non-recursive refers to bidirectional.

Time-varying covariate. Predictor measured at each occasion.

Fixed effects regression model. Regression model of difference scores, with y_{2-1} regressed on x_{2-1}.

Wiener model. Autoregressive model with paths only between each adjacent time point for any number of repeated measures. Synonymous with the simplex model and random walk model.

Random walk. Autoregressive model with paths only between each adjacent time point for any number of repeated measures. Synonymous with the simplex model and the Wiener model.

Simplex model. Correlational structure of repeated measurements, implying a particular mathematical structure. A simplex model is an autoregressive model for repeated measures, with autoregressive paths only between each adjacent time point.

Quasi-simplex model. Simple model of observed variables that attempts to remove measurement error via estimation of measurement residuals with single-indicator latent variables. Also known as the quasi-Markov simplex model.

Markov chain process. Term generally used to refer to simplex models with binary variables, though it can be used more generally for simplex-related models.

First-order autoregression. Refers to autoregression between adjacent time points (y_t regressed on y_{t-1}). Also lag 1 autoregression.

Stationarity. Involves equality or invariance in relationships among variables over time, usually operationalized as constant parameter estimates over time.

Temporal inertia. Tendency for a variable to remain stable over time. Used in the relative level sense implying high autocorrelation.

Equilibrium. Steady state of a system reached once variables cease to change over time.

Negative feedback. The relationship between an initial time point and the subsequent time point when entropy or regression toward the mean occur. Can be used to determine the speed at which the system will reach equilibrium.

Chapter 6

Latent state-trait model. Structural model that partitions stable and unstable systematic variance.

Common consistency coefficient. Estimate of the proportion of variance due to stable factors.

Occasion-specific coefficient. Estimate of the proportion of variance due to state factors.

Method specificity coefficient. Estimate of the proportion of variance due to method factors.

Target coefficient. Ratio of chi-square values used to evaluate whether a set of constraints, such as adding a higher-order factor, lead to a decrement in fit that is of a large magnitude.

Residual method factor approach. Creates one fewer method factors than loadings, as one loading and its corresponding method serve as the comparison or referent.

Chapter 7

Intraindividual variation. Variation in scores across time within an individual case.

Interindividual variation. Variation across cases. May refer to variation in intercept or slope coefficients.

Fixed effects. Estimates of the average intercept or slope value.

Random effects. Estimates of the variance of intercepts of slopes across cases.

Intraclass correlation coefficient (ICC). An index of the similarity of scores within an individual (or a group). Greater similarity within individuals implies greater heterogeneity between individuals, so the ICC is defined as the proportion of total variation that is interindividual variation.

Conditional intraclass correlation coefficient. The intraclass correlation for intercepts when a covariate is included in the model.

Reliability coefficient. An estimate of the stability of the intercept or slope.

Occasion-specific nonsystematic variance. In a growth curve model, variance in observed variables that is not explainable by other variables.

Occasion-specific systematic variance. In a growth curve model, variance in observed variables that is explainable by other variables omitted from the model.

Heterogeneity of variance. In growth curve modeling, term that refers to allowing different estimates of occasion-specific residual variance over time.

Homogeneity of variance. In growth curve modeling, a term that refers to constraining estimates of occasion-specific residual variance to be equal over time.

Cross-level interaction. The interaction implied when a time-invariant covariate is a predictor of the slope variable. In multilevel analysis, the time-invariant covariate is a

level-2 variable, because it is measured at the case level, whereas time is a level-1 variable because it is estimated for each time point within a case.

Simple trajectory. Simple slope for a growth curve model – the expected value of the slope factor η_1 at a particular value of the covariate x.

Time-varying covariates. Predictors measured at each occasion.

Variable definition method. Using individually-varying time scores in growth curve analysis.

Cohort sequential design. Longitudinal design of different groups, usually a time-related grouping such as age, observed over the same intervals but starting at different points in time. Also known as accelerated design.

Second-order growth curve models. Latent growth curve models using latent variables with multiple indicators at each occasion.

Autoregressive latent trajectory model. Latent growth curve model with an autoregressive component.

Chapter 8

Latent piecewise growth curve model. Method of modeling discontinuous change in separate segments of a longitudinal study.

Increment (decrement) model. A piecewise growth model that estimates a base rate slope across the full length of the study and a slope representing a relative change in rate after the event. Same as the added growth model.

Instantaneous rate of change. Estimate of the linear effect in the quadratic model, representing the rate of change as the increment in time score approaches 0.

Orthogonal polynomials. Time codes used in testing nonlinear models, constructed so that linear, quadratic, or higher-order effects are independent of one another.

Local linear approximation. Approximation of the first or second derivative through difference scores.

Amplifying effect. When a linear effect has a positive association with a nonlinear effect.

Damping effect. When a linear effect has a negative association with a nonlinear effect in a dynamic systems model.

Latent basis model. Growth curve model fitting the form of the data by freely estimating loadings for the growth factor.

Chapter 9

Self-feedback effect. Term for a lagged effect added to the latent difference score model.

Dual change score. Latent difference score model with self-feedback loop combined with slope and intercept factors.

Coupling effect. Lagged pathways between difference score factors in the simultaneous latent difference score model.

Bivariate dual change score. Latent difference score model involving intercept and slope factors, dual change components, lagged effects, and covariates.

Chapter 10

Local independence. Assumption in latent class analysis that there is no residual association among indicators. Same as conditional independence.

Response probabilities. The conditional probability that the indicator will be equal to 1, given latent class membership.
Homogeneity. In latent class analysis, patterns of responses by individuals are similar.
Latent class separation. When latent classes maximally distinguish among cases.
Entropy. In latent class analysis, refers to a type of average of class membership probabilities. High values represent high accuracy in classification.
Latent profile analysis. Latent class models with continuous indicators.
Mixture model. Statistical modeling framework that involves combining two or more distributions.
Structural equation mixture models. Models that combine continuous and categorical variables.
Class enumeration. Determining the number of latent classes.
Latent Markov model. Simplex model with latent class variables.
Stationarity. Parameters in a latent transition model are equal over time.
Latent stage sequential modeling. Latent transition models that investigate hypotheses about a particular pattern of transitions, such as the mover–stayer model.
Mover–stayer model. Specific latent state sequential model that classifies individuals into a class of individuals who make some transitions and a class of individuals who make no transitions.
Auxiliary variables. Variables in a latent class or latent transition model serving as predictors of outcomes.
Growth mixture models. Structural models that use latent classes to group latent growth trajectories.
Overextraction. Tendency to conclude too many latent classes, because a model with more classes is likely to fit better than a model with fewer classes.
Structural equation mixture models. Structural equation models that include categorical latent variables.

Chapter 11

Time series. Any sequence of measurement over time.
Period. The number of time units it takes for a value to repeat.
Frequency. The number of complete cycles per unit of time.
Amplitude. The magnitude of difference between the lowest and highest values within a given cycle.
Integrating. Removing trends, usually through differencing.
Stationarity. Parameters do not change over time. May include mean, the variance, or the autocovariance.
White noise. Unsystematic variation over time that remains when means and autocorrelations are stationary.
Autocorrelation function. Histogram of the magnitude of correlation values for each lag order. Also known as a correlogram.
Prewhitening. Removal of changes in means or autocorrelations so that parameters are stationary and only white noise remains.
AR(1). First-order autoregression process.
MA(1). First-order moving average model.
ARMA(1,1). First-order autoregressive first-order moving average model.
Slutsky–Yule effect. Moving averages may create cyclical patterns even if there is no underlying autoregressive process.

Integration. The process of detrending data, often through differencing.
Ergodicity. The condition that the covariance of a series approaches zero as lag length increases. Under this condition, the mean of a series will converge to the ensemble mean as the number of time points approaches infinity.

Chapter 12

Event history analysis. Synonym for survival analysis.
Failure time analysis. Synonym for survival analysis.
Discrete-time survival analysis. Survival analysis of discrete time points, usually applied to fewer regular intervals than continuous-time survival analysis. Usually uses a special specification of the logistic regressions model.
Continuous-time survival analysis. Survival analysis models for a time metric that has many values. Cox regression is the most popular form.
Cox regression. Continuous-time survival analysis model developed by Sir David Cox. Sometimes referred to as the proportional hazard model.
Censored. Data for which some events are not observed because of the time limitations of the study. Data may be left, right, or interval censored. Right censoring is most commonly addressed with time series analysis.
Right censoring. Occurs when the study observation period ends before the outcome of the event can be observed for all cases.
Left censoring. Occurs when the observation period begins after some events may have already occurred.
Interval censoring. Timing of the event is unknown within intervals of the study.
Uninformative censoring. Censoring is unrelated to the probability of the occurrence of the event after controlling for other factors.
Informative censoring. When the probability of the event's occurrence is independently related to whether or not the case has been censored.
Hazard rate. Risk rate of experiencing the event for any time interval.
Risk pool. Cases that are still at risk for experiencing the event at a certain interval or time point.
Survivor or survival function. Cumulative proportion of those still in the risk pool (i.e., those that have not experience the event).
Median survival time. Point at which the survival risk dips below .50, giving the expected time it takes for the event to occur.
Life table. Summary table of hazard and survivor values by time intervals.
Person-period data format. Data structure in which each repeated measure appears as a separate record in the data set. Sometimes called the "long" format. Used with conventional logistic regression discrete-time survival analysis.
Proportional hazard assumption. The hazard function considered across time applies equally well to each case where the ratio of one case's hazard relative to any other case maintains the same proportion over time.
Proportional odds assumption. Assumption that the odds of the hazard must be parallel over time for values of the covariate.
Baseline hazard. Hazard for a referent group. "Baseline" in this context does not refer to the first time point.
Ties. Cases with the same time to event.
Exact method. Method of handling ties in survival analysis.

Censor variable. Variable in Cox regression analysis used to indicate whether or not an event has occurred during the course of the study.

Chapter 13

Missing completely at random (MCAR). Missing data mechanisms in which missingness is unrelated to any variables in the data set.

Missing at random (MAR). Missingness is related to other variables in the data set but not to the values of the variable that is missing data.

Missing not at random (MNAR). Neither MCAR and MAR conditions are met. The variable is related to the probability of missingness even after accounting for other variables.

Missing data mechanisms. Missing data condition – whether data are MCAR, MAR, or NMAR.

Ignorable missingness. When the variable with missing data meets the MCAR or MAR condition.

Nonignorable missingness. Usually used synonymously with MNAR, but includes an additional stipulation that parameters that govern the missing data mechanism (conditional distributions) are equal for the observed values and the missingness.

Univariate missingness. Data are missing among different cases for individual variables in the data set. Sometimes referred to 'as' item nonresponse."

Unit missingness. Data are missing among the same cases for a set of variables. Sometimes referred to as "multivariate missingness."

Monotone missing. Pattern of missing data in which values are missing from one point in time forward, as when cases drop out of the study.

Planned missingness. Data are missing by design. May included modules in a survey asked of only a subset of respondents, longitudinal data collected at individually varying intervals, or cohort sequential designs.

Observed data coverage. The proportion of cases present (nonmissing), usually considered univariately or bivariately.

Outcome dependent. The missingness indicators are modeled as a function of the outcome.

Parameter dependent. The missingness indicators are modeled as a function of the parameters in the model (e.g., intercept and slope factors).

Pattern mixture models. The approach to analysis when data are thought to be MNAR in which missing data patterns are grouped by latent class analysis or are modeled as a function of continuous random factors.

Complete-case identification. Parameter estimates are based on the complete cases.

Neighboring case identification. Parameter estiamtes are based on other missing data groups that are the most similar.

Available case identification. Parameter estimates are based on a weighted average of all other patterns.

Future time dependence. Missingness at later time points may still be dependent on later values under this identifying restriction

Index

Page numbers in **bold** indicate tables and in *italics* indicate figures.