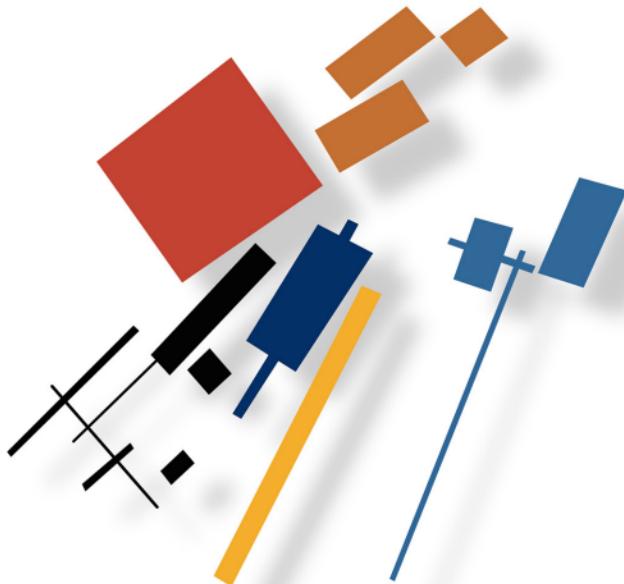


FOURTH EDITION

# Principles and Practice of Structural Equation Modeling



REX B. KLINE



THE GUILFORD PRESS

# Principles and Practice of Structural Equation Modeling

# Methodology in the Social Sciences

David A. Kenny, Founding Editor

Todd D. Little, Series Editor

[www.guilford.com/MSS](http://www.guilford.com/MSS)

This series provides applied researchers and students with analysis and research design books that emphasize the use of methods to answer research questions. Rather than emphasizing statistical theory, each volume in the series illustrates when a technique should (and should not) be used and how the output from available software programs should (and should not) be interpreted. Common pitfalls as well as areas of further development are clearly articulated.

## RECENT VOLUMES

APPLIED MISSING DATA ANALYSIS

*Craig K. Enders*

APPLIED META-ANALYSIS FOR SOCIAL SCIENCE RESEARCH

*Noel A. Card*

DATA ANALYSIS WITH Mplus

*Christian Geiser*

INTENSIVE LONGITUDINAL METHODS: AN INTRODUCTION  
TO DIARY AND EXPERIENCE SAMPLING RESEARCH

*Niall Bolger and Jean-Philippe Laurenceau*

DOING STATISTICAL MEDIATION AND MODERATION

*Paul E. Jose*

LONGITUDINAL STRUCTURAL EQUATION MODELING

*Todd D. Little*

INTRODUCTION TO MEDIATION, MODERATION, AND CONDITIONAL  
PROCESS ANALYSIS: A REGRESSION-BASED APPROACH

*Andrew F. Hayes*

BAYESIAN STATISTICS FOR THE SOCIAL SCIENCES

*David Kaplan*

CONFIRMATORY FACTOR ANALYSIS FOR APPLIED RESEARCH, SECOND EDITION

*Timothy A. Brown*

PRINCIPLES AND PRACTICE OF STRUCTURAL EQUATION MODELING, FOURTH EDITION

*Rex B. Kline*

# Principles and Practice of Structural Equation Modeling

FOURTH EDITION

Rex B. Kline

*Series Editor's Note by Todd D. Little*



THE GUILFORD PRESS  
New York      London

© 2016 The Guilford Press  
A Division of Guilford Publications, Inc.  
370 Seventh Avenue, Suite 1200, New York, NY 10001  
[www.guilford.com](http://www.guilford.com)

All rights reserved

No part of this book may be reproduced, translated, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission from the publisher.

Printed in the United States of America

This book is printed on acid-free paper.

Last digit is print number: 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data is available from the publisher

978-1-4625-2334-4 (paperback)  
978-1-4625-2335-1 (hardcover)

*For my family—  
Joanna, Julia Anne, and Luke Christopher  
and Stephen Albert Hamilton Wright (1957–2014)*

I sense the world might be more dreamlike, metaphorical, and poetic than we currently believe—but just as irrational as sympathetic magic when looked at in a typically scientific way. I wouldn't be surprised if poetry—poetry in the broadest sense, in the sense of a world filled with metaphor, rhyme, and recurring patterns, shapes, and designs—is how the world works. The world isn't logical, it's a song.

—DAVID BYRNE (2009)

# Series Editor's Note

Rex Kline has assembled a fourth edition that retains all the wonderful features of his bestselling earlier editions, and he seamlessly integrates recent advances in structural equation modeling (SEM). Rex is a scholar of SEM and has a special gift—of being able to communicate complex statistical concepts in language that all readers can grasp. The accessible style of writing and the many pedagogical features of the book (e.g., chapter-end annotated reading lists, exercises with answers) make it a “must have” for any user of SEM. It is a resource that keeps improving and expanding with each new edition and is the resource I recommend first on this subject—whether the question comes from a beginner or an experienced user.

As a scholar of modern statistical practice and techniques, Rex has studied the developments and advances in the world of SEM generally, and he has covered “hot” topics, such as Pearl’s structural causal modeling. His coverage of Pearl’s graph theory approach to causal reasoning, as many of the reviewers of prepublication drafts of the fourth edition have also noted, is both easy to understand and comprehensive. It’s so good, he ought to get a prize for best in presentation! In this new edition, he takes us through causal mediation analysis, conditional process modeling, and confirmatory factor analysis with categorical indicators. Other additions to this masterpiece of pedagogy include insightful discussions of significance testing, the use of bootstrap estimation, and the principles of measurement theory.

Although Rex suggests in his Introduction that no single book can cover all of SEM, his book is about as thorough as they come. His didactic approach is refreshing and engaging, and the breadth and depth of material covered is simply impressive. As he notes and you will feel, Rex is a researcher talking to you as a fellow researcher, carefully explaining in conceptually driven terms the logic and principles that underlie the world of SEM. The wealth of examples provide entry points for researchers across a

broad array of disciplines. This book will speak to you regardless of your field or specific area of expertise.

As always, the support materials that Rex provides are thorough: he covers now six different SEM software packages (Amos, EQS, lavaan for R, LISREL, Mplus, and Stata)! The Appendix material is a treasure trove of useful building blocks, from the elements of LISREL notation, to practical advice, to didactic presentation of complex ideas and procedures. Rex has assembled real-world examples of troublesome data to demonstrate how to handle the analysis problems that inevitably pop up. These features have always been a mainstay of earlier editions, but they have now been expanded to cover even more topics. Rex bookends all this material with an introductory chapter that truly sets the stage for the journey through the land of SEM and a concluding chapter that covers very practical best-practice advice for every step along the way.

Enjoy Rex Kline's classic for a fourth time!

TODD D. LITTLE  
*At 28,000 feet*  
*On my way to Wit's End*  
*Lakeside, Montana*

# Acknowledgments

It was an honor to work once again with such talented people. The Methodology and Statistics publisher at The Guilford Press, C. Deborah Laughton, continues to delight with her uncanny ability to give just the right feedback at exactly the right time. Her enthusiasm and focus help to keep everything on track. Any writer would be blessed to have C. Deborah on his or her side. The names of reviewers of earlier drafts were revealed to me only after the writing was complete, and their original comments were not associated with their names. Thank you very much to all the people listed next who devoted their time and effort to communicate their impressions about various chapters. Their comments were invaluable in revising the fourth edition:

- Ryan Bowles, Human Development and Family Studies, College of Social Science, Michigan State University
- Chris L. S. Coryn, The Evaluation Center, Western Michigan University
- Christine DiStefano, Educational Studies, College of Education, University of South Carolina
- Debbie L. Hahs-Vaughn, Department of Educational and Human Sciences, College of Education and Human Performance, University of Central Florida
- Donna Harrington, School of Social Work, University of Maryland
- Michael R. Kotowski, School of Communication Studies, College of Communication and Information, University of Tennessee
- Richard Wagner, Department of Psychology, Florida State University
- Craig S. Wells, Department of Educational Policy, Research and Administration, College of Education, University of Massachusetts Amherst

- Tiffany Whittaker, Department of Educational Psychology, College of Education, University of Texas at Austin
- John Willse, Department of Educational Research Methodology, School of Education, University of North Carolina at Greensboro

A special thanks goes to Judea Pearl, of the Computer Science Department at the University of California, Los Angeles. He kindly answered many questions about graph theory and, along with Bryant Chen, gave helpful suggestions on earlier drafts of Chapter 8 about the structural causal model. Series Editor Todd D. Little, of the Institute for Measurement, Methodology, Analysis, and Policy at Texas Tech University, provided insightful comments and suggestions for the final version of the manuscript. I always learn something new when working with a good copyeditor, and this time with Betty Pessagno serving as the copyeditor was no exception. Her work with the original manuscript improved the quality of the presentation, just as she did for the third edition. At the Guilford Press, it was a pleasure to work again with Production Editor William Meyer and with Art Director Paul Gordon, who designed the elegant book cover. Chuck Huber (StataCorp), Linda Muthén (Muthén & Muthén), and Peter Bentler and Eric Wu (Multivariate Software) commented on earlier drafts of descriptions of, respectively, Stata, Mplus, and EQS.

And, once again, my deepest thanks to my wife, Joanna, and our children, Julia and Luke, for all their love and support while writing this book. With all this sustenance available to me, any limitations that remain in the book are clearly my own.

# Contents

<b>Introduction</b>	1
Book Website	2
Pedagogical Approach	2
Principles over Computer Tools	3
Symbols and Notation	3
Life's a Journey, Not a Destination	3
Plan of the Book	4
 <b>PART I. CONCEPTS AND TOOLS</b>	
<b>1 • Coming of Age</b>	<b>7</b>
Preparing to Learn SEM	7
Definition of SEM	9
Importance of Theory	10
A Priori, but Not Exclusively Confirmatory	11
Probabilistic Causation	11
Observed Variables and Latent Variables	12
Data Analyzed in SEM	13
SEM Requires Large Samples	14
Less Emphasis on Significance Testing	17
SEM and Other Statistical Techniques	17
SEM and Other Causal Inference Frameworks	18
Myths about SEM	20
Widespread Enthusiasm, but with a Cautionary Tale	21
Family History	23
Summary	24
Learn More	24

<b>2 • Regression Fundamentals</b>	<b>25</b>
Bivariate Regression	25
Multiple Regression	30
Left-Out Variables Error	35
Suppression	36
Predictor Selection and Entry	37
Partial and Part Correlation	39
Observed versus Estimated Correlations	41
Logistic Regression and Probit Regression	44
Summary	47
Learn More	47
Exercises	48
<b>3 • Significance Testing and Bootstrapping</b>	<b>49</b>
Standard Errors	49
Critical Ratios	51
Power and Types of Null Hypotheses	52
Significance Testing Controversy	54
Confidence Intervals and Noncentral Test Distributions	57
Bootstrapping	60
Summary	62
Learn More	62
Exercises	63
<b>4 • Data Preparation and Psychometrics Review</b>	<b>64</b>
Forms of Input Data	64
Positive Definiteness	67
Extreme Collinearity	71
Outliers	72
Normality	74
Transformations	77
Relative Variances	81
Missing Data	82
Selecting Good Measures and Reporting about Them	88
Score Reliability	90
Score Validity	93
Item Response Theory and Item Characteristic Curves	94
Summary	95
Learn More	96
Exercises	96
<b>5 • Computer Tools</b>	<b>97</b>
Ease of Use, Not Suspension of Judgment	97
Human–Computer Interaction	98

---

Tips for SEM Programming	100
SEM Computer Tools	101
Other Computer Resources for SEM	111
Computer Tools for the SCM	112
Summary	113
Learn More	113
<b>PART II. SPECIFICATION AND IDENTIFICATION</b>	
<b>6 • Specification of Observed Variable (Path) Models</b>	117
Steps of SEM	117
Model Diagram Symbols	121
Causal Inference	122
Specification Concepts	126
Path Analysis Models	129
Recursive and Nonrecursive Models	135
Path Models for Longitudinal Data	138
Summary	141
Learn More	142
Exercises	142
<b>APPENDIX 6.A.</b> LISREL Notation for Path Models	143
<b>7 • Identification of Observed-Variable (Path) Models</b>	145
General Requirements	145
Unique Estimates	148
Rule for Recursive Models	149
Identification of Nonrecursive Models	150
Models with Feedback Loops and All Possible Disturbance Correlations	150
Graphical Rules for Other Types of Nonrecursive Models	153
Respecification of Nonrecursive Models That Are Not Identified	155
A Healthy Perspective on Identification	157
Empirical Underidentification	157
Managing Identification Problems	158
Path Analysis Research Example	159
Summary	159
Learn More	160
Exercises	160
<b>APPENDIX 7.A.</b> Evaluation of the Rank Condition	161
<b>8 • Graph Theory and the Structural Causal Model</b>	164
Introduction to Graph Theory	164
Elementary Directed Graphs and Conditional Independences	166

Implications for Regression Analysis	170
Basis Set	173
Causal Directed Graphs	174
Testable Implications	176
Graphical Identification Criteria	177
Instrumental Variables	180
Causal Mediation	181
Summary	184
Learn More	185
Exercises	185
<b>APPENDIX 8.A.</b> Locating Conditional Independences in Directed Cyclic Graphs	186
<b>APPENDIX 8.B.</b> Counterfactual Definitions of Direct and Indirect Effects	187
<b>9 • Specification and Identification of Confirmatory Factor Analysis Models</b>	188
Latent Variables in CFA	188
Factor Analysis	189
Characteristics of EFA Models	191
Characteristics of CFA Models	193
Other CFA Specification Issues	195
Identification of CFA Models	198
Rules for Standard CFA Models	201
Rules for Nonstandard CFA Models	202
Empirical Underidentification in CFA	206
CFA Research Example	206
Summary	207
Learn More	207
Exercises	209
<b>APPENDIX 9.A.</b> LISREL Notation for CFA Models	210
<b>10 • Specification and Identification of Structural Regression Models</b>	212
Causal Inference with Latent Variables	212
Types of SR Models	213
Single Indicators	214
Identification of SR Models	217
Exploratory SEM	219
SR Model Research Examples	220
Summary	223
Learn More	225
Exercises	225
<b>APPENDIX 10.A.</b> LISREL Notation for SR Models	226

**PART III. ANALYSIS**

<b>11 • Estimation and Local Fit Testing</b>	231
Types of Estimators	231
Causal Effects in Path Analysis	232
Single-Equation Methods	233
Simultaneous Methods	235
Maximum Likelihood Estimation	235
Detailed Example	239
Fitting Models to Correlation Matrices	253
Alternative Estimators	255
A Healthy Perspective on Estimation	258
Summary	259
Learn More	259
Exercises	260
<b>APPENDIX 11.A.</b> Start Value Suggestions for Structural Models	261
<b>12 • Global Fit Testing</b>	262
State of Practice, State of Mind	262
A Healthy Perspective on Global Fit Statistics	263
Model Test Statistics	265
Approximate Fit Indexes	266
Recommended Approach to Fit Evaluation	268
Model Chi-Square	270
RMSEA	273
SRMR	277
Tips for Inspecting Residuals	278
Global Fit Statistics for the Detailed Example	278
Testing Hierarchical Models	280
Comparing Nonhierarchical Models	286
Power Analysis	290
Equivalent and Near-Equivalent Models	292
Summary	297
Learn More	298
Exercises	298
<b>APPENDIX 12.A.</b> Model Chi-Squares Printed by LISREL	299
<b>13 • Analysis of Confirmatory Factor Analysis Models</b>	300
Fallacies about Factor or Indicator Labels	300
Estimation of CFA Models	301
Detailed Example	304
Respecification of CFA Models	309
Special Topics and Tests	312
Equivalent CFA Models	315
Special CFA Models	319

Analyzing Likert-Scale Items as Indicators	323
Item Response Theory as an Alternative to CFA	332
Summary	333
Learn More	333
Exercises	334
<b>APPENDIX 13.A.</b> Start Value Suggestions for Measurement Models	335
<b>APPENDIX 13.B.</b> Constraint Interaction in CFA Models	336
<b>14 • Analysis of Structural Regression Models</b>	338
Two-Step Modeling	338
Four-Step Modeling	339
Interpretation of Parameter Estimates and Problems	340
Detailed Example	341
Equivalent SR Models	348
Single Indicators in a Nonrecursive Model	349
Analyzing Formative Measurement Models in SEM	352
Summary	361
Learn More	362
Exercises	362
<b>APPENDIX 14.A.</b> Constraint Interaction in SR Models	363
<b>APPENDIX 14.B.</b> Effect Decomposition in Nonrecursive Models and the Equilibrium Assumption	364
<b>APPENDIX 14.C.</b> Corrected Proportions of Explained Variance for Nonrecursive Models	365
<b>PART IV. ADVANCED TECHNIQUES AND BEST PRACTICES</b>	
<b>15 • Mean Structures and Latent Growth Models</b>	369
Logic of Mean Structures	369
Identification of Mean Structures	373
Estimation of Mean Structures	374
Latent Growth Models	374
Detailed Example	375
Comparison with a Polynomial Growth Model	387
Extensions of Latent Growth Models	390
Summary	392
Learn More	392
Exercises	393
<b>16 • Multiple-Samples Analysis and Measurement Invariance</b>	394
Rationale of Multiple-Samples SEM	394
Measurement Invariance	396
Testing Strategy and Related Issues	399
Example with Continuous Indicators	403

---

Example with Ordinal Indicators	411
Structural Invariance	420
Alternative Statistical Techniques	420
Summary	421
Learn More	421
Exercises	422
<b>APPENDIX 16.A. Welch–James Test</b>	423
<b>17 • Interaction Effects and Multilevel Structural Equation Modeling</b>	424
Interactive Effects of Observed Variables	424
Interactive Effects in Path Analysis	431
Conditional Process Modeling	432
Causal Mediation Analysis	435
Interactive Effects of Latent Variables	437
Multilevel Modeling and SEM	444
Summary	450
Learn More	450
Exercises	451
<b>18 • Best Practices in Structural Equation Modeling</b>	452
Resources	452
Specification	454
Identification	457
Measures	458
Sample and Data	458
Estimation	461
Respecification	463
Tabulation	464
Interpretation	465
Avoid Confirmation Bias	466
Bottom Lines and Statistical Beauty	466
Summary	467
Learn More	467
<b>Suggested Answers to Exercises</b>	469
<b>References</b>	489
<b>Author Index</b>	510
<b>Subject Index</b>	516
<b>About the Author</b>	534

The companion website [www.guilford.com/kline-materials](http://www.guilford.com/kline-materials) provides downloadable data, syntax, and output for all the book's examples in six widely used SEM computer tools and links to related web pages.



# Introduction

It is an honor to present the fourth edition of this book. Like the previous editions, this one introduces structural equation modeling (SEM) in a clear, accessible way for readers without strong quantitative backgrounds. New examples of the application of SEM are included in this edition, and all the examples cover a wide range of disciplines, including education, psychometrics, human resources, and psychology, among others. Some examples were selected owing to technical problems in the analysis, but such examples give a context for discussing how to handle problems that can crop up in SEM. So not all applications of SEM described in this book are picture perfect, but neither are actual research problems.

The many changes in this edition are intended to enhance the pedagogical presentation and cover recent developments. The biggest changes are as follows.

1. This is one of the first introductory books to introduce Judea Pearl's structural causal model (SCM), an approach that offers unique perspectives on causal modeling. It is also part of new developments in causal mediation analysis.
2. Computer files for all detailed examples are now available for a total of six widely used SEM computer tools, including Amos, EQS, lavaan for R, LISREL, Mplus, and Stata. Computer tools for the SCM are also described.
3. Presentations on model specification, identification, and estimation are reorganized to separate coverage of observed variable (path) models from that of latent variable models. The specification and identification of path models are covered before these topics are dealt with for latent variable models. Later chapters that introduce estimation and hypothesis testing do not assume knowledge of latent variable models. This organization makes it easier for instructors who prefer to cover the specification, identification, and analysis of path models before doing so for latent variable models.

4. Two changes concern the technique of confirmatory factor analysis (CFA). The analysis of ordinal data in CFA is covered in more detail with two new examples, one of which concerns the topic of measurement invariance. The topic just mentioned is now covered in its own chapter in this edition.

## **BOOK WEBSITE**

The address for this book's website is [www.guilford.com/kline](http://www.guilford.com/kline). From the site, you can freely access or download the following resources:

- Computer files—data, syntax, and output—for all detailed examples in this book.
- Links to related web pages, including sites for computer programs and calculating webpages that perform certain types of analyses.

The website promotes a learning-by-doing approach. The availability of both syntax and data files means that you can reproduce the analyses in this book using the corresponding software program. Even without access to a particular program, such as Mplus, you can still download and open on your own computer the Mplus output files for a particular example and view the results. This is because all computer files on the website are either plain text (ASCII) files that require nothing more than a basic text editor to view their contents or they are PDF (Portable Document Format) files that can be viewed with the freely available Adobe Reader. Even if you use a particular SEM computer tool, it is still worthwhile to review the files on the website generated by other programs. This is because it can be helpful to consider the same analysis from somewhat different perspectives. Some of the exercises for this book involve extensions of the analyses for these examples, so there are plenty of opportunities for practice with real data sets.

## **PEDAGOGICAL APPROACH**

You may be reading this book while participating in a course or seminar on SEM. This context offers the potential advantage of the structure and support available in a classroom setting, but formal coursework is not the only way to learn about SEM. Another is self-study, a method through which many researchers learn about what is, for them, a new statistical technique. (This is how I first learned about SEM, not in classes.) I assume that most readers are relative newcomers to SEM or that they already have some knowledge but wish to hone their skills.

Consequently, I will speak to you (through my author's voice) as one researcher to another, not as a statistician to the quantitatively naïve. For example, the instructional language of statisticians is matrix algebra, which conveys a lot of information in a rela-

tively short amount of space, but you must already be familiar with linear algebra to decode the message. There are other, more advanced works about SEM that emphasize matrix representations (Bollen, 1989; Kaplan, 2009; Mulaik, 2009b), and these works can be consulted when you are ready. Instead, fundamental concepts about SEM are presented here using the language of researchers: words and figures, not matrix equations. I will not shelter you from some of the more technical aspects of SEM, but I aim to cover requisite concepts in an accessible way that supports continued learning.

## **PRINCIPLES OVER COMPUTER TOOLS**

You may be relieved to know that you are not at a disadvantage at present if you have no experience using an SEM computer tool. This is because the presentation in this book is not based on the symbolism or syntax associated with a particular software package. In contrast, some other books are linked to specific SEM computer tools. They can be invaluable for users of a particular program, but perhaps less so for others. Instead, key principles of SEM that users of *any* computer tool must understand are emphasized here. In this way, this book is more like a guide to writing style than a handbook about how to use a particular word processor. Besides, becoming proficient with a particular software package is just a matter of practice. But without strong conceptual knowledge, the output one gets from a computer tool for statistical analyses—including SEM—may be meaningless or, even worse, misleading.

## **SYMBOLS AND NOTATION**

As with other statistical techniques, there is no gold standard for notation in SEM, but the symbol set associated with the original syntax of LISREL is probably the most widely used in advanced works. For this reason, this edition introduces LISREL symbolism, but these presentations are optional; that is, I do not force readers to memorize LISREL symbols in order to get something out of this book. This is appropriate because the LISREL notational system can be confusing unless you have memorized the whole system. I use a few key symbols in the main text, but the rest of LISREL notation is described in chapter appendices.

## **LIFE'S A JOURNEY, NOT A DESTINATION**

Learning to use a new set of statistical techniques is also a kind of journey, one through a strange land, at least at the beginning. Such journeys require a commitment of time and the willingness to tolerate the frustration of trial and error, but this is one journey that you do not have to make alone. Think of this book as a travel atlas or even someone to advise you about language and customs, what to see and pitfalls to avoid, and what

lies just over the horizon. I hope that the combination of a conceptually based approach, many examples, and the occasional bit of sage advice presented in this book will help to make the statistical journey a little easier, maybe even enjoyable. (Imagine that!)

## **PLAN OF THE BOOK**

The topic of SEM is very broad, and not every aspect of it can be covered in a single volume. With this reality in mind, I will now describe the topics covered in this book. Part I introduces fundamental concepts and computer tools. Chapter 1 lays out the basic features of SEM. It also deals with myths about SEM and outlines its relation to other causal inference frameworks. Chapters 2 and 3 review basic statistical principles and techniques that form a groundwork for learning about SEM. These topics include regression analysis, statistical significance testing, and bootstrapping. How to prepare data for analysis in SEM and select good measures is covered in Chapter 4, and computer tools for SEM and the SCM are described in Chapter 5.

Part II consists of chapters about the specification and identification phases in SEM. Chapters 6 and 7 cover observed variable models, or path models. Chapter 8 deals with path analysis from the perspective of the SCM and causal graph theory. Chapters 9 and 10 are about, respectively, CFA models and structural regression (SR) models. The latter (SR models) have features of both path models and measurement models. Part III is devoted to the analysis. Chapters 11 and 12 deal with principles of estimation and hypothesis testing that apply to any type of structural equation model. The analysis of CFA models is considered in Chapter 13, and analyzing SR models is the subject of Chapter 14. Actual research problems are considered in these presentations.

Part IV is about advanced techniques and best practices. The analysis of means in SEM is introduced in Chapter 15, which also covers latent growth models. How to analyze a structural equation model with data from multiple samples is considered in Chapter 16, which also deals with the topic of measurement invariance in CFA. Estimation of the interactive effects of latent variables, conditional process analysis, causal mediation analysis, and the relation between multilevel modeling and SEM are all covered in Chapter 17. Chapter 18 offers best practice recommendations in SEM. This chapter also mentions common mistakes with the aim of helping you to avoid them.

## **Part I**

# Concepts and Tools



# 1

## Coming of Age

---

This book is your guide to the principles, assumptions, strengths, limitations, and application of structural equation modeling (SEM) for researchers and students without extensive quantitative backgrounds. Accordingly, the presentation is conceptually rather than mathematically oriented, the use of formulas and symbols is kept to a minimum, and many examples are offered of the application of SEM to research problems in various disciplines, including psychology, education, health sciences, and other areas. When you finish reading this book, I hope that you will have acquired the skills to begin to use SEM in your own research in an informed, principled way. The following observation attributed to the playwright George Bernard Shaw is relevant here: Life isn't about finding yourself, life is about creating yourself. Let's go create something together.

---

### **PREPARING TO LEARN SEM**

Listed next are recommendations for the best ways to get ready to learn about SEM. I offer these suggestions in the spirit of giving you a healthy perspective at the beginning of this task, one that empowers your sense of being a researcher.

#### **Know Your Area**

Strong familiarity with the theoretical and empirical literature in your research area is the single most important thing you could bring to SEM. This is because everything, from the specification of your initial model to modification of that model in subsequent reanalyses to interpretation of the results, must be guided by your domain knowledge. So you need first and foremost to be a *researcher*, not a statistician or a computer geek. This is true for most kinds of statistical analysis in that the value of the product (numeri-

cal results) depends on the quality of the ideas (your hypotheses) on which the analysis is based. Otherwise, that familiar expression about computer data analysis, garbage in–garbage out, may apply.

### **Know Your Measures**

Kühnel (2001) reminds us that learning about SEM has the by-product that researchers must deal with fundamental issues of measurement. Specifically, the analysis of measures with strong psychometric characteristics, such as good score reliability and validity, is essential in SEM. For example, it is impossible to analyze a structural equation model with latent variables that represent hypothetical constructs without thinking about how to measure those constructs. When you have just a single measure of a construct, it is especially critical for this single indicator to have good psychometric properties. Similarly, the analysis of measures with deficient psychometrics could bias the results.

### **Review Fundamental Statistical Concepts and Techniques**

Before learning about SEM, you should have a good understanding of (1) principles of regression techniques, including multiple regression, logistic regression, and probit regression; (2) the correct interpretation of results from tests of statistical significance; and (3) data screening and measure selection. These topics are reviewed over the next few chapters, but it may help to know now why they are so important.

Some kinds of statistical estimates in SEM are interpreted exactly as regression coefficients. The values of these coefficients are adjusted for correlated predictors just as they are in standard regression techniques. The potential for bias due to omitted predictors that covary with measured predictors is basically the same in SEM as in regression analysis. Results of significance tests are widely misunderstood in perhaps most analyses, including SEM, and you need to know how to avoid making common mistakes. Preparing the data for analysis in SEM requires doing a thorough preparation, screening for potential problems, and taking remedial action, if needed.

### **Use the Best Research Computer in the World . . .**

Which is the human brain; specifically—*yours*. At the end of the analysis in SEM—or any type of statistical analysis—it is you as the researcher who must evaluate the degree of support for the hypotheses, explain any unexpected findings, relate the results to those of previous studies, and reflect on implications of the findings for future research. These are all matters of judgment. A statistician or computer geek could help you to select data analysis tools or write program syntax, but not the rest without your content expertise. As aptly stated by Pedhazur and Schmelkin (1991), “no amount of proficiency will do you any good, if you do not think” (p. 2).

## Get a Computer Tool for SEM

Obviously, you need a computer program to conduct the analysis. In SEM, many choices of computer tools are now available, some for no cost. Examples of free computer programs or procedures include *Onyx*, a graphical environment for creating and testing structural equation models, and various SEM packages such as *lavaan* or *sem* for R, which is an open-source language and environment for statistical computing and graphics. Commercial options for SEM include Amos, EQS, LISREL, and Mplus, which are all free-standing programs that do not require a larger computing environment. Procedures for SEM that require the corresponding environment include the CALIS procedure of SAS/STAT, the *sem* and *gsem* commands in Stata, the RAMONA procedure of Systat, and the SEPATH procedure in STATISTICA. All the computer tools just mentioned and others are described in Chapter 5. The website for this book has links to home pages for SEM computer tools (see the Introduction).

## Join the Community

An Internet electronic mail network called SEMNET is dedicated to SEM.<sup>1</sup> It serves as an open forum for discussion and debate about the whole range of issues associated with SEM. It also provides a place to ask questions about analyses or about more general issues, including philosophical ones (e.g., the nature of causality, causal inference). Subscribers to SEMNET come from various disciplines, and they range from newcomers to seasoned veterans, including authors of many works cited in this book. Sometimes the discussion gets a little lively (sparks can fly), but so it goes in scientific discourse. Whether you participate as a lurker, or someone who mainly reads posts, or as an active poster, SEMNET offers opportunities to learn something new. There is even a theme song for SEM, the hilarious *Ballad of the Casual Modeler*, by David Rogosa (1988). You can blame me if the tune gets stuck in your head.

## DEFINITION OF SEM

The term **structural equation modeling** (SEM) does not designate a single statistical technique but instead refers to a family of related procedures. Other terms such as **covariance structure analysis**, **covariance structure modeling**, or **analysis of covariance structures** are also used in the literature to classify these techniques under a single label. These terms are essentially interchangeable, but only structural equation modeling is used throughout this book.

Pearl (2012) defines SEM as a causal inference method that takes three inputs (I) and generates three outputs (O). The inputs are

---

<sup>1</sup>[www2.gsu.edu/~mkteer/semnet.html](http://www2.gsu.edu/~mkteer/semnet.html)

- I-1. A set of qualitative causal hypotheses based on theory or results of empirical studies that are represented in a structural equation model. The hypotheses are typically based on assumptions, only some of which can actually be verified or tested with the data.
- I-2. A set of queries or questions about causal relations among variables of interest such as, what is the magnitude of the direct effect of  $X$  on  $Y$  (represented as  $X \rightarrow Y$ ), controlling for all other presumed causes of  $Y$ ? All queries follow from model specification.
- I-3. Most applications of SEM are in nonexperimental designs, but data from experimental or quasi-experimental designs can be analyzed, too—see Bergsma, Croon, and Hagenaars (2009) for more information.

The outputs of SEM are

- O-1. Numeric estimates of model parameters for hypothesized effects including, for example,  $X \rightarrow Y$ , given the data.
- O-2. A set of logical implications of the model that may not directly correspond to a specific parameter but that still can be tested in the data. For example, a model may imply that variables  $W$  and  $Y$  are unrelated, controlling for certain other variables in the model.
- O-3. The degree to which the testable implications of the model are supported by the data.

The next few sections elaborate on the inputs and outputs of SEM.

## IMPORTANCE OF THEORY

As in other statistical techniques, the quality of the outputs of SEM depend on the validity of the researcher's ideas (the first input, I-1). Thus, the point of SEM is to *test a theory* by specifying a model that represents predictions of that theory among plausible constructs measured with appropriate observed variables (Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulian, 2007). If such a model does not ultimately fit the data, this outcome is interesting because there is value in reporting models that challenge or debunk theories.

Beginners sometimes mistakenly believe that the point of SEM is to *find a model that fits the data*, but this outcome by itself is not very impressive. This is because *any* model, even one that is grossly wrong (misspecified), can be made to fit the data by making it more complicated (adding parameters). In fact, if a structural equation model is specified to be as complex as possible, it will perfectly fit the data. This is a general characteristic of statistical modeling, not just of SEM. But the point is that “success” in SEM is

determined by whether the analysis deals with substantive theoretical issues regardless of whether or not a model is retained. In contrast, whether or not a scientifically meaningless model fits the data is irrelevant (Millsap, 2007).

## A PRIORI, BUT NOT EXCLUSIVELY CONFIRMATORY

Computer tools for SEM require you to provide a lot of information about things such as the directionalities of causal effects among variables (e.g.,  $X \rightarrow Y$  vs.  $Y \rightarrow X$ ). These a priori specifications reflect your hypotheses, and in total they make up the model to be analyzed. In this sense, SEM can be viewed as confirmatory. But as often happens, the data may be inconsistent with the model, which means that you must either abandon your model or modify the hypotheses on which it is based. In a **strictly confirmatory** application, the researcher has a single model that is either retained or rejected based on its correspondence to the data (Jöreskog, 1993), and that's it. But on few occasions will the scope of model testing be so narrow.

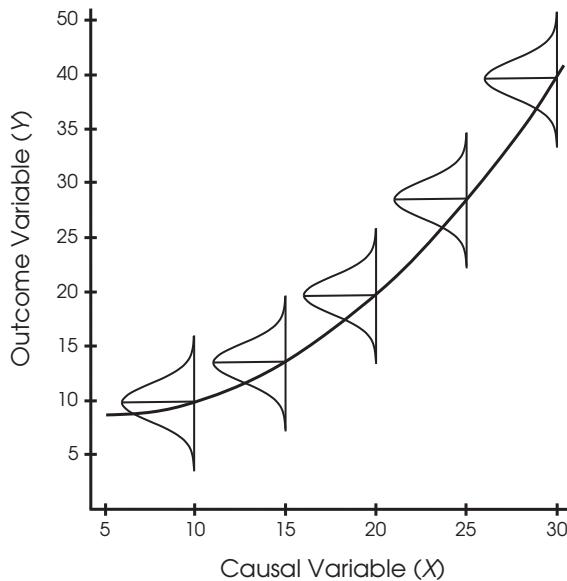
A second, somewhat less restrictive context involves the testing of **alternative models**, and it refers to situations in which more than one a priori model is available (Jöreskog, 1993). Alternative models usually include the same observed variables but represent different patterns of effects among them. This context requires sufficient bases to specify more than one model; the particular model with acceptable correspondence to data may be retained, but the rest will be rejected. A third context, **model generation**, is probably the most common and occurs when an initial model does not fit the data and is subsequently modified. The respecified model is then tested again with the same data (Jöreskog, 1993). The goal of this process is to "discover" a model with three attributes: It makes theoretical sense, it is reasonably parsimonious, and it has acceptably close correspondence to the data.

## PROBABILISTIC CAUSATION

Models analyzed in SEM generally assume probabilistic causality, not **deterministic causality**. The latter means that given a change in a causal variable, the same consequence is observed in all cases on the outcome variable. In contrast, **probabilistic causality** allows for changes to occur in outcomes at some probability  $< 1.0$ . Estimation of these probabilities (effects) with sample data are typically based on specific distributional assumptions, such as normality. Causality as a functional relation between two quantitative variables is preserved in this viewpoint, but causal effects are assumed to shift a probability distribution (Mulaik, 2009b).

An example of a probabilistic causal model is presented in Figure 1.1. The specific functional relation between the two variables depicted in the figure is

$$\hat{Y} = .05X^2 - .50X + 10.00 \quad (1.1)$$



**FIGURE 1.1.** Example of a probabilistic relation between a continuous cause ( $X$ ) and an outcome ( $Y$ ).

where  $\hat{Y}$  is the predicted average score, given  $X$ , in normal distributions where observed scores on  $Y$  vary around  $\hat{Y}$ . In the figure, as  $X$  increases, the predictive distribution for  $Y$  increases in both curvilinear and linear ways, but there is error variance at every point in a probabilistic causal relation.

## OBSERVED VARIABLES AND LATENT VARIABLES

A key feature of SEM is the explicit distinction between observed (manifest) variables and latent variables. Observed variables represent your data—that is, variables for which you have collected scores and entered in a data file. These variables can be categorical or continuous, but all latent variables in SEM are continuous. There are other statistical methods for analyzing categorical latent variables, but SEM deals with continuous latent variables only.

Latent variables in SEM generally correspond to **hypothetical constructs**, or explanatory entities presumed to reflect a continuum that is not directly observable. An example is the construct of *intelligence*. There is no single, definitive measure of intelligence. Instead, researchers use different types of observed variables, such as tasks of verbal reasoning or memory capacity, to assess facets of intelligence. Latent variables in SEM can represent a wide range of phenomena. For example, constructs about attributes of people (e.g., intelligence, anxiety); higher-level units of analysis (e.g., groups, neighborhoods, geographic regions); measures, such as method effects (e.g., self-report

vs. observational); or sources of information (e.g., teachers vs. students) can all be represented as latent variables in SEM.

An observed variable used as an indirect measure of a construct is an **indicator**, and the statistical realization of a construct based on analyzing scores from its indicators is a **factor**. The explicit distinction between indicators and factors in SEM allows one to test a wide variety of hypotheses about measurement. Suppose that a researcher believes that variables  $X_1-X_3$  tap a common domain that is distinct from the one measured by  $X_4-X_5$ . In SEM, it is relatively easy to specify a model where  $X_1-X_3$  are indicators of one factor and  $X_4-X_5$  are indicators of a different factor. If the fit of the model just described to the data is poor, the hypothesis behind its specification will be rejected. The ability to analyze both observed and latent variables distinguishes SEM from more standard statistical techniques, such as the analysis of variance (ANOVA) and multiple regression, which analyze observed variables only.

Another latent variable category in SEM corresponds to residual or error terms, which can be associated with either observed variables or factors specified as outcomes (dependent variables). In the case of indicators, a residual term represents variance not explained by the factor that the corresponding indicator is supposed to measure. Part of this unexplained variance is due to random measurement error or score unreliability. The explicit representation of measurement error is a special characteristic of SEM. This is not to say that SEM can compensate for gross psychometric flaws—no technique can—but this property lends a more realistic quality to the analysis. Some more conventional statistical techniques make unreasonable assumptions in this area. For example, it is assumed in multiple regression that all predictors are measured without error. This assumption is routinely violated in practice. In diagrams of structural equation models, residual terms may be represented using the same symbols as for substantive latent variables. This is because error variance must be estimated, given the model and data. In this sense error, variance is not directly observable in the raw data and is thus latent.

The capability to analyze in SEM observed or latent variables as either causes or outcomes permits great flexibility in the types of hypotheses that can be tested. But models in SEM are not *required* to have substantive latent variables. (Most structural equation models have error terms represented as latent variables.) That is, the evaluation of models that concern effects among only observed variables is certainly possible in SEM. This describes the technique of path analysis, the original member of the SEM family.

## DATA ANALYZED IN SEM

The basic datum of SEM is the covariance, which is defined for two observed continuous variables X and Y as follows:

$$\text{cov}_{XY} = r_{XY} SD_X SD_Y \quad (1.2)$$

where  $r_{XY}$  is the Pearson correlation and  $SD_X$  and  $SD_Y$  are their standard deviations.<sup>2</sup> A covariance thus represents the strength of the linear association between  $X$  and  $Y$  and their variabilities, albeit with a single number. Because the covariance is an unstandardized statistic, its value has no fixed lower or upper bound. For example, covariances of, say, -1,003.26 or 13.58 are possible, given the scales of the original scores. In any event,  $\text{cov}_{XY}$  expresses more information than  $r_{XY}$ , which says something about association in a standardized metric only.

To say that the covariance is the basic statistic of SEM implies that the analysis has two goals: (1) to understand patterns of covariances among a set of observed variables and (2) to explain as much of their variance as possible with the researcher's model. The part of a structural equation model that represents hypotheses about variances and covariances is the **covariance structure**. The next several chapters outline the rationale of analyzing covariance structures, but essentially all models in SEM have covariance structures.

Some researchers, especially those who use ANOVA as their main analytical tool, have the impression that SEM is concerned solely with covariances. This view is too narrow because means can be analyzed in SEM, too. But what really distinguishes SEM is that means of latent variables can also be estimated. In contrast, ANOVA is concerned with means of observed variables only. It is also possible to estimate in SEM effects traditionally associated with ANOVA, including between-groups and within-groups (e.g., repeated measures) mean contrasts. For example, in SEM one can estimate the magnitude of group mean differences on latent variables, something that is not really feasible in ANOVA. When means are analyzed along with covariances in SEM, the model has both a covariance structure and a **mean structure**, and the mean structure often represents the estimation of means on latent variables. Means are not analyzed in most SEM studies—that is, a mean structure is not required—but the option to do so provides additional flexibility.

## SEM REQUIRES LARGE SAMPLES

Attempts have been made to adapt SEM techniques to work in smaller samples (Jung, 2013), but it is still generally true that SEM is a large-sample technique. Implications of this property are considered throughout the book, but certain types of estimates in SEM, such as standard errors for effects of latent variables, may be inaccurate when the sample size is not large. The risk for technical problems in the analysis is greater, too.

Because sample size is such an important issue, let us now consider the bottom-line question: What is a “large enough” sample size in SEM? It is impossible to give a single answer because several factors affect sample size requirements:

---

<sup>2</sup>The covariance of a variable with itself is just its variance, such as  $\text{cov}_{XX} = s_x^2$ .

1. More complex models, or those with more parameters, require bigger sample sizes than simpler models with fewer parameters. This is because models with more parameters require more estimates, and larger samples are necessary in order for the computer to estimate the additional parameters with reasonable precision.

2. Analyses in which all outcome variables are continuous and normally distributed, all effects are linear, and there are no interactions require smaller sample sizes compared with analyses in which some outcomes are not continuous or have severely non-normal distributions or there are curvilinear or interactive effects. This issue also speaks to the availability of different estimation methods in SEM, some of which need very large samples because of assumptions they make—or do not make—about the data.

3. Larger sample sizes are needed if score reliability is low; that is, less precise data requires larger samples in order to offset the potential distorting effects of measurement error. Latent variable models can control measurement error better than observed variable models, so fewer cases may be needed when there are multiple indicators for constructs of interest. The amount of missing data also affects sample size requirements. As expected, higher levels of missing data require larger sample sizes in order to compensate for loss of information.

4. There are also special sample size considerations for particular kinds of structural equation models. In factor analysis, for example, larger samples may be needed if there are relatively few indicators per factor, the factors explain unequal proportions of the variance across the indicators, some indicators covary appreciably with multiple factors, the number of factors is increased, or covariances between factors are relatively low.

Given all of these influences, there is no simple rule of thumb about sample size that works across all studies. Also, sample size requirements in SEM can be considered from at least two different perspectives, (1) the number of cases required in order for the results to have adequate statistical precision versus (2) minimum sample sizes needed in order for significance tests in SEM to have reasonable power. Recall that power is the probability of rejecting the null hypothesis in significance testing when the alternative hypothesis is true in the population. Power  $>.85$  or so may be an adequate minimum, but even higher levels may be needed if the consequences of a Type II error, or the failure to reject a false null hypothesis, are serious. Depending on the model and analysis, sample size requirements for adequate statistical power in SEM can be *much* greater than those for adequate statistical precision.

Results of a recent computer simulation (Monte Carlo) study by Wolf, Harrington, Clark, and Miller (2013) illustrate the difficulty with “one-size-fits-all” heuristics about sample size requirements in SEM. These authors studied a relatively small range of structural equation models, including factor analysis models, observed-variable versus latent-variable models of mediation, and single-indicator versus multiple-indicator mea-

surement models. Minimum sample sizes for both precision and power varied widely across the different models and extent of missing data. For example, minimum sample sizes for factor analysis models ranged from 30 to 460 cases, depending on the number of factors (1–3), the number of indicators per factor (3–8), the average correlation between indicators and factors (.50–.80), the magnitude of factor correlations (.30–.50), and the extent of missing data (2–20% per indicator).

In a later chapter I will show you how to estimate minimum sample sizes in power analysis for SEM, but now I want to suggest at least a few rough guidelines about sample size requirements for statistical precision. For latent variable models where all outcomes are continuous and normally distributed and where the estimation method is maximum likelihood—the default method in most SEM computer tools—Jackson (2003) describes the **N:*q* rule**. In this heuristic, Jackson suggested that researchers think about minimum sample sizes in terms of the ratio of the number of cases (*N*) to the number of model parameters that require statistical estimates (*q*). A recommended sample-size-to-parameters ratio would be 20:1. For example, if a total of *q* = 10 parameters require estimates, then a minimum sample size would be 20*q*, or *N* = 200. Less ideal would be an *N:**q* ratio of 10:1, which for the example just given for *q* = 10 would be a minimum sample size of 10*q*, or *N* = 100. As the *N:**q* ratio falls below 10:1 (e.g., *N* = 50 for *q* = 10 for a 5:1 ratio), so does the trustworthiness of the results. The risk for technical problems in the analysis is also greater.

It is even more difficult to suggest a meaningful absolute minimum sample size, but it helps to consider typical sample sizes in SEM studies. A median sample size may be about 200 cases based on reviews of studies in different research areas, including operations management (Shah & Goldstein, 2006) and education and psychology (MacCallum & Austin, 2000). But *N* = 200 may be too small when analyzing a complex model or outcomes with non-normal distributions, using an estimation method other than maximum likelihood, or finding that there are missing data. With *N* < 100, almost any type of SEM may be untenable unless a very simple model is analyzed, but models so basic may be uninteresting. Barrett (2007) suggested that reviewers of journal submissions routinely reject for publication any SEM analysis where *N* < 200 unless the population studied is restricted in size. This recommendation is not standard practice, but it highlights the fact that analyzing small samples in SEM is problematic.

Most published SEM studies are probably based on samples that are too small. For example, Loehlin (2004) said that the results of power analyses in SEM are “frequently sobering” because researchers often learn that their sample sizes are too small for adequate statistical power. Westland (2010) reviewed a total of 74 SEM studies published in four different journals in management information systems. He estimated that (1) the average sample size across these studies, about *N* = 375, was only 50% of the minimum size needed to support the conclusions; (2) the median sample size, about *N* = 260, was only 38% of the minimum required and also reflected substantial negative skew in undersampling; and (3) results in about 80% of all studies were based on insufficient sample sizes. We revisit sample size requirements in later chapters, but many, if not most, published SEM studies are based on samples that are too small.

## LESS EMPHASIS ON SIGNIFICANCE TESTING

A proper role for significance testing in SEM is *much* smaller compared with more standard techniques such as ANOVA and multiple regression. One reason is that SEM features the evaluation of entire models, which brings a higher-level perspective to the analysis. Results of significance testing of individual effects represented in the model may be of interest in some studies, but at some point the researcher must make a decision about the whole model: Should it be rejected?—modified?—if so, how? Thus, there is a sense in SEM that the view of the entire model takes precedence over that of specific effects represented in the model.

Another reason is the large-sample requirement in SEM. It can happen in large samples that a result is “highly significant” (e.g.,  $p < .0001$ ) but trivial in effect size. By the same token, virtually all effects that are not zero will be significant in a sufficiently large sample. In fact, if the sample size is large, then a significant result just basically confirms a large sample, which is a tautology, or a needless repetition of the same sense in different words. But if sample size is too small—which is true in many SEM studies—the power of significance tests may be so low that it is impossible to correctly interpret certain results from significance testing.

A third reason for limiting the role of significance testing in SEM is that observed statistical significance, or  $p$  values, for effects of latent variables are *estimated* by the computer, but this estimate could change if, say, a different estimation method is used or sometimes across different computer programs for the same model and data. Differences in estimated  $p$  values across different software packages are not usually great, but slight differences in  $p$  can make big differences in hypothesis testing, such as  $p = .051$  versus  $p = .049$  for the same effect when testing at the .05 level.

A fourth reason is not specific to SEM, but concerns statistical data analysis in general: Researchers should be more concerned with estimating effect sizes and their precisions than with the outcomes of significance testing (Kline, 2013a). Also, SEM gives better estimates of effect size than traditional techniques for observed variables, including ANOVA and multiple regression. For all these reasons, I agree with the view of Rodgers (2010) that SEM is part of a “quiet methodological revolution” that involves the shift from significance testing about individual effects to the evaluation of whole statistical models.

## SEM AND OTHER STATISTICAL TECHNIQUES

You may know that ANOVA is just a special case of multiple regression. The two techniques are based on the same underlying mathematical model that belongs to a larger family known as the **general linear model** (GLM). The multivariate techniques of MANOVA (i.e., multivariate ANOVA) and canonical variate analysis (canonical correlation), among others, are also part of the GLM. The whole of the GLM can be seen as just a restricted case of SEM for analyzing observed variables (Fan, 1997). So learning about

SEM really means extending your repertoire of data analysis skills to the next level, one that offers even more flexibility than the GLM.

Latent variables analyzed in SEM are assumed to be continuous. Other statistical techniques analyze categorical latent variables, or **classes**, which represent subpopulations where membership is not known but is inferred from the data. Thus, a goal of the analysis is to identify the number and nature of latent classes. The technique of **latent class analysis** is a type of factor analysis but for categorical indicators and latent variables (Hagenaars & McCutcheon, 2002). A special kind of latent class model that represents the shift from one of two different states, such as from nonmastery to mastery of a skill, is a **latent transition model**. In **latent class regression**, a criterion is predicted by estimated class membership and other variables that covary with class membership. In contrast to standard regression analysis of observed variables, it is not assumed in latent class regression that the prediction equation holds for all cases.

Until recently, SEM was generally viewed as a relatively distinct family of techniques from those just mentioned for analyzing categorical latent variables. But this view is changing because of attempts to express latent variable models within a common mathematical framework (Bartholomew, 2002). For example, Muthén (2001) describes the analysis of **mixture models** with latent variables that may be continuous or categorical. When both are present in the same model, the analysis is basically SEM conducted across different inferred subpopulations (classes). Skrondal and Rabe-Hesketh (2004) outline **generalized linear latent and mixed models** (GLAMM), which feature (1) a response model that associates different types of observed variables (continuous, dichotomous, etc.) with latent variables; (2) a structural model where observed or latent variables can be predictors or outcomes; and (3) a distribution model for continuous or discrete latent variables, including nonparametric models. This framework also includes multilevel models for hierarchical data sets.

It is also true that modern SEM computer tools can analyze a greater variety of latent variable models. For example, the Mplus program can analyze all basic kinds of SEM models and mixture models, too. Both kinds of analyses just mentioned can be combined with a multilevel analysis in Mplus. The gsem (generalized SEM) command in Stata can analyze models based on the GLAMM framework. Both Mplus and LISREL have capabilities for analyzing item characteristic curves of the type generated in item response theory (IRT). Computer tools or procedures like those just mentioned blur the distinction between SEM, latent class analysis, multilevel modeling, mixture models analysis, and IRT analysis.

## SEM AND OTHER CAUSAL INFERENCE FRAMEWORKS

The SEM family is related to two other approaches to causal inference, the **potential outcomes model** (POM)—also called the **Neyman–Rubin model** after Jerzy Neyman and Donald Rubin—and Judea Pearl’s **structural causal model** (SCM). Briefly, the POM elaborates on the role of counterfactuals in causal inference (Rubin, 2005). A **counter-**

**factual** is a hypothetical or conditional statement that expresses not what has happened but what could or might happen under different conditions (e.g., “I would not have been late, if I had correctly set the alarm”). There are two basic counterfactuals in treatment outcome studies: (1) what would be the outcomes of control cases, if they were treated; and (2) what would be the outcomes of treated cases, if they were not treated? If each case was either treated or not treated, these potential outcomes are not observed. This means that the observed data—outcomes for treated versus control cases—is a subset of all possible combinations.

The POM is concerned with conditions under which causal effects may be estimated from data that do not include all potential outcomes. In experimental designs, random assignment guarantees over replications the equivalence of the treated and control groups. Thus, any observed difference in outcome over replication studies can be attributed to the treatment. But things are more complicated in quasi-experimental or nonexperimental designs where the assignment mechanism is both nonrandom and unknown. In this case, the average observed difference may be a confounded estimator of treatment effects versus selection factors. In such designs, the POM distinguishes between equations for the observed data versus those for causal parameters. This helps to clarify how estimators based on the data may differ from estimators based on the causal model (MacKinnon, 2008). The POM has been applied many times in randomized clinical trials and in mediation analysis, a topic covered later in this book.

Some authors describe the POM as a more disciplined method for causal inference than SEM (Rubin, 2009), but such claims are problematic for two reasons (Bollen & Pearl, 2013). First, it is possible to express counterfactuals in SEM as predicted values for outcome variables, once we fix values of its causal variables to constants that represent the conditions in counterfactual statements (Kenny, 2014b). This means that various alternative estimators in the POM are available in SEM, too. Second, the POM and SEM are logically equivalent in that a theorem in one framework can be expressed as a theorem in the other. That the two systems encode causal hypotheses in different ways—in SEM as functional relations among observed or latent variables and in the POM as statistical relations among counterfactual (latent) variables—is just a superficial difference (Pearl, 2012).

The SCM is based on graph theory for causal modeling. It originated in the computer science literature with Pearl’s work on Bayesian networks and machine learning. It has been elaborated by Pearl since the 1980s as a method for causal inference that integrates both the POM and SEM into a single comprehensive framework that also extends the capabilities of both (Pearl, 2009b). This is why Hayduk et al. (2003) described the SCM as the future of SEM and also why I introduce readers to it in this book. The SCM is well known in disciplines such as epidemiology, but it is less familiar in areas such as psychology and education. This is unfortunate because the SCM has features described next that make it worthwhile to learn about it.

The SCM is graphical in nature; specifically, causal hypotheses are represented in directed graphs where either observed or latent variables can be depicted. Unlike graphs (model diagrams) in SEM, which are basically static entities that require data in order to

be analyzed, there are special computer tools in the SCM for analyzing directed graphs *without* data. This capability permits the researcher to test his or her *ideas* before collecting the data. For example, the analysis of a directed graph may indicate that particular causal effects cannot be estimated unless additional variables are measured. It is easier to deal with this problem when the study is being planned than after the data are collected. Computer tools for the SCM also help the researcher to find testable implications of the causal hypotheses represented in the graph. No special software is required to analyze the data. This is because testable implications for continuous variables can be evaluated using partial correlations, which can be estimated with standard computer tools for statistical analysis.

The SCM is described in some works using the notation and semantics of mathematics, including theorems, lemmas, and proofs (Pearl, 2009b). This style adds rigor to the presentation, but it can be challenging for beginners. Readers are helped to learn about the SCM in the same pedagogical style used in the rest of this book (i.e., concepts over equations).

## MYTHS ABOUT SEM

Two myths about SEM have already been stated, notably the false belief that SEM analyzes only covariances and the idea that SEM is relevant only in nonexperimental studies. Bollen and Pearl (2013) describe some additional myths that are also easily dismissed. One is the incorrect perception that curvilinear or interactive effects cannot be analyzed in SEM. It is actually no problem to analyze such effects in SEM for either observed or latent variables, but the researcher must make the appropriate specifications (e.g., Equation 1.1) so that the computer can estimate such effects. Another myth is that only continuous outcomes in linear causal models can be analyzed in SEM. There are special estimation methods in contemporary SEM computer programs that analyze noncontinuous outcomes, such as binary or ordinal variables, in models where causal relations are linear or nonlinear. These capabilities are described in later chapters.

Bollen and Pearl (2013) consider other myths about the role of SEM in causal inference. Some of these myths are overly pessimistic evaluations of SEM, such as the view that the POM is inherently superior to SEM for causal inference. Other myths are too optimistic. For example, it is a myth that SEM somehow permits researchers to infer causation from associations (covariances) alone. Nothing could be further from the truth; specifically, SEM is not some magical statistical method that allows researchers to specify a causal model, collect data, tinker with the model until its correspondence with the data is acceptable, and then conclude that the model corresponds to reality. Yes, it is true that some researchers misuse SEM in this way, but inferring causation from correlation requires a principled, disciplined approach that takes account of theory, design, data, replication, and causal assumptions, only some of which are empirically verifiable.

The most that could be concluded in SEM is that the model is consistent with the data, but we cannot generally claim that our model is proven. In this way, SEM is a **disconfirmatory procedure** that can help us to reject false models (those with poor fit to the data), but it does not *confirm* the veracity of the researcher's model. The presence of equivalent or near-equivalent models that explain the data just as well as the researcher's preferred model is one reason. Another is the possibility of specification error, or the representation in the model of false hypotheses. Bollen (1989) put it like this (emphasis in original):

*If a model is consistent with reality, then the data should be consistent with the model. But if the data are consistent with the model, this does not imply that the model corresponds to reality.*  
(p. 68)

Retaining a model in SEM makes causal assumptions more tentatively plausible, but any such conclusions must withstand replication and the criticism of other researchers who suggest alternative models for the same data (Bollen & Pearl, 2013).

A related myth is that SEM and regression techniques are equivalent. There are superficial resemblances in the equations analyzed, but Bollen and Pearl (2013) remind us that SEM and regression are fundamentally different. In regression, the roles of predictor and criterion are theoretically interchangeable. For example, there is no special problem in bivariate regression with specifying  $X$  as a predictor of  $Y$  in one analysis ( $Y$  is regressed on  $X$ ) and then in a second analysis with regressing  $X$  on  $Y$ . There is no such ambiguity in SEM, where the specification that  $X$  affects  $Y$  and not the reverse is a causal link that reflects theory and also depends on other assumptions in the analysis. Thus, the semantics of SEM and regression are distinct. Yes, there are certain types of models in SEM, such as path models, where standard regression techniques can be used to estimate presumed causal effects, but the context of SEM is causal modeling, not mere statistical prediction.

## **WIDESPREAD ENTHUSIASM, BUT WITH A CAUTIONARY TALE**

One cannot deny that SEM is increasingly “popular” among researchers. This is evident by the growing numbers of computer tools for SEM, formal courses at the graduate level, continuing education seminars, and journal articles in which authors describe the results of SEM analyses. It is also difficult to look through an issue of a research journal in psychology, education, or other areas and not find at least one article that concerns SEM.

It is not hard to understand enthusiasm for SEM. As described by David Kenny in the Series Editor Note in the second edition of this book, researchers love SEM because it addresses questions they want answered and it “thinks” about research problems the way that researchers do. But there is evidence that many—if not most—published

reports of the application of SEM have at least one flaw so severe that it compromises the scientific value of the article.

MacCallum and Austin (2000) reviewed about 500 SEM studies in 16 different psychology research journals, and they found problems with the reporting in most studies. For example, in about 50% of the articles, the reporting of parameter estimates was incomplete (e.g., unstandardized estimates omitted); in about 25% the type of data matrix analyzed (e.g., correlation vs. covariance matrix) was not described; and in about 10% the model specified or the indicators of factors were not clearly specified. Shah and Goldstein (2006) reviewed 93 articles in four operations management research journals. In most articles, they found that it was hard to determine the model actually tested or the complete set of observed variables. They also found that the estimation method used was not mentioned in about half of the articles, and in 31 out of 143 studies they reported that the model described in the text did not match the statistical results reported in text or tables.

Nothing in SEM protects against **equivalent models**, which explain the data just as well as the researcher's preferred model but make differing causal claims. The problem of equivalent models is relevant in probably most SEM applications, but most authors of SEM studies do not even mention it (MacCallum & Austin, 2000). Ignoring equivalent models is a serious kind of **confirmation bias** whereby researchers test a single model, give an overly positive evaluation of the model, and fail to consider other explanations of the data (Shah & Goldstein, 2006). The potential for confirmation bias is further strengthened by the lack of replication, a point considered next.

It is rare when SEM analyses are replicated across independent samples either by the same researchers who collected the original data (internal replication) or by other researchers who did not (external replication). The need for large samples in SEM complicates replication, but most of the SEM research literature is made up of one-shot studies that are never replicated. It is critical to eventually replicate a structural equation model if it is ever to represent anything beyond a mere statistical exercise. Kaplan (2009) notes that despite over 40 years of application of SEM in the behavioral sciences, it is rare that results from SEM analyses are used for policy or clinically relevant prediction studies.

The ultimate goal of SEM—or any other type of method for statistical modeling—should be to attain what I call **statistical beauty**, which means that the final retained model (if any)

1. has a clear theoretical rationale (i.e., it makes sense);
2. differentiates between what is known and what is unknown—that is, what is the model's range of convenience, or limits to its generality?—and
3. sets conditions for posing new questions.

That most applications of SEM fall short of these goals should be taken as an incentive by all of us to do better.

## FAMILY HISTORY

Because SEM is a collection of related techniques, it does not have a single source. Part of its origins date to the early years of the 20th century with the development of what we now call exploratory factor analysis, usually credited to Charles Spearman (1904). A few years later, the biogeneticist Sewell Wright (e.g., 1921, 1934) developed the basics of path analysis. Wright demonstrated how observed covariances could be related to the parameters of both direct and indirect effects among a set of observed variables. In doing so, he showed how these effects could be estimated from sample data. Wright also invented path diagrams, or graphical representations of causal hypotheses that we still use to this day. The technique of path analysis was subsequently introduced to the behavioral sciences by various authors, including H. M. Blalock (1961) and O. D. Duncan (1966) in sociology, among others (see Wolfle, 2003).

The measurement (factor analysis) and structural (path analysis) approaches were integrated in the early 1970s in the work of basically three authors: K. G. Jöreskog, J. W. Keesling, and D. Wiley, into a framework that Bentler (1980) called the **JWK model**. One of the first widely available computer programs able to analyze models based on the JWK framework—now called SEM—was LISREL, developed by K. G. Jöreskog and D. Sörbom in the 1970s. The first publicly available version for mainframe computers, LISREL III, was published in 1974, and LISREL has been subsequently updated many times.

The 1980s and 1990s witnessed the development of more computer programs and a rapid expansion of the use of SEM techniques in many different areas of the behavioral sciences. Examples from this time include works about latent variable models of growth and change over time (Duncan, Duncan, Strycker, Li, & Alpert, 1999) and also about the estimation of the curvilinear and interactive effects of latent variables (Schumaker & Marcoulides, 1998). Muthén's (1984) descriptions of methods for ordinal data further extended the range of application of SEM. Another major development concerned the convergence of SEM and techniques for multilevel modeling (Muthén, 1994).

Since the 2000s, there has been a surge of interest in Bayesian methods in the behavioral sciences. Bayesian statistics are a set of methods for the orderly expression and revision of support for hypotheses as new evidence is gathered and combined with extant knowledge. Unlike traditional significance testing, which estimates the probabilities of data under null hypotheses, Bayesian methods can generate estimated probabilities of hypotheses, given the data. They also generally require the researcher to specify the exact forms of the distributions for hypothesized effects (parameters) both prior to synthesizing new data (prior distributions) and after (posterior distributions). Methods in SEM based on Bayesian information criteria, used to compare two different (competing) models for the same observed variables, are covered later in the book. Bayesian capabilities are available in some SEM computer tools, such as Amos and Mplus. See Kaplan and Depaoli (2012) for more information about Bayesian SEM.

Given the history just reviewed, it is safe to say that SEM has come of age; that is, it is a mature set of techniques. And with maturity should come awareness of one's limita-

tions, the motivation to compensate for them, and openness to new perspectives. This is why in later chapters we deal with topics, such as Pearl's SCM, that are not covered in most introductions to SEM. Just as life is a process of continual learning, so is causal modeling.

## SUMMARY

The SEM family of techniques has its origins in regression analyses of observed variables and in factor analyses of latent variables. Essential features include its a priori nature, the potential to explicitly distinguish between observed and latent variables, and the capacity to analyze covariances or means in experimental or nonexperimental designs. More and more researchers are using SEM, but in too many studies there are serious problems with the way SEM is applied or with how the analysis is described or the results are reported. How to avoid getting into trouble with SEM is a major theme in later chapters. Pearl's structural causal model unifies SEM with the potential outcomes model and extends the capabilities of both. It also brings new rigor to causal inference in the behavioral sciences. The ideas introduced in this chapter set the stage for reviewing fundamental principles of regression that underlie SEM in the next chapter.

## LEARN MORE

A concise history of SEM is given in the chapter by Matsueda (2012), and Bollen and Pearl (2013) describe various myths about SEM. Wolfe (2003) traces the introduction of path analysis to the social sciences.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). New York: Springer.

Matsueda, R. L. (2012). Key advances in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 3–16). New York: Guilford Press.

Wolfe, L. M. (2003). The introduction of path analysis to the social sciences, and some emergent themes: An annotated bibliography. *Structural Equation Modeling*, 10, 1–34.

## 2

# Regression Fundamentals

---

Knowing about regression analysis will help you to learn about SEM. Although the techniques considered next analyze observed variables only, their basic principles make up a core part of SEM. This includes the dependence of the results on not only what is measured (the data), but also on what is not measured, or omitted relevant variables, a kind of specification error. Some advice: *Even if you think that you already know a lot about regression, you should nevertheless read this chapter carefully.* This is because many readers tell me that they learned something new after hearing about the issues outlined here. Next I assume that standard deviations ( $SD$ ) for continuous variables are calculated as the square root of the sample variance  $s^2 = SS/df$ , where  $SS$  refers to the sum of squared deviations from the mean and the overall degrees of freedom are  $df = N - 1$ . Standardized scores, or normal deviates, are calculated as  $z = (X - M)/SD$  for a continuous variable  $X$ .

---

## BIVARIATE REGRESSION

Presented in Table 2.1 are scores on three continuous variables. Considered next is bivariate regression for variables  $X$  and  $Y$ , but later we deal with the multiple regression analysis that also includes variable  $W$ . The unstandardized bivariate regression equation for predicting  $Y$  from  $X$ —also called regressing  $Y$  on  $X$ —takes the form

$$\hat{Y} = B_X X + A_X \quad (2.1)$$

where  $\hat{Y}$  refers to predicted scores. Equation 2.1 describes a straight line where  $B_X$ , the unstandardized regression coefficient for predictor  $X$ , is the slope of the line, and  $A_X$  is the constant or intercept term, or the value of  $\hat{Y}$ , if  $X = 0$ . For the data in Table 2.1,

**TABLE 2.1. Example Data Set for Bivariate Regression and Multiple Regression**

Case	X	W	Y	Case	X	W	Y
A	16	48	100	K	18	50	102
B	14	47	92	L	19	51	115
C	16	45	88	M	16	52	92
D	12	45	95	N	16	52	102
E	18	46	98	O	22	50	104
F	18	46	101	P	12	51	85
G	13	47	97	Q	20	54	118
H	16	48	98	R	14	53	105
I	18	49	110	S	21	52	111
J	22	49	124	T	17	53	122

Note.  $M_X = 16.900$ ,  $SD_X = 3.007$ ;  $M_W = 49.400$ ,  $SD_W = 2.817$ ;  $M_Y = 102.950$ ,  $SD_Y = 10.870$ ;  $r_{XY} = .686$ ,  $r_{XW} = .272$ ,  $r_{WY} = .499$ .

$$\hat{Y} = 2.479X + 61.054$$

which says that a 1-point increase in  $X$  predicts an increase in  $Y$  of 2.479 points and that  $\hat{Y} = 61.054$ , given  $X = 0$ . Exercise 1 asks you to calculate these coefficients for the data in Table 2.1.

The predicted scores defined by Equation 2.1 make up a composite, or a weighted linear combination of the predictor and the intercept. The values of  $B_X$  and  $A_X$  in Equation 2.1 are generally estimated with the method of **ordinary least squares** (OLS) so that the **least squares criterion** is satisfied. The latter means that the sum of squared residuals, or  $\sum(Y - \hat{Y})^2$ , is as small as possible in a particular sample. Consequently, OLS estimation capitalizes on chance variation, which implies that values of  $B_X$  and  $A_X$  will vary over samples. As we will see later, capitalization on chance is a greater problem in smaller versus larger samples.

Coefficient  $B_X$  in Equation 2.1 is related to the Pearson correlation  $r_{XY}$  and the standard deviations of  $X$  and  $Y$  as follows:

$$B_X = r_{XY} \left( \frac{SD_Y}{SD_X} \right) \quad (2.2)$$

A formula for  $r_{XY}$  is presented later, but for now we can see in Equation 2.2 that  $B_X$  is just a rearrangement of the expression for the covariance between  $X$  and  $Y$ , or  $\text{cov}_{XY} = r_{XY}SD_XSD_Y$ . Thus,  $B_X$  corresponds to the covariance structure of Equation 2.1. Because  $B_X$  reflects the original metrics of  $X$  and  $Y$ , its value will change if the scale of either variable is altered (e.g.,  $X$  is measured in centimeters instead of inches). For the same reason, values of  $B_X$  are not limited to a particular range. For example, it may be possible to derive

values of  $B_X$  such as  $-7.50$  or  $1,225.80$ , depending on the raw score metrics of  $X$  and  $Y$ . Consequently, a numerical value of  $B_X$  that appears large does not necessarily mean that  $X$  is an important or strong predictor of  $Y$ .

The intercept  $A_X$  of Equation 2.1 is related to both  $B_X$  and the means of both variables:

$$A_X = M_Y - B_X M_X \quad (2.3)$$

The term  $A_X$  represents the mean structure of Equation 2.1 because it conveys information about the means of both variables albeit with a single number. As stated,  $\hat{Y} = A_X$  when  $X = 0$ , but sometimes scores of zero are impossible on certain predictors (e.g., there is no IQ score of zero in conventional metrics for such scores). If so, scores on  $X$  may be **centered**, or converted to mean deviations  $x = X - M_X$ , before analyzing the data. (Scores on  $Y$  are not centered.) Once centered,  $x = 0$  corresponds to a score that equals the mean in the original (uncentered) scores, or  $X = M_X$ . When regressing  $Y$  on  $x$ , the value of the intercept  $A_x$  equals  $\hat{Y}$  when  $x = 0$ ; that is, the intercept is the predicted score on  $Y$  when  $X$  takes its average value. Although centering generally changes the value of the intercept ( $A_X \neq A_x$ ), centering does *not* affect the value of the unstandardized regression coefficient ( $B_X = B_x$ ). Exercise 2 asks you to prove this point for the data in Table 2.1.

Regression residuals, or  $Y - \hat{Y}$ , sum to zero and are uncorrelated with the predictor, or

$$r_{X(Y-\hat{Y})} = 0 \quad (2.4)$$

The equality represented in Equation 2.4 is required in order for the computer to calculate unique values of the regression coefficient and intercept in a particular sample. Conceptually, assuming independence of residuals and predictors, or the **regression rule** (Kenny & Milan, 2012), permits estimation of the explanatory power of the latter (e.g.,  $B_X$  for  $X$  in Equation 2.1) controlling for omitted (unmeasured) predictors. Bollen (1989) referred to this assumption as **pseudo-isolation** of the measured predictor  $X$  from all other unmeasured predictors of  $Y$ . This term describes the essence of statistical control where  $B_X$  is estimated, assuming that  $X$  is unrelated to all other measured or unmeasured predictors of  $Y$ .

The predictor and criterion in bivariate regression are theoretically interchangeable; that is, it is possible to regress  $Y$  on  $X$  or to regress  $X$  on  $Y$  in two separate analyses. Regressing  $X$  on  $Y$  would make less sense if  $X$  were measured before  $Y$  or if  $X$  is known to cause  $Y$ . Otherwise, the roles of predictor and criterion are not fixed in regression. The unstandardized regression equation for regressing  $X$  on  $Y$  is

$$\hat{X} = B_Y Y + A_Y \quad (2.5)$$

where the regression coefficient and intercept in Equation 2.5 are defined, respectively as follows:

$$B_Y = r_{XY} \left( \frac{SD_X}{SD_Y} \right) \quad \text{and} \quad A_Y = M_X - B_Y M_Y \quad (2.6)$$

The expression for  $B_Y$  is nothing more than a different rearrangement of the same covariance, or  $\text{cov}_{XY} = r_{XY} SD_X SD_Y$ , compared with the expression for  $B_X$  (see Equation 2.2). For the data in Table 2.1, the unstandardized regression equation for predicting  $X$  from  $Y$  is

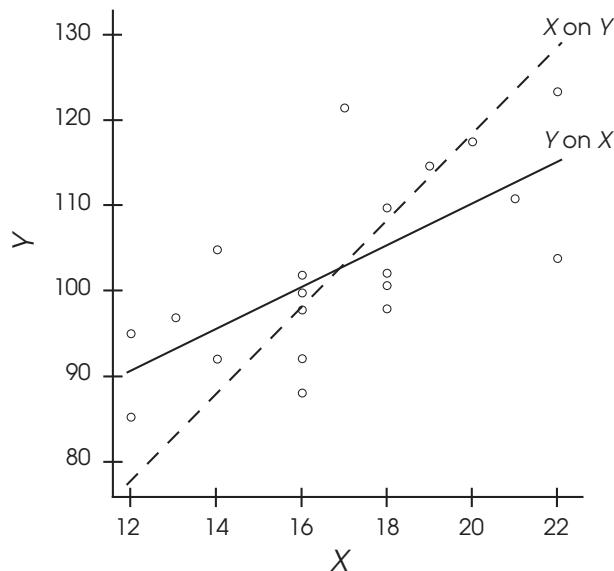
$$\hat{X} = .190 Y - 2.631$$

which says that a 1-point increase in  $Y$  predicts an increase in  $X$  of .190 points and that  $\hat{X} = -2.631$ , given  $Y = 0$ . Presented in Figure 2.1 are the unstandardized equations for regressing  $Y$  on  $X$  and for regressing  $X$  on  $Y$  for the data in Table 2.1. In general, the two possible unstandardized prediction equations in bivariate regression are not identical. This is because the  $Y$ -on- $X$  equation minimizes residuals on  $Y$ , but the  $X$ -on- $Y$  equation minimizes residuals on  $X$ .

The equation for regressing  $Y$  on  $X$  when both variables are standardized (i.e., their scores are normal deviates,  $z$ ) is

$$\hat{z}_Y = r_{XY} z_X \quad (2.7)$$

where  $\hat{z}_Y$  is the predicted standardized score on  $Y$  and the Pearson correlation  $r_{XY}$  is the standardized regression coefficient. There is no intercept or constant term in Equation 2.7 because the means of standardized variables equal zero. (Variances of standardized



**FIGURE 2.1.** Unstandardized prediction lines for regressing  $Y$  on  $X$  and for regressing  $X$  on  $Y$  for the data in Table 2.1.

variables are 1.0.) For the data in Table 2.1,  $r_{XY} = .686$ . Given  $z_X = 1.0$  and  $r_{XY} = .686$ , then  $\hat{z}_Y = .686(1.0)$ , or .686; that is, a score one standard deviation above the mean on  $X$  predicts a score almost seven-tenths of a standard deviation above the mean on  $Y$ . A standardized regression coefficient thus equals the expected difference on  $Y$  in standard deviation units, given an increase on  $X$  of one full standard deviation. Unlike the unstandardized regression coefficient  $B_X$  (see Equation 2.2), the value of the standardized regression coefficient ( $r_{XY}$ ) is unaffected by the scale on either  $X$  or  $Y$ . It is true that (1)  $r_{XY} = .686$  is also the standardized coefficient when regressing  $z_X$  on  $z_Y$ , and (2) the standardized prediction equation in this case is  $\hat{z}_X = r_{XY} z_Y$ .

There is a special relation between  $r_{XY}$  and the unstandardized predicted scores. If  $Y$  is regressed on  $X$ , for example, then

1.  $r_{XY} = r_{Y\hat{Y}}$ ; that is, the bivariate correlation between  $X$  and  $Y$  equals the bivariate correlation between  $Y$  and  $\hat{Y}$ ;
2. the observed variance in  $Y$  can be represented as the exact sum of the variances of the predicted scores and the residuals, or  $s_Y^2 = s_{\hat{Y}}^2 + s_{Y-\hat{Y}}^2$ ; and
3.  $r_{XY}^2 = s_{\hat{Y}}^2 / s_Y^2$ , which says that the squared correlation between  $X$  and  $Y$  equals the ratio of the variance of the predicted scores over the variance of the observed scores on  $Y$ .

The equality just stated is the basis for interpreting squared correlations as proportions of explained variance, and a squared correlation is the **coefficient of determination**. For the data in Table 2.1,  $r_{XY}^2 = .686^2 = .470$ , so we can say that  $X$  explains about 47.0% of the variance in  $Y$ , and vice versa. Exercise 3 asks you to verify the second and third equalities just described for the data in Table 2.1.

When replication data are available, it is actually better to compare unstandardized regression coefficients, such as  $B_X$ , across different samples than to compare standardized regression coefficients, such as  $r_{XY}$ . This is especially true if those samples have different variances on  $X$  or  $Y$ . This is because the correlation  $r_{XY}$  is standardized based on the variability in a particular sample. If variances in a second sample are not the same, then the basis of standardization is not constant over the first and second samples. In contrast, the metric of  $B_X$  is that of the raw scores for variables  $X$  and  $Y$ , and these metrics are constant over samples.

Unstandardized regression coefficients are also better when the scales of all variables are meaningful rather than arbitrary. Suppose that  $Y$  is the time to complete an athletic event and  $X$  is the number of hours spent in training. Assuming a negative covariance, the value of  $B_X$  would indicate the predicted decrease in performance time for every additional hour of training. In contrast, standardized coefficients describe the effect of training on performance in standard deviation units, which discard the original—and meaningful—scales of  $X$  and  $Y$ . The assumptions of bivariate regression are essentially the same as those of multiple regression. They are considered in the next section.

## MULTIPLE REGRESSION

The logic of multiple regression is considered next for the case of two continuous predictors,  $X$  and  $W$ , and a continuous criterion  $Y$ , but the same ideas apply if there are three or more predictors. The form of the unstandardized equation for regressing  $Y$  on both  $X$  and  $W$  is

$$\hat{Y} = B_X X + B_W W + A_{X,W} \quad (2.8)$$

where  $B_X$  and  $B_W$  are the **unstandardized partial regression coefficients** and  $A_{X,W}$  is the intercept. The coefficient  $B_X$  estimates the change in  $Y$ , given a 1-point change in  $X$  while controlling for  $W$ . The coefficient  $B_W$  has the analogous meaning for the other predictor. The intercept  $A_{X,W}$  equals the predicted score on  $Y$  when the scores on *both* predictors are zero, or  $X = W = 0$ . If zero is not a valid score on either predictor, then  $Y$  can be regressed on centered scores ( $x = X - M_X$ ,  $w = W - M_W$ ) instead of the original scores. If so, then  $\hat{Y} = A_{x,w}$ , given  $X = M_X$  and  $W = M_W$ . As in bivariate regression, centering does not affect the values of the regression coefficients for each predictor in Equation 2.8 (i.e.,  $B_X = B_x$ ,  $B_W = B_w$ ).

The overall multiple correlation is actually just the Pearson correlation between the observed and predicted scores on the criterion, or  $R_{YX,W} = r_{Y\hat{Y}}$ . Unlike bivariate correlations, though, the range of  $R$  is 0–1.0. The statistic  $R^2$  equals the proportion of variance explained in  $Y$  by both predictors  $X$  and  $W$ , controlling for their intercorrelation. For the data in Table 2.1, the unstandardized regression equation is

$$\hat{Y} = 2.147X + 1.302W + 2.340$$

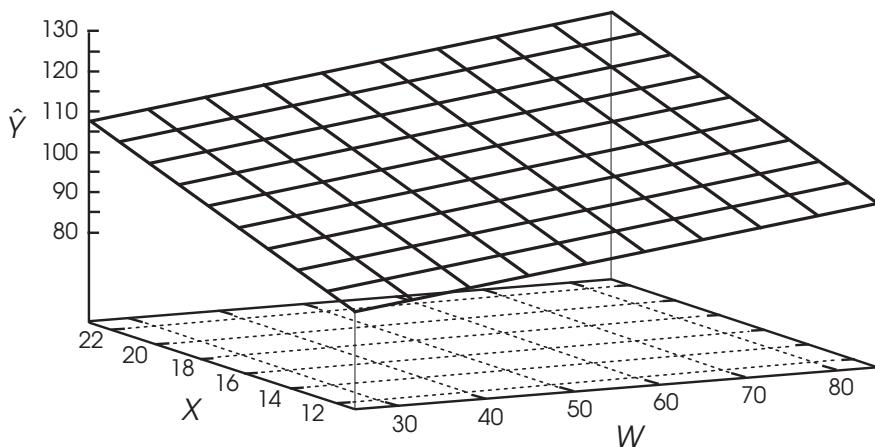
and the multiple correlation equals .759. Given these results, we can say that

1. a 1-point change in  $X$  predicts an increase in  $Y$  of 2.147 points, controlling for  $W$ ;
2. a 1-point change in  $W$  predicts an increase in  $Y$  of 1.302 points, controlling for  $X$ ;
3.  $\hat{Y} = 2.340$ , given  $X = W = 0$ ; and
4. the predictors explain  $.759^2 = .576$ , or about 57.6% of the total variance in  $Y$ , after taking account of their intercorrelation ( $r_{XW} = .272$ ; Table 2.1).

The regression equation just described defines a plane in three dimensions where the slope along the  $X$ -axis is 2.147, the slope along the  $W$ -axis is 1.302, and the  $Y$ -intercept for  $X = W = 0$  is 2.340. This regression surface is plotted in Figure 2.2 over the range of scores in Table 2.1.

Equations for the unstandardized partial regression coefficients for each of two continuous predictors are

$$B_X = b_X \left( \frac{SD_X}{SD_Y} \right) \quad \text{and} \quad B_W = b_W \left( \frac{SD_W}{SD_Y} \right) \quad (2.9)$$



**FIGURE 2.2.** Unstandardized regression surface for predicting  $Y$  from  $X$  and  $W$  for the data in Table 2.1.

where  $b_X$  and  $b_W$  for  $X$  and  $W$  are, respectively, their **standardized partial regression coefficients**, also known as **beta weights**. Their formulas are listed next:

$$b_X = \frac{r_{XY} - r_{WY}r_{XW}}{1 - r_{XW}^2} \quad \text{and} \quad b_W = \frac{r_{WY} - r_{XY}r_{XW}}{1 - r_{XW}^2} \quad (2.10)$$

In the numerators of Equation 2.10, the bivariate correlation of each predictor with the criterion is adjusted for the correlation of the other predictor with the criterion and for correlation between the two predictors. The denominators in Equation 2.10 adjust the total standardized variance by removing the proportion shared by the two predictors. If the values of  $r_{XY}$ ,  $r_{WY}$ , and  $r_{XW}$  vary over samples, then values of coefficients in Equations 2.8–2.10 will also change.

Given three or more predictors, the formulas for the regression coefficients are more complicated but follow the same principles (see Cohen et al., 2003, pp. 636–642). If there is just a single predictor  $X$ , then  $b_X = r_{XY}$ . The intercept in Equation 2.8 can be expressed as a function of the unstandardized partial regression coefficients and the means of all three variables as follows:

$$A_{X,W} = M_Y - B_X M_X - B_W M_W \quad (2.11)$$

The regression equation for standardized variables is

$$\hat{z}_Y = b_X r_{XY} + b_W r_{WY} \quad (2.12)$$

For the data in Table 2.1,  $b_X = .594$ , which says that the difference on  $Y$  is expected to be about .60 standard deviations large, given a difference on  $X$  of one full standard deviation, while we are controlling for  $W$ . The result  $b_W = .337$  has the analogous meaning

except that X is now statistically controlled. Because all variables have the same metric in the standardized solution, we can directly compare values of  $b_X$  with  $b_W$  and correctly infer that the relative predictive power of X is about 1.75 times that of W because the ratio  $.594/.337 = 1.76$ . In general, values of  $b$  can be directly compared across different predictors within the same sample, but unstandardized coefficients ( $B$ ) are preferred for comparing results for the same predictor over different samples.

The statistic  $R_{Y \cdot X, W}^2$  can also be expressed as a function of the beta weights and bivariate correlations of the predictors with the criterion. With two predictors,

$$R_{Y \cdot X, W}^2 = b_X r_{XY} + b_W r_{WY} \quad (2.13)$$

The role of beta weights as corrections for predictor overlap is also apparent in this equation. Specifically, if  $r_{XW} = 0$  (the predictors are independent), then  $b_X = r_{XY}$  and  $b_W = r_{WY}$  (Equation 2.10). This means that  $R_{Y \cdot X, W}^2$  is just the sum of  $r_{XY}^2$  and  $r_{WY}^2$ . But if  $r_{XW} \neq 0$  (the predictors covary), then  $b_X$  and  $b_W$  do not equal the corresponding bivariate correlations and  $R_{Y \cdot X, W}^2$  is not the simple sum of  $r_{XY}^2$  and  $r_{WY}^2$  (it is less). Exercise 4 asks you to verify some of the facts about multiple regression just stated for the data in Table 2.1.

Standard regression analyses do not require raw data files. This is because regression equations and values of  $R^2$  can be calculated from summary statistics (e.g., Equation 2.13), and many regression computer procedures read summary statistics as the input data. For example, the SPSS syntax listed next reads the summary statistics in Table 2.1 and specifies the regression of Y on X and W. Four-decimal accuracy is recommended for matrix input:

```
comment table 2.1, regress y on x, w.  
matrix data variables=x w y/contents=mean sd n corr  
/format=lower nodiagonal.  
begin data  
16.9000 49.4000 102.9500  
3.0070 2.8172 10.8699  
20 20 20  
.2721  
.6858 .4991  
end data.  
regression matrix=in(*)/variables=x w y/dependent=y  
/enter.
```

A drawback to conducting regression analyses with summaries statistics is that residuals cannot be calculated for individual cases.

## Corrections for Bias

The statistic  $R^2$  is a positively biased estimator of  $\rho^2$  (rho-squared), the population proportion of explained variance. The degree of bias is greater in smaller samples or when

the number of predictors is large relative to the number of cases. For example, if  $N = 2$  in bivariate regression and there are no tied scores on  $X$  or  $Y$ , then  $r^2$  must equal 1.0. Now suppose that  $N = 100$  and  $k = 99$ , where  $k$  is the number of predictor variables. With so many predictors—in fact, the maximum number for  $N = 100$ —the value of  $R^2$  must equal 1.0 because there can be no error variance with so many predictors, and this is true even for random numbers.

There are many corrections that downward adjust  $R^2$  as a function of  $N$  and  $k$ . Perhaps the most familiar is Wherry's (1931) equation:

$$\hat{R}^2 = 1 - (1 - R^2) \left( \frac{N-1}{N-k-1} \right) \quad (2.14)$$

where  $\hat{R}^2$  is the **shrinkage-corrected estimate of  $\rho^2$** . In small samples it can happen that  $\hat{R}^2 < 0$ ; if so, then  $\hat{R}^2$  is interpreted as though its value were zero. As the sample size increases for a constant number of predictors, values of  $\hat{R}^2$  and  $R^2$  are increasingly similar, and in very large samples they are essentially equal; that is, it is unnecessary to correct for positive bias in very large samples. Exercise 5 asks you to apply the Wherry correction to the data in Table 2.1.

## Assumptions

The statistical and conceptual assumptions of regression are strict, probably more so than many researchers realize. They are summarized next:

1. *Regression coefficients reflect unconditional linear relations only.* The estimate for  $B_X$  in Equation 2.8 assumes that the linear relation between  $X$  and  $Y$  remains constant over all levels of (a)  $X$  itself, (b) the other measured predictor,  $W$ , and (c) all unmeasured predictors. But if the relation between  $X$  and  $Y$  is appreciably curvilinear or conditional, the value of  $B_X$  could misrepresent predictive power. A conditional relation implies interaction, where the covariance between  $X$  and  $Y$  changes over the levels of at least one other predictor, measured or unmeasured. A curvilinear relation of  $X$  to  $Y$  is also conditional in the sense that the shape of the regression surface changes over the levels of  $X$  (e.g., Figure 1.1). How to represent curvilinear or interactive effects in regression analysis and SEM is considered in Chapter 17.

2. *All predictors are perfectly reliable (no measurement error).* This very strong assumption is necessary because there is no direct way in standard regression analysis to represent less-than-perfect score reliability for the predictors. Consequences of minor violations of this requirement may not be critical, but more serious ones can result in substantial bias. This bias can affect not only the regression weights of predictors measured with error but also those of other predictors. It is difficult to anticipate the direction of this **propagation of measurement error**. Depending on sample intercorrelations, some absolute regression weights may be biased upward (too large), but others may be

biased in the other direction (too small), or **attenuation bias**. There is no requirement that the criterion be measured without error, but the use of a psychometrically deficient measure of it can reduce the value of  $R^2$ . Note that measurement error in the criterion only affects the standardized regression coefficients, not the unstandardized ones. If the predictors are also measured with error, too, then these effects for the criterion could be amplified, diminished, or canceled out, but it is best not to hope for the absence of bias; see McDonald, Behson, and Seifert (2005) for more information about measurement error in regression analysis.

**3. Significance tests in regression assume that the residuals are normally distributed and homoscedastic.** The homoscedastic characteristic means that the residuals have constant variance across all levels of the predictors. Distributions of residuals can be heteroscedastic (the opposite of homoscedastic) or non-normal due to outliers, severe non-normality in the observed scores, more measurement error at some levels of the criterion or predictors, or a specification error. The residuals should always be inspected in regression analyses (see Cohen, Cohen, West, & Aiken, 2003, chap. 4). Reports of regression analyses without comment on the residuals are inadequate. Exercise 6 asks you to inspect the residuals for the multiple regression analysis of the data in Table 2.1. Although there is no requirement in regression for normal distributions of the original scores, values of multiple correlations and absolute partial regression coefficients are reduced if the distributions for a predictor and the criterion have very different shapes, such as very positively skewed on one versus very negatively skewed on the other.

**4. There are no causal effects among the predictors (i.e., there is a single equation).** Because predictors and criteria are theoretically interchangeable in regression, such analyses can be viewed as strictly predictive. But sometimes the analysis is explicitly or implicitly motivated by causal hypotheses, where a researcher views the regression equation as a prototypical causal model with the predictors as causes and the criterion as their outcome (Cohen et al., 2003). If predictors in standard regression analyses are viewed as causal, then we must assume there are no causal effects between them. Specifically, standard regression analyses do not allow for indirect causal effects where one predictor, such as  $X$ , affects another, such as  $W$ , which in turn affects the criterion,  $Y$ . The indirect effect just described would be represented in SEM by the presumed causal order

$$X \rightarrow W \rightarrow Y$$

From a regression perspective, (1) variable  $W$  is both a predictor (of  $Y$ ) and an outcome (of  $X$ ), and (2) there are actually two equations, one for  $W$  another for  $Y$ . But standard regression techniques analyze a single equation at a time, in this case for just  $Y$ , and thus yield estimates of direct effects only. If there are appreciable indirect effects but such effects are not estimated, then estimates of direct effects in standard regression analyses can be very wrong (Achen, 2005). The idea behind this type of bias is elaborated in Chapter 8 which discusses graph theory in causal inference.

5. *There is no specification error.* A few different kinds of potential mistakes involve specification error. These include the failure to estimate the correct functional form of relations between predictors and the criterion, such as assuming unconditional linear effects only when there are sizable curvilinear or interactive effects. Use of the incorrect estimation method is another kind of error. For example, OLS estimation is for continuous criteria, but dichotomous outcomes (e.g., pass-fail) require different methods, such as those used in logistic regression. Including predictors that are irrelevant in the population is a specification error. The concern is that an irrelevant predictor could in a particular sample relate to the criterion by sampling error alone, and this chance covariance may distort values of regression coefficients for other predictors. Omitting from the regression equation predictors that (1) account for some unique proportion of criterion variance and (2) covary with measured predictors is **left-out variables error**, described next.

## LEFT-OUT VARIABLES ERROR

—or more lightheartedly described as the “heartbreak of L.O.V.E.” (Mauro, 1990), this is a potentially serious specification error. As covariances between measured (included) and unmeasured (excluded) predictors increase, results based on the included predictors only tend to become progressively more biased. Suppose that  $r_{XY} = .40$  and  $r_{WY} = .60$  for, respectively, predictors X and W. A researcher measures only X and specifies it as the sole predictor of Y in a bivariate regression. In this analysis for the included predictor, the standardized regression coefficient is  $r_{XY} = .40$ . But if the researcher had the foresight to also measure W, the omitted predictor, and specify it along with X as predictors in a multiple regression analysis (e.g., Equation 2.8), the beta weight for X in this analysis,  $b_X$ , may not equal .40. If not, then  $r_{XY}$  as a standardized regression coefficient with X as the sole predictor does not reflect the true relation of X to Y compared with  $b_X$  derived with both predictors in the equation.

The difference between  $r_{XY}$  and  $b_X$  varies with  $r_{XW}$ , the correlation between the included and omitted predictors. Specifically, if the included and omitted predictors are unrelated ( $r_{XW} = 0$ ), there is no difference, or  $r_{XY} = b_X = .40$  in this example because there is no correction for correlated predictors. Specifically, given

$$r_{XY} = .40, r_{WY} = .60, \text{ and } r_{XW} = 0$$

you can verify, using Equations 2.10 and 2.13, that the multiple regression results with both predictors are

$$b_X = .40, b_W = .60, \text{ and } R^2_{Y,X,W} = .52$$

So we conclude that  $r_{XY} = b_X = .40$  regardless of whether or not W is included in the regression equation, given  $r_{XW} = 0$ .

Now suppose that

$$r_{XY} = .40, r_{WY} = .60, \text{ and } r_{XW} = .60$$

Now we assume that the correlation between the included predictor  $X$  and the omitted predictor  $W$  is .60, not zero. In the bivariate analysis with  $X$  as the sole predictor,  $r_{XY} = .40$  (the same as before), but now the results of the multiple regression analysis are

$$b_X = .06, b_W = .56, \text{ and } R^2_{Y,X,W} = .36$$

Here the value of  $b_X$  is much lower than that of  $r_{XY}$ , respectively, .06 versus .40. This happens because coefficient  $b_X$  controls for  $r_{XW} = .60$ , whereas  $r_{XY}$  does not; thus,  $r_{XY}$  overestimates the relation between  $X$  and  $Y$  compared with  $b_X$ .

Omitting a predictor correlated with others in the equation does not always result in overestimation of the predictive power of an included predictor. For example, if  $X$  is the included predictor and  $W$  is the omitted predictor, it is also possible for the absolute value of  $r_{XY}$  in the bivariate analysis to be less than that of  $b_X$  when both predictors are included in the equation; that is,  $r_{XY}$  underestimates the relation indicated by  $b_X$ . It is also possible for  $r_{XY}$  and  $b_X$  to have different signs. Both cases just mentioned indicate suppression. But overestimation due to omission of a predictor may occur more often than underestimation (suppression). Also, the pattern of bias may be more complicated when there are several omitted variables (e.g., overestimation for some measured predictors, underestimation for others).

Predictors are typically excluded because they are not measured. This means that it is difficult to actually know by how much and in what direction(s) regression coefficients may be biased relative to what their values would be if all relevant predictors were included. But it is unrealistic to expect the researcher to know and be able to measure all relevant predictors. In this sense, all regression equations are probably misspecified to some degree. If omitted predictors are uncorrelated with included predictors, the consequences of left-out variables error may be slight; otherwise, the consequences may be more serious. Careful review of theory and research is the main way to avoid serious specification error by decreasing the potential number of left-out variables.

## SUPPRESSION

Perhaps the most general definition is that suppression occurs when either (1) the absolute value of a predictor's beta weight is greater than that of its bivariate correlation with the criterion or (2) the two have different signs (see also Shieh, 2006). So defined, suppression implies that the estimated relation between a predictor and a criterion while controlling for other predictors is a "surprise," given the bivariate correlations. Suppose that  $X$  is the amount of psychotherapy,  $W$  is the degree of depression, and  $Y$  is the number of prior suicide attempts. The bivariate correlations in a hypothetical sample are

$$r_{XY} = .19, r_{WY} = .49, \text{ and } r_{XW} = .70$$

Based on these results, it might seem that psychotherapy is harmful because of its positive association with suicide attempts ( $r_{XY} = .19$ ). When both predictors (psychotherapy and depression) are analyzed in multiple regression, however, the results are

$$b_X = -.30, b_W = .70, \text{ and } R^2_{Y,X,W} = .29$$

The beta weight for psychotherapy (−.30) has the opposite sign of its bivariate correlation (.19), and the beta weight for depression (.70) exceeds its bivariate correlation (.49).

The results just described are due to controlling for other predictors. Here, people who are more depressed are more likely to be in psychotherapy ( $r_{XW} = .70$ ) and also more likely to try to harm themselves ( $r_{WY} = .49$ ). Correcting for these associations in multiple regression indicates that the relation of psychotherapy to suicide attempts is actually negative once depression is controlled. It is also true that the relation of depression to suicide is even stronger once psychotherapy is controlled. Omit either psychotherapy or depression from the analysis—a left-out variables error—and the bivariate results with the remaining predictor are misleading.

The example just described concerns **negative suppression**, where the predictors have positive bivariate correlations with the criterion and with each other, but one receives a negative beta weight in the multiple regression analysis. A second type is **classical suppression**, where one predictor is uncorrelated with the criterion but receives a nonzero beta weight controlling for another predictor. For example, given the following correlations in a hypothetical sample,

$$r_{XY} = 0, r_{WY} = .60, \text{ and } r_{XW} = .50$$

the results of a multiple regression analysis are

$$b_X = -.40, b_W = .80, \text{ and } R^2_{Y,X,W} = .48$$

This example of classical suppression (i.e.,  $r_{XY} = 0, b_X = -.40$ ) demonstrates that bivariate correlations of zero can mask true predictive relations once other variables are controlled. There is also **reciprocal suppression**, which can occur when two variables correlate positively with the criterion but negatively with each other. Some cases of suppression can be modeled in SEM as the result of inconsistent direct versus indirect effects of causally prior variables on outcome variables. These possibilities are explored later in the book.

## PREDICTOR SELECTION AND ENTRY

An implication of suppression is that predictors should not be selected based on values of bivariate correlations with the criterion. These **zero-order associations** do not con-

trol for other predictors, so their values can be misleading compared with partial regression coefficients for the same variables. For the same reason, whether or not bivariate correlations with the criterion are statistically significant is also irrelevant concerning predictor selection. Although regression computer procedures make it easy to do so, researchers should avoid mindlessly dumping long lists of explanatory variables into regression equations in order to control for their effects (Achen, 2005). The risk is that even small but undetected nonlinearities or indirect effects among predictors can seriously bias partial regression coefficients. It is better to judiciously select the smallest number of predictors—those deemed essential based on extant theory or results of prior empirical studies.

Once selected, there are two basic ways to enter predictors into the equation: One is to enter all predictors at once, or **simultaneous (direct) entry**. The other is to enter them over a series of steps, or **sequential entry**. Entry order can be determined according to one of two different standards, theoretical (rational) or empirical (statistical). The rational standard corresponds to **hierarchical regression**, where you tell the computer a fixed order for entering the predictors. For example, sometimes demographic variables are entered at the first step, and then entered at the second step is a psychological variable of interest. This order not only controls for the demographic variables but also permits evaluation of the predictive power of the psychological variable, over and beyond that of the simple demographic variables. The latter can be estimated as the increase in the squared multiple correlation, or  $\Delta R^2$ , from that of step 1 with demographic predictors only to that of step 2 with all predictors in the equation.

An example of the statistical standard is **stepwise regression**, where the computer selects predictors for entry based solely on statistical significance; that is, which predictor, if entered into the equation, would have the smallest  $p$  value for the test of its partial regression coefficient? After selection, predictors at a later step can be removed from the equation according to  $p$  values (e.g., if  $p \geq .05$  for a predictor in the equation at a particular step). The stepwise process stops when there could be no statistically significant  $\Delta R^2$  by adding more predictors. Variations on stepwise regression include **forward inclusion**, where selected predictors are not later removed from the equation, and **backward elimination**, which begins with all predictors in the equation and then automatically removes them, but such methods are directed by the computer, not you.

Problems of stepwise and related methods are so severe that they are actually banned in some journals (Thompson, 1995), and for good reasons, too. One problem is extreme capitalization on chance. Because every result in these methods is determined by  $p$  values in a particular sample, the findings are unlikely to replicate. Another problem is that not all stepwise regression procedures report  $p$  values that are corrected for the total number of variables that were considered for inclusion. Consequently,  $p$  values in stepwise computer output are generally too low, and absolute values of test statistics are too high; that is, the computer's choices could actually be wrong. Even worse, such methods give the false impression that the researcher does not have to think about predictor selection. Stepwise and related methods are anachronisms in modern data analysis. Said more plainly, death to stepwise regression, think for yourself—see Whittingham, Stephens, Bradbury, and Freckleton (2006) for more information.

Once a final set of rationally selected predictors has been entered into the equation, they should *not* be subsequently removed if their regression coefficients are not statistically significant. To paraphrase Loehlin (2004), the researcher should *not* feel compelled to drop every predictor that is not significant. In smaller samples, the power of significance tests may be low, and removing a nonsignificant predictor can substantially alter the solution. If you had good reason for including a predictor, then it is better to leave it in the equation until replication indicates that the predictor does not appreciably relate to the criterion.

## PARTIAL AND PART CORRELATION

The concept of partial correlation concerns the idea of **spuriousness**: If the observed relation between two variables is wholly due to one or more common cause(s), their association is spurious. Consider these bivariate correlations between vocabulary breadth ( $Y$ ), foot length ( $X$ ), and age ( $W$ ) in a hypothetical sample of elementary school children:

$$r_{XY} = .50, r_{WY} = .60, \text{ and } r_{XW} = .80$$

Although the correlation between foot length  $X$  and vocabulary breadth  $Y$  is fairly substantial (.50), it is hardly surprising because both are caused by a third variable, age  $W$  (i.e., maturation).

The **first-order partial correlation**  $r_{XYW}$  removes the influence of a third variable  $W$  from both  $X$  and  $Y$ . The formula is

$$r_{XYW} = \frac{r_{XY} - r_{XW} r_{WY}}{\sqrt{(1 - r_{XW}^2)(1 - r_{WY}^2)}} \quad (2.15)$$

Applied to the hypothetical correlations just listed, the partial correlation between foot length and vocabulary breadth controlling for age is  $r_{XYW} = .043$ . (You should verify this result.) Because the association between  $X$  and  $Y$  disappears when  $W$  is controlled, their bivariate relation may be spurious. Presumed spurious associations due to common causes are readily represented in SEM.

Equation 2.15 for partial correlation can be extended to control for two or more external variables. For example, the **second-order partial correlation**  $r_{XYWZ}$  estimates the association between  $X$  and  $Y$  controlling for both  $W$  and  $Z$ . There is a related coefficient called **part correlation** or **semipartial correlation** that controls for external variables out of either of two other variables, but not both. The formula for the **first-order part correlation**  $r_{Y(X,W)}$ , for which the association between  $X$  and  $W$  is controlled but not for the association between  $Y$  and  $W$ , is

$$r_{Y(X,W)} = \frac{r_{XY} - r_{WY} r_{XW}}{\sqrt{1 - r_{XW}^2}} \quad (2.16)$$

Given the same bivariate correlations among these three variables reported earlier, the part correlation between vocabulary breadth ( $Y$ ) and foot length ( $X$ ) controlling only foot length for age ( $W$ ) is  $r_{Y(X,W)} = .033$ . This result (.033) is somewhat smaller than the partial correlation for these data, or  $r_{XYW} = .043$ . In general,  $r_{XYW} \geq r_{Y(X,W)}$ ; if  $r_{XW} = 0$ , then  $r_{XYW} = r_{Y(X,W)}$ .

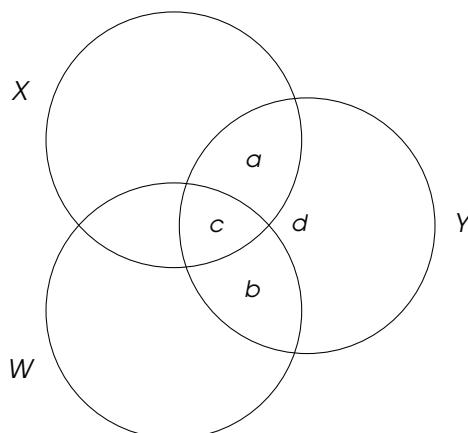
Relations among the squares of the various correlations just described can be illustrated with a Venn-type diagram like the one in Figure 2.3. The circles represent total standardized variances of the criterion  $Y$  and predictors  $X$  and  $W$ . The regions in the figure labeled  $a-d$  make up the total standardized variance of  $Y$ , so

$$a + b + c + d = 1.0$$

Areas  $a$  and  $b$  represent the proportions of variance in  $Y$  uniquely explained by, respectively,  $X$  and  $W$ , but area  $c$  represents the simultaneous overlap (redundancy) of the predictors with the criterion.<sup>1</sup> Area  $d$  represents the proportion of unexplained variance. The squared bivariate correlations of the predictors with the criterion and the overall squared multiple correlation can be expressed as sums of the areas  $a$ ,  $b$ ,  $c$ , or  $d$  in Figure 2.3, as follows:

$$r_{XY}^2 = a + c \quad \text{and} \quad r_{WY}^2 = b + c$$

$$R_{Y \cdot X, W}^2 = a + b + c = 1.0 - d$$



**FIGURE 2.3.** Venn diagram for the standardized variances of predictors  $X$  and  $W$  and criterion  $Y$ .

<sup>1</sup>Note that interpretation of the area  $c$  in Figure 2.3 as a proportion of variance generally holds when all bivariate correlations are positive and there is no suppression. Otherwise, the value  $c$  can be a negative, but there is no such thing as a negative proportion of variance.

The squared part correlations match up directly with the unique areas  $a$  and  $b$  in Figure 2.3. Each of these areas also equals the *increase* in the total proportion of explained variance that occurs by adding a second predictor to the equation (i.e.,  $\Delta R^2$ ); that is,

$$r_{Y(X,W)}^2 = a = R_{Y \cdot X, W}^2 - r_{WY}^2 \quad (2.17)$$

$$r_{Y(W \cdot X)}^2 = b = R_{Y \cdot X, W}^2 - r_{XY}^2$$

The squared partial correlations correspond to areas  $a$ ,  $b$ , and  $d$  in Figure 2.3, and each estimates the proportion of variance in the criterion explained by one predictor but not the other. The formulas are

$$\begin{aligned} r_{XY \cdot W}^2 &= \frac{a}{a+d} = \frac{R_{Y \cdot X, W}^2 - r_{WY}^2}{1 - r_{WY}^2} \\ r_{WY \cdot X}^2 &= \frac{b}{b+d} = \frac{R_{Y \cdot X, W}^2 - r_{XY}^2}{1 - r_{XY}^2} \end{aligned} \quad (2.18)$$

For the data in Table 2.1,  $r_{Y(X,W)}^2 = .327$  and  $r_{XY \cdot W}^2 = .435$ . In words, predictor  $X$  uniquely explains .327, or 32.7% of the total variance of  $Y$  (squared part correlation). Of the variance in  $Y$  not already explained by  $W$ , predictor  $X$  accounts for .435, or 43.5% of the remaining variance (squared partial correlation). Exercise 7 asks you to calculate and interpret the corresponding results for the other predictor,  $W$ , and the same data.

When predictors are correlated—which is just about always—beta weights, partial correlations, and part correlations are alternative ways to describe in standardized terms the relative explanatory power of each predictor controlling for the rest. None is more “correct” than the others because each gives a different perspective on the same data. Note that unstandardized regression coefficients ( $B$ ) are preferred when comparing results for the same predictors and criterion across different samples.

## OBSERVED VERSUS ESTIMATED CORRELATIONS

The Pearson correlation estimates the degree of linear association between two continuous variables. Its equation is

$$r_{XY} = \frac{\text{cov}_{XY}}{SD_X SD_Y} = \frac{\sum_{i=1}^N z_{X_i} z_{Y_i}}{df} \quad (2.19)$$

where  $df = N - 1$ . Rodgers and Nicewander (1988) described a total of 11 other formulas, each of which represents a different conceptual or computational definition of  $r$ , but all of which yield the same result for the same data.

A continuous variable is one for which, theoretically, any value is possible within the limits of its score range. This includes values with decimals, such as 3.75 seconds or 13.60 kilograms. In practice, variables with a range of at least 15 points or so are usually considered as continuous even if their scores are discrete, or integers only (e.g., scores of 10, 11, 12, etc.). For example, the PRELIS program of LISREL—used for data preparation—automatically classifies a variable with less than 16 levels as ordinal.

The statistic  $r$  has a theoretical maximum absolute value of 1.0. But the practical upper limit for  $|r|$  is  $< 1.0$  if the relation between  $X$  and  $Y$  is not unconditionally linear, there is measurement error in either  $X$  or  $Y$ , or distributions for  $X$  versus  $Y$  have different shapes. The amount of variation in samples (i.e.,  $SD_X$  and  $SD_Y$  in Equation 2.19) also affects the value of  $r$ . In general, restriction of range on either  $X$  or  $Y$  through sampling or case selection (e.g., only cases with higher scores on  $X$  are studied) tends to reduce values of  $|r|$ , but not always (see Huck, 1992). The presence of outliers, or extreme scores, can also distort the value of  $r$ ; see Goodwin and Leech (2006) for more information.

There are other forms of the Pearson correlation for observed variables that are either natural dichotomies, such as male versus female for chromosomal sex, or ordinal (ranks). For example:

1. The **point-biserial correlation** ( $r_{pb}$ ) estimates the association between a dichotomy and a continuous variable (e.g., treatment vs. control, weight).
2. The **phi coefficient** ( $\hat{\phi}$ ) is for two dichotomies (e.g., treatment vs. control, survived vs. died).
3. **Spearman's rank order correlation** or **Spearman's rho** ( $\hat{\rho}$ ) is for two ranked variables (e.g., finish order in a race, rank by amount of training time).

Computational formulas for all these special forms are just rearrangements of Equation 2.19 for  $r$  (e.g., Kline, 2013a, pp. 138, 166).

All forms of the Pearson correlation estimate associations between observed (measured) variables. Other, non-Pearson correlations assume that the underlying, or latent, variables are continuous and normally distributed. For example:

1. The **biserial correlation** ( $r_{bis}$ ) is for a naturally continuous variable, such as weight, and a dichotomy, such as recovered—not recovered, that theoretically represents a dichotomized continuous latent variable. For example, presumably degrees of recovery were collapsed when the observed dichotomy was created. The value of  $r_{bis}$  estimates what the Pearson  $r$  would be if the dichotomized variable were continuous and normally distributed.
2. The **polyserial correlation** is the generalization of  $r_{bis}$  that does basically the same thing for a naturally continuous variable and a theoretically continuous-

but-polytomized variable (i.e., categorized into three or more levels). Likert-type response scales for survey or questionnaire items, such as *agree*, *undecided*, or *disagree*, are examples of a polytomized response continuum about the degree of agreement.

3. The **tetrachoric correlation** ( $r_{tet}$ ) for two dichotomized variables estimates what  $r$  would be if both measured variables were continuous and normally distributed.
4. The **polychoric coefficient** is the generalization of the tetrachoric correlation that estimates  $r$  but for ordinal observed variables with two or more levels.

Computing polyserial or polychoric correlations is complicated and requires special software, such as PRELIS in LISREL. These programs generally use a form of maximum likelihood estimation that assumes normality of the latent continuous variables, and error variance tends to increase rapidly as the number of categories on the observed variables decreases from about five to two; that is, dichotomized continuous variables generate the greatest imprecision.

The PRELIS program can also analyze **censored variables**, for which values occur outside of the range of measurement. Suppose that a scale registers values of weight between 1 and 300 pounds only. For objects that weigh either less than 1 pound or more than 300 pounds, the scale tells us only that the measured weight is, respectively, at most 1 pound or at least 300 pounds. In this example, the hypothetical scale is both left censored and right censored because the values less than 1 or more than 300 are not registered on the scale. There are other possibilities for censoring, but scores on censored variables are either exactly known (e.g., weight = 250) or partially known in that they fall within an interval (e.g., weight  $\geq 300$ ). The technique of **censored regression**, better known in economics than in the behavioral sciences, analyzes censored outcomes.

In SEM, Pearson correlations are normally analyzed as part of analyzing covariances when outcome variables are continuous. But noncontinuous outcome variables can be analyzed in SEM, too. One option is to calculate polyserial or polychoric correlations from the raw data and then fit the model to these predicted Pearson correlations. Special methods for analyzing noncontinuous variables in SEM are considered later in Chapters 13 and 16.

In both regression and SEM, it is generally a bad idea to categorize predictors or outcomes that are continuous in order to form **pseudo-groups** (e.g., “low” vs. “high” based on a mean split). Categorization not only discards numerical information about individual differences in the original distribution but it also tends to reduce absolute values of sample correlations when population distributions are normal. The degree of this reduction is greater as the cutting point moves further away from the mean. But if population correlations are low and the sample size is small, then categorization can actually increase absolute sample correlations. Categorization can also create artifactual main or interactive effects, especially when cutting points are arbitrary. In general, it is

better to analyze continuous variables as they are and without categorizing them—see Royston, Altman, and Sauerbrei (2006) for more information.

## LOGISTIC REGRESSION AND PROBIT REGRESSION

Some options to analyze dichotomous outcomes in SEM are based on **logistic regression**. Just as in standard multiple regression, the predictors in logistic regression can be either continuous or categorical. But the prediction equation in logistic regression is a **logistic function**, or a sigmoid function with an “S” shape. It is a type of **link function**, or a transformation that relates the observed outcomes to the predicted outcomes in a regression analysis. Each method of regression has its own special kind of link function. In standard multiple regression with continuous variables, the link function is the **identity link**, which says that observed scores on the criterion  $Y$  are in the same units as  $\hat{Y}$ , the predicted scores (e.g., Figure 2.1). For noncontinuous outcomes, though, original and predicted scores are in different metrics.

Suppose that a total of 32 patients with the same disorder are administered a daily treatment for a varying number of days (5–60). After treatment, the patients are rated as recovered (1) or not recovered (0). Presented in Table 2.2 are the hypothetical raw data for this example. I used Statgraphics Centurion (StatPoint Technologies, 1982–2013)<sup>2</sup> to plot the logistic function with 95% confidence limits for these data that is presented in Figure 2.4. This function generates  $\hat{\pi}$ , the predicted probability of recovery, given the number of days treated,  $X$ . The confidence limits for these predictions are so wide because the sample size is small (see the figure). Because predicted probabilities are estimated from the data, they correspond to a latent continuous variable, and in this sense logistic regression (and probit regression, too) can be seen as a latent variable technique.

The estimation method in logistic regression is not OLS. Instead, it is usually a form of maximum likelihood estimation that is applied after transforming the dichotomous outcome variable into a **logit**, which is the natural logarithm (i.e., natural base  $e$ , or about 2.7183) of the **odds** of the target outcome,  $\hat{\omega}$ . The quantity  $\hat{\omega}$  is the ratio of the probability for the target event, such as recovered, over the probability for the other event, such as not recovered. Suppose that 60% of patients recover after treatment, but the rest, or 40%, do not recover, or

$$\hat{\pi} = .60 \text{ and } 1 - \hat{\pi} = .40$$

The odds of recovery are thus  $\hat{\omega} = .60/.40$ , or 1.50; that is, the odds are 3:2 in favor of recovery. Odds are converted back to probabilities by dividing the odds by 1.0 plus the odds. For example,  $\hat{\omega} = 1.50$ , so  $\hat{\pi} = 1.50/2.50 = .60$ , which is the probability of recovery.

Coefficients for predictors in logistic regression are calculated by the computer in a log metric, but each coefficient can be converted to an **odds ratio**, which estimates

---

<sup>2</sup>[www.statgraphics.com/downloads.htm](http://www.statgraphics.com/downloads.htm)

**TABLE 2.2. Example Data Set for Logistic Regression and Probit Regression**

Status	<i>n</i>	Number of days in treatment ( <i>X</i> )
Not recovered ( <i>Y</i> = 0)	16	6, 7, 9, 10, 11, 13, 15, 16, 18, 19, 23, 25, 26, 28, 30, 32
Recovered ( <i>Y</i> = 1)	16	27, 30, 33, 35, 36, 39, 41, 42, 44, 46, 47, 49, 51, 53, 55, 56

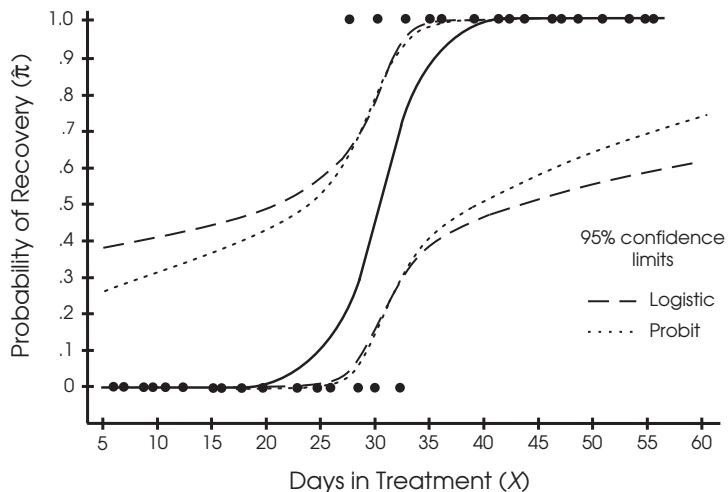
the difference in the odds of the target outcome, given a 1-point increase in the predictor, controlling for all other predictors. I submitted the data in Table 2.2 to the Logistic Regression procedure in Statgraphics Centurion. The prediction equation in a log metric is

$$\text{logit}(\hat{\pi}) = \ln\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = \ln(\hat{\omega}) = .455X - 13.701$$

where .455 is the coefficient for the predictor *X*, number of treatment days, and -13.701 is the intercept. Taking the antilogarithm of the coefficient for days in treatment, or

$$\ln^{-1}(.455) = e^{.455} = 1.576$$

gives us the odds ratio, or 1.576. This result says that for each additional day of treatment, the odds for recovery increase by 57.6%. But this rate of increase is not linear; instead, the rate at which a logistic curve ascends or descends changes according to



**FIGURE 2.4.** Predicted probability of recovery with 95% confidence limits for the data in Table 2.2.

values of the predictor. For these data, the greatest rate of change in predicted recovery occurs between 30 and 40 days of treatment. But at the extremes ( $X < 30$  or  $X > 40$ ), the rate of change in the probability of recovery is much less—see Figure 2.4. The inverse logit function presented next generates the logistic curve plotted in the figure:

$$\hat{\pi} = \text{logit}^{-1}(.455X - 13.701) = \frac{e^{.455X - 13.701}}{1 + e^{.455X - 13.701}}$$

An alternative method is **probit regression**, which analyzes binary outcomes in terms of a **probit function**, where probit stands for “probability unit.” A probit model assumes that the observed dichotomy  $Y = 1$  for the target outcome versus  $Y = 0$  for other events is determined by a normal continuous latent variable  $Y^*$  with a mean of zero and variance of 1.0 such that

$$Y = \begin{cases} 1 & \text{if } Y^* \geq 0 \\ 0 & \text{if } Y^* < 0 \end{cases} \quad (2.20)$$

The equation in probit regression generates  $\hat{Y}^*$  in the metric of normal deviates ( $z$  scores). Next, the computer uses the equation for the cumulative distribution function of the normal curve ( $\Phi$ ) to calculate predicted probabilities of the target outcome  $\hat{\pi}$  from values of  $\hat{Y}^*$  for each case:

$$\hat{\pi} = \Phi(\hat{Y}^*) \quad (2.21)$$

Equation 2.21 is known as the **normal ogive model**.<sup>3</sup>

I analyzed the data in Table 2.2 using the Probit Analysis procedure in Statgraphics Centurion. The prediction equation is

$$\hat{Y}^* = .268X - 8.072$$

The coefficient for  $X$ , .268, estimates in standard deviation units the amount of change in recovery, given a one-day increase in treatment. That is, the  $z$ -score for recovery increases by .268 for each additional day of treatment. Again, this rate of change is not constant because the overall relation is nonlinear (Figure 2.4). Predicted probabilities of recovery for this example are generated by the probit function

$$\hat{\pi} = \Phi(.268X - 8.072)$$

The 95% confidence limits for the probit function are somewhat different than those for the logistic function for the data in Table 2.2—see Figure 2.4.

---

<sup>3</sup>You can see the equation for  $\Phi$  at [http://en.wikipedia.org/wiki/Normal\\_distribution](http://en.wikipedia.org/wiki/Normal_distribution)

Logistic regression and probit regression applied in the same large samples tend to give similar results but in different metrics for the coefficients. The scaling factor that converts results from the logistic model to the same metric as the normal ogive (probit) model is approximately 1.7. For example, the ratio of the coefficients for the predictor in, respectively, the logistic and probit analyses of the data in Table 2.2 is  $.455/.268 = 1.698$ , or 1.7 at single-decimal accuracy. The two procedures may generate appreciably different results if there are many cases at the extremes (predicted probabilities are close to either 0 to 1.0) or if the sample is small. Probit regression is more computationally intensive than logistic regression, but this difference is relatively unimportant for modern microcomputers with fast processors and ample memory. It can happen that computer procedures for probit regression may fail to generate a solution in smaller samples. Agresti (2007) describes additional techniques for categorical data.

## SUMMARY

You should know about regression analysis before learning the basics of SEM. For both sets of techniques, the results are affected not only by what is measured (i.e., the data) but also by what is not measured, especially if omitted predictors covary with included predictors, which is a specification error. Accordingly, you should carefully select predictors after review of theory and results of prior studies in the area. In regression, those predictors should have adequate psychometric characteristics because there is no allowance for measurement error. The same restriction does not apply in SEM, but use of grossly inadequate measures in SEM can seriously bias the results, too. When selecting predictors, the role of judgment should be greater than that of significance testing, which capitalizes on sample-specific variation. The role of significance testing and the technique of bootstrapping in SEM are considered in the next chapter.

## LEARN MORE

The book by Cohen, Cohen, West, and Aiken (2003) is considered by many as a kind of "bible" for multiple regression. Royston, Altman, and Sauerbrei (2006) explain why categorizing predictor or outcome variables is a bad idea. Shieh (2006) describes suppression in more detail.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York: Routledge.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25, 127–141.

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and Psychological Measurement*, 66, 435–447.

**EXERCISES**

All questions concern the data in Table 2.1.

1. Calculate the unstandardized regression equation for predicting  $Y$  from  $X$  based on the descriptive statistics.
2. Show that centering scores on  $X$  does not change the value of the unstandardized regression coefficient for predicting  $Y$  but does affect the value of the intercept.
3. Show that  $s_Y^2 = s_{\hat{Y}}^2 + s_{Y-\hat{Y}}^2$  and  $r_{XY}^2 = s_{\hat{Y}}^2 / s_Y^2$  when  $X$  is the only predictor of  $Y$ .
4. Calculate the unstandardized regression equation and the standardized regression equation for predicting  $Y$  from both  $X$  and  $W$ . Also calculate  $R_{Y \cdot X, W}^2$ .
5. Calculate  $\hat{R}_{Y \cdot X, W}^2$ .
6. Construct a histogram of the residuals for the regression of  $Y$  on both  $X$  and  $W$ .
7. Compute and interpret  $r_{WY \cdot X}^2$  and  $r_{Y(X \cdot W)}^2$ .

# 3

## Significance Testing and Bootstrapping

---

This chapter addresses statistical significance testing and the technique of bootstrapping, with special attention to their roles in SEM. Significance testing has become increasingly controversial over the years. This is due both to the inherent limitations of significance testing and to the failure of most researchers to understand what statistical significance means. Estimation of confidence intervals (interval estimation) as an alternative to significance testing is described. Two different methods for calculating confidence intervals for statistics with complex distributions are outlined: noncentrality interval estimation and bootstrapping. Some types of fit statistics in SEM are distributed as noncentral test distributions, and bootstrapping is a computer-based resampling procedure with application in SEM.

---

### STANDARD ERRORS

The standard error is a standard deviation in a **sampling distribution**, the probability distribution for a sample statistic based on all possible random samples selected from the same population and each based on the same  $N$ . The standard error estimates **sampling error**, or the difference between sample statistics and the corresponding population parameter. Given constant variability among cases, standard error varies inversely with  $N$ . This means that distributions of statistics from larger samples are generally narrower (less variable) than distributions of the same statistic from smaller samples.

There are textbook formulas for standard errors of statistics with simple distributions. By “simple” I mean that (1) the statistic estimates a single parameter and (2) the basic shape of its distribution does not change as a function of that parameter. For example, means have simple distributions, and the equation for their standard error is

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \quad (3.1)$$

where  $\sigma$  is the population standard deviation among cases. Given  $\sigma$ , the value of  $\sigma_M$  decreases as  $N$  increases (see Figure 3.1). An original normal distribution along with two different sampling distributions of means for  $N = 5$  and  $N = 25$  are depicted. There is greater variation of sample means around the population mean  $\mu$  when the sample size is smaller. The value of  $\sigma_M$  must be estimated, if  $\sigma$  is unknown. The estimator is

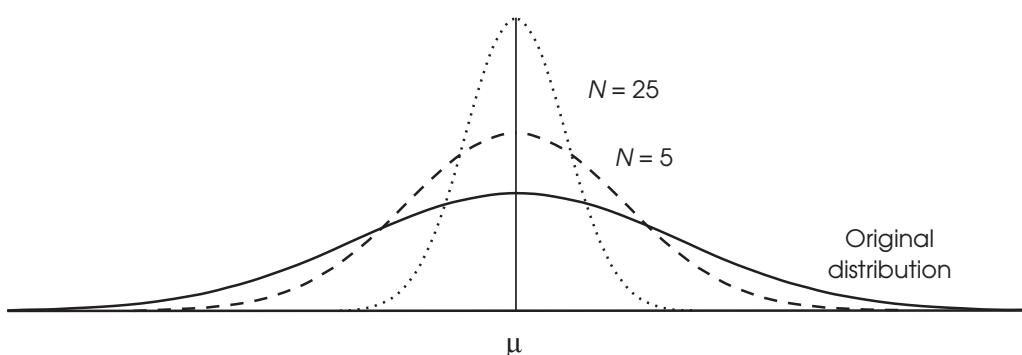
$$SE_M = \frac{SD}{\sqrt{N}} \quad (3.2)$$

Note that  $SE_M$  itself has a standard error. This is because the value of  $SE_M$  will vary over random samples drawn from the same population.

Standard errors for statistics from observed variables estimate sampling error under the exacting assumptions stated next:

1. The method of sampling is random, or at least haphazard enough to generate representative samples over replications.
2. There is no other source of error besides sampling error.
3. Standard errors for parametric statistics often assume normality or homoscedasticity.

The problem with the assumptions just stated is that they are false in most studies. For example, true random sampling requires a list of all members in a population, but such lists are rare. Most samples in human research are ad hoc (convenience) samples made up of participants who happen to be available. What standard errors measure in such samples is generally unknown. Scores are affected by multiple types of error, including sampling error, measurement error, and, in treatment outcome studies, implementation error, or deviations from a treatment protocol, such as due to poor



**FIGURE 3.1.** An original distribution of scores and two distributions of random sample means each based on different sample sizes,  $N = 5$  and  $N = 25$ .

patient or therapist compliance. Other types of error include specification error, such as left-out variables error, and a host of threats to internal validity (e.g., confounding), external validity (e.g., interference due to multiple treatments), and construct validity (e.g., scores are not reliable) (Shadish, Cook, & Campbell, 2001). But standard errors generally assume that the scores are perfect in every way except for the vagaries of random sampling.

The normality assumption refers to population distributions, but normal distributions in actual studies are rare. Many, if not most, empirical distributions are not even symmetrical, much less normal, and departures from normality are often strikingly large (Micceri, 1989). Geary (1947, p. 214) suggested that the disclaimer, “Normality is a myth; there never was, and never will be, a normal distribution,” should appear in all statistics textbooks. Ratios across different groups of largest to smallest variances as large as 8:1 are not uncommon (Keselman et al., 1998), so perhaps homoscedasticity is a myth, too. Even small departures from distributional assumptions can appreciably distort standard errors in small or unrepresentative samples. There are robust estimators with fewer distributional assumptions (Erceg-Hurn & Mirosevich, 2008), but their standard errors assume random sampling, too.

## CRITICAL RATIOS

The basic form of a significance test is the **critical ratio**, the ratio of a statistic over its standard error. Assuming large samples and normality, a critical ratio is interpreted as a deviate in a normal curve ( $z$ ) with a mean of zero and a standard deviation that equals the standard error. A heuristic is that if  $|z| > 2.00$ , the null hypothesis ( $H_0$ ) that the corresponding parameter is zero is rejected at the .05 level ( $p < .05$ ) for a two-tailed test ( $H_1$ ). The precise value of  $|z|$  for the .05 level is 1.96, and for the .01 level it is 2.58. For example, given

$$M = 5.00, SD = 25.00, N = 100, H_0: \mu = 0, \text{ and } H_1: \mu \neq 0,$$

$$SE_M = \frac{25.00}{\sqrt{100}} = 2.50 \quad \text{and} \quad z = \frac{5.00}{2.50} = 2.00$$

For  $z = 2.00$  and assuming random sampling and no other error besides sampling error,  $p = .046$ , so the null hypothesis that the population mean is zero is rejected at the .05 level.

In small samples, the ratio  $M/SE_M$  approximates a  $t$  distribution, which necessitates the use of special tables to determine the critical values of  $t$  for the .05 or .01 levels.<sup>1</sup> These distributions are **central t distributions** where the null hypothesis is assumed to be true. Such distributions have a single parameter,  $df$ , the degrees of freedom, which are

---

<sup>1</sup>Within large samples,  $t$  and  $z$  for the same statistic are essentially equal, and their values are asymptotic in very large samples.

$N - 1$  for a single mean. Other forms of the  $t$  test for means have different  $df$  values. For instance,  $df = N - 2$ , where  $N$  is the total number of cases when means from two independent samples are compared, but all central  $t$  distributions assume a true null hypothesis. There are central distributions for other test statistics, such as  $F$  or  $\chi^2$ , and tables of critical values for these familiar test statistics can be found in many statistics textbooks.

In some SEM computer programs, standard errors are calculated for the unstandardized solution only. You can see this fact when you look through the computer output and find no standard errors printed for standardized estimates. This means that results of significance tests ( $z$ ) are available only for the unstandardized estimates. Researchers often assume that  $p$  values for unstandardized estimates also apply to the corresponding standardized estimates. For samples that are large and representative, this assumption may not be unreasonable. But you should know that the  $p$  value for an unstandardized estimate does not automatically apply to its standardized counterpart. This is because standardized estimates have their own standard errors, and their critical ratios may not correspond to the same probabilities as the critical ratios for the corresponding unstandardized results. This explains why you should (1) always report the unstandardized solution including the standard errors and (2) not associate  $p$  values for unstandardized estimates with the corresponding standardized estimates. An example follows.

Suppose that the values of an unstandardized estimate, its standard error, and the standardized estimate are, respectively, 4.20, 2.00, and .60. In a large sample, the unstandardized estimate would be significant at the .05 level because  $z = 4.20/2.00$ , or 2.10, which exceeds the critical value (1.96) at  $p < .05$ . Whether the standardized estimate of .60 is also significant at the same level is unknown because it has no standard error. Consequently, it would be inappropriate to report the standardized coefficient as

\* .60\*

where the asterisk designates  $p < .05$ . It is better to report both the unstandardized and standardized estimates and also the standard error of the former, like this:

✓ 4.20\* (2.10) .60

where the asterisk is associated with the unstandardized estimate (4.20), not the standardized one (.60).

## POWER AND TYPES OF NULL HYPOTHESES

The failure to reject a null hypothesis, such as  $p \geq .05$  when testing at the .05 level, is meaningful only if (1) the power of the test is adequate and (2) the null hypothesis is at least plausible to some degree. **Power** is the probability of getting statistically significant results over random samples when the null hypothesis is false. Power is also the

complement of probability of a Type II error (failing to reject  $H_0$  when it is false), often designated as  $\beta$ , so  $1 - \beta = \text{power}$ . Whatever increases power decreases  $\beta$ , and vice versa. Power varies directly with the magnitude of the population effect size and your sample size. Other factors that affect power include:

1. The level of statistical significance  $\alpha$  (e.g., .05 vs. .01) and the directionality of  $H_1$  (i.e., one- or two-tailed tests).
2. Whether samples are independent or dependent (i.e., between-subjects or within-subjects design).
3. The particular test statistic used.
4. The reliability of the scores.

The following combination generally leads to the greatest power: a large sample, specification of  $\alpha = .05$ , a one-tailed (directional)  $H_1$ , a within-subjects design, a parametric test statistic (e.g.,  $t$ ) rather than a nonparametric statistic (e.g., Mann–Whitney  $U$ ), and scores that are very reliable.

Power should be estimated when the study is planned but before the data are collected. Some granting agencies require such a priori estimates of power in applications for funds. If power is low, there is little point in carrying out the study, if outcomes of significance testing are important. For example, if power is only about .50, then the likelihood over random samples of rejecting a false null hypothesis is no greater than guessing the outcome of a coin toss. In this case, tossing a coin instead of conducting the study would be just as likely to give the correct decision in the long run and would save time and money, too.

Unfortunately, only about 10% of researchers report the a priori power of their analyses (Ellis, 2010). This is a problem because without knowing power estimates, one is unable to correctly interpret results that are not statistically significant. That is, do such results indicate lack of support for the researcher's hypothesis or just the expected consequence of inadequate power? There is free software for power analysis, so the widespread failure to estimate and report power is bewildering.<sup>2</sup> How to estimate power in SEM is described in Chapter 12, but power for certain kinds of significance tests in SEM is often quite low even in large samples.

The type of null hypothesis tested most often is a **nil hypothesis**, which says that the value of a parameter or the difference between two or more parameters is zero. A nil hypothesis for the  $t$  test of a mean contrast is

$$H_0: \mu_1 - \mu_2 = 0$$

which predicts that two population means are exactly equal. The problem with nil hypotheses is that it is unlikely that the value of any parameter (or difference between

---

<sup>2</sup>[www.gpower.hhu.de/en.html](http://www.gpower.hhu.de/en.html)

two parameters) is exactly zero, especially if zero means the total absence of an effect or association. It is possible for the  $t$  test to specify a **non-nil hypothesis**, such as

$$H_0: \mu_1 - \mu_2 = 5.0$$

but doing so is rare in practice. As the name suggests, a non-nil hypothesis predicts that a population effect or association is not zero.

It is more difficult to specify and test non-nil hypotheses for other test statistics, such as  $F$  when comparing three or more means. This is because computer programs almost always assume a nil hypothesis. Nil hypotheses may be appropriate in new research areas where it is unknown whether effects exist at all, but such hypotheses are less suitable in more established areas where it is already known that certain effects are not zero. If so, then (1) an implausible nil hypothesis is an uninteresting “straw man” argument (a fallacy) that is easily rejected, and (2)  $p$  values in significance testing are too low. This happens because the data seem more exceptional than they really are compared with evaluating the same data under a more realistic non-nil hypothesis.

## SIGNIFICANCE TESTING CONTROVERSY

Until recently, significance testing was both routine and expected (i.e., everybody did it). But significance testing has been increasingly criticized as unscientific and unempirical (Kline, 2013a; Lambdin, 2012). Some authors in **statistics reform** suggest that overreliance on significance testing can lead to **trained incapacity**, or the inability of researchers to understand their own results due to inherent limitations of significance tests and myriad associated cognitive distortions (Ziliak & McCloskey, 2008). Essential criticisms of significance testing are listed next:

1. Outcomes of significance tests— $p$  values—are wrong in most studies.
2. Researchers do not understand  $p$  values.
3. Most applications of significance testing are incorrect.
4. Significance tests do not tell researchers what they want to know.

The fact that  $p$  values are calculated under implausible assumptions (e.g., random sampling, normality, no measurement error) was mentioned earlier in the section on standard errors. Distributional assumptions are rarely verified because researchers mistakenly believe that significance tests are robust even in small, unrepresentative samples (Hoekstra, Kiers, & Johnson, 2013). If assumptions are checked, the wrong methods are used, including significance tests that supposedly verify distributional assumptions of other significance tests, such as Levene’s test for homoscedasticity. The problem with such tests is that their results are often wrong due in part to their own unrealistic assumptions (Erceg-Hurn & Mirosevich, 2008).

Most researchers misinterpret statistical significance. For example, about 80–90% of psychology professors endorse false beliefs about statistical significance, no better than psychology undergraduate students in introductory statistics courses (Haller & Krauss, 2002). These comparably high rates of misinterpretation suggest an ongoing cycle of misinformation, where instructors transmit false information to students, who then perpetuate the myths to the next generation. Most false beliefs about  $p$  values involve overinterpretation that favor the researcher's hypotheses, which is a form of confirmation bias—see Topic Box 3.1 for a review of the “Big Five” misinterpretations of statistical significance. Exercises 1–3 ask you to comment on examples of incorrect definitions of  $p$  values.

### TOPIC BOX 3.1

#### Cognitive Errors in Significance Testing

First, we consider the correct interpretation of  $p$  values, which is actually quite narrow in scope. They represent the conditional probability:

$$p \left( \begin{array}{l} \text{Result or} \\ \text{more extreme} \end{array} \middle| \begin{array}{l} H_0 \text{ true, random sampling,} \\ \text{all other assumptions} \end{array} \right)$$

which is the likelihood of a sample result or one even more extreme assuming random sampling under a true null hypothesis and where all other assumptions are met (distributional requirements, no error other than sampling error, independent scores, etc.). Most of what contributes to a  $p$  value are those *even more extreme* results that were not actually observed; that is,  $p$  values are only partially empirical. Two correct interpretations for the case  $p < .05$  are given next. Other correct definitions are probably just variations of the ones that follow:

1. Suppose the study were repeated by drawing many random samples from the same population(s) where the null hypothesis is true (i.e., every result happens by chance). Less than 5% of these hypothetical results would be even more inconsistent with  $H_0$  than the actual result.
2. Less than 5% of test statistics from many random samples are further away from the mean of the sampling distribution under  $H_0$  than the one for the observed result. In other words, the odds are less than 1 to 19 of getting a result from a random sample even more extreme than the observed one.

Described next are what I call the “Big Five” misinterpretations of  $p$  values. The **odds against chance fallacy** is the false belief that  $p$  indicates the probability that a particular result happened by chance (i.e., due to sampling error). Remember that  $p$  is calculated for a range of results, most unobserved, and not for any single

result. Also,  $p$  is calculated assuming that  $H_0$  is already true, so the probability that sampling error is the only explanation is already taken to be 1.0. Thus, it is illogical to view  $p$  as measuring the likelihood of sampling error. Besides, the probability that sample results are affected by error of some kind—sampling, measurement, or specification error, among others—is virtually 1.0. From this perspective, virtually all sample results are wrong (Ioannidis, 2005). That is, our data routinely lie, they lie through multiple types of error, and it is only when results are averaged over studies, such as in the technique of meta-analysis, that some of these errors begin to cancel out. Significance testing in individual studies in no way helps in this process.

The **local Type I error fallacy** for the case where  $p < .05$  and  $\alpha = .05$  (i.e.,  $H_0$  is rejected) says that the likelihood that the decision just taken to reject the null hypothesis is a Type I error is less than 5%. This belief is false because any particular decision to reject  $H_0$  is either correct or incorrect, so no probability (error other than 0 or 1.0) is associated with it. Only with sufficient replication could we determine whether or not the decision to reject  $H_0$  in a particular study was correct. The **inverse probability fallacy** is the false belief that  $p$  is the probability that the null hypothesis is true. This error stems from forgetting that  $p$  values are probabilities of data under the null hypothesis, not the other way around.

Two other fallacies concern the complements of  $p$  values, or  $1 - p$ . The **valid research hypothesis fallacy** is the false belief that  $1 - p$  is the probability that the alternative hypothesis is true. The quantity  $1 - p$  is a probability, but it is just the likelihood of getting a result even *less* extreme under  $H_0$  than the one actually found. The **replicability fallacy** is that  $1 - p$  is the likelihood of finding the same result in another random sample. If this fallacy were true, knowing the likelihood of replication would be very useful. Unfortunately,  $p$  is just the probability of data in a particular sample under a specific null hypothesis. In general, replication is a matter of experimental design and whether some effect actually exists in the population (i.e., it is an empirical question). Kline (2013a, chap. 4) describes additional false beliefs about  $p$  values.

Most researchers fail to report the power of their significance tests. Another misuse comes from treating the conventional levels of statistical significance, .05 or .01, as golden rules that apply to all studies and disciplines. The value of  $\alpha$  sets the risk of Type I error, or the probability over random samples that a true null hypothesis will be rejected, but Type II error is often more serious. An example is when a nil hypothesis is known to be false before even collecting the data. In this case, the effective level of  $\alpha$  is zero, and Type II error is the only possible kind of error. Another example is when a treatment for an illness is beneficial, but the results are not significant at  $p < .05$ , the highest conventional level of  $\alpha$ . Type II error in this context means that a beneficial treatment is not detected. There is actually no requirement to specify an arbitrary level

of  $\alpha$  (i.e., .05 or .01) that does not properly balance the risk of Type I error against that of Type II error (Hurlbert & Lombardi, 2009).

Armstrong (2007) argued that significance testing does not foster progress in science even if such tests are properly conducted. This is because their results do not tell researchers what they wish to know, including the likelihood that some hypothesis is true, given the data; the probability that a Type I error has occurred, given that the null hypothesis was just rejected; the prospects for replication; and whether the findings are actually important. An alternative is to describe replicated results in terms of their effect sizes and precisions (confidence intervals) and interpret their substantive significance using language relevant to stakeholders in a particular research context (Aguinis et al., 2010). Given all the problems just considered, significance testing is actually banned in some research journals such as *Basic and Applied Social Psychology* (Trafimow & Marks, 2015).

## **CONFIDENCE INTERVALS AND NONCENTRAL TEST DISTRIBUTIONS**

Interval estimation is an alternative to significance testing. It involves reporting effect sizes with confidence intervals (error bars, margins of error) that indicate a range of results considered equivalent within the limits of sampling error to the specific result found (i.e., the point estimate). For statistics with simple distributions, the width of either side of a

$$100 \times (1 - \alpha) \%$$

confidence interval is determined by the product of the standard error and the critical value of a central test statistic at the  $\alpha$  level of statistical significance for a two-tailed alternative hypothesis. For example, given

$$M = 100.00, SD = 9.00, N = 25, \text{ and } SE_M = 1.80$$

the 95% confidence interval is

$$100.00 \pm (1.80) t_{2\text{-tail}, \alpha = .05} \quad (24)$$

where  $t_{2\text{-tail}, \alpha = .05}$  (24) is the positive two-tailed critical value in a central  $t$  distribution at the .05 level of statistical significance, which for  $df = 24$  is 2.064.<sup>3</sup> The 95% confidence interval is thus

$$100.00 \pm 1.80 (2.064), \text{ or } 100.00 \pm 3.72$$

---

<sup>3</sup>See the calculating webpage at [www.usablestats.com/calcs/tinv](http://www.usablestats.com/calcs/tinv)

which defines the interval [96.28, 103.72]. This interval specifies a range of values considered equivalent to the observed mean within the limits of sampling error at the 95% confidence level. The point estimate of 100.00 falls at the exact center of the interval, and the whole interval explicitly conveys the idea that a margin of error is associated with the corresponding statistic (100.00). Note that the interval [96.28, 103.72] is based on a single estimate of  $\sigma_M$ , or  $SE_M = 1.80$ . But this quantity (1.80) is itself just a point estimate, and the value of  $SE_M$  in a different sample will almost certainly not be 1.80. This means that the interval [96.28, 103.72] is actually too narrow (i.e., more precise than it seems), if we also consider sampling error in  $SE_M$ .

Because confidence intervals are based on the same standard errors as significance tests—and rely on the same unrealistic assumptions—researchers should not overinterpret their lower or upper bounds. Suppose a 95% confidence interval based on  $M = 2.50$  is [0, 5.00], which includes zero. This fact can be misinterpreted, such as wrongly concluding that  $\mu = 0$ . But zero is only one value within a range of estimates, so it has no special status. This means that the hypothesis that  $\mu = 0$  is not favored any more than the hypothesis that  $\mu = 5.00$  (or that  $\mu$  equals any other value in the range 0–5.0). Confidence intervals are subject to sampling error, too, so zero may not fall within the 95% confidence interval in a replication sample. Do not believe that confidence intervals are just significance tests in disguise (Thompson, 2006). This is because null hypotheses are required for significance tests, but not for confidence intervals, and many null hypotheses have little scientific value.

Statistics with complex distributions may not follow central distributions. For example, if  $\rho^2 = 0$ , then distributions of  $R^2$  follow **central F distributions** with  $k$  and  $N - k - 1$  degrees of freedom, where  $k$  is the number of predictors. Central F distributions assume  $\rho^2 = 0$  and provide the critical values for the familiar  $F$  test in multiple regression or ANOVA. But if  $\rho^2 > 0$ , the sampling distribution for  $R^2$  is defined by **non-central F distributions**, which have an additional parameter, called the **noncentrality parameter**. This parameter indicates the degree to which the null hypothesis that  $\rho^2 = 0$  is false. Noncentral F distributions take the form

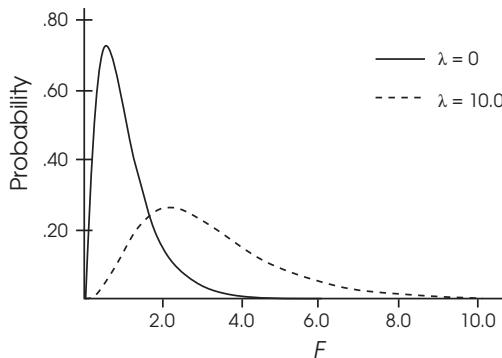
$$F(k, N - k - 1, \lambda) \quad (3.3)$$

where  $\lambda$  is the noncentrality parameter. The latter is related to  $\rho^2$  and the sample size, or

$$\lambda = N \left( \frac{\rho^2}{1 - \rho^2} \right) \quad (3.4)$$

If  $\rho^2 = 0$ , then  $\lambda = 0$ , which indicates no departure from the nil hypothesis. Equation 3.4 can be rearranged to express  $\rho^2$  as a function of  $\lambda$  and the sample size:

$$\rho^2 = \frac{\lambda}{N + \lambda} \quad (3.5)$$



**FIGURE 3.2.** Distributions of central  $F$  and noncentral  $F$  for 5 and 20 degrees and where the noncentrality parameter ( $\lambda$ ) equals 0 for central  $F$  and  $\lambda = 10.0$  for noncentral  $F$ .

Presented in Figure 3.2 are two  $F$  distributions where the degrees of freedom are 5 and 20. For the central  $F$  distribution in the left part of the figure,  $\lambda = 0$ . But  $\lambda = 10.0$  for the noncentral  $F$  distribution in the right side of the figure. Note in the figure that (1) both distributions are positively skewed, but the central  $F$  distribution has greater skew than the noncentral  $F$  distribution. Also, (2) the noncentral  $F$  distribution has a greater expected value—the weighted average of all possible values—than the central  $F$  distribution. This is because the noncentral  $F$  distribution in the figure assumes that  $\rho^2 > 0$ , but the central  $F$  distribution is for  $\rho^2 = 0$ .

Steiger and Fouladi (1997) showed that if we can obtain a confidence interval for  $\lambda$ , we can also obtain a confidence based on  $R^2$  (which estimates  $\rho^2$ ) using Equation 3.5. To do so, we use a computer tool that finds  $\lambda_L$ , the lower bound of the confidence interval for  $\lambda$ . For the 95% level, the lower bound  $\lambda_L$  equals the value of  $\lambda$  for the noncentral  $F$  distribution in which the observed  $F$  falls at the 97.5th percentile. The upper bound  $\lambda_U$  equals the value of  $\lambda$  for the noncentral  $F$  distribution in which the observed  $F$  falls at the 2.5th percentile. But we need to find which particular noncentral  $F$  distributions are the most consistent with the data, and this is the problem solved with the right computer tool. An example follows.

I used J. Steiger's Noncentral Distributional Calculator (NDC), a freely available Windows application for noncentrality interval estimation.<sup>4</sup> For the data in Table 2.1,

$$R_{Y,X,W}^2 = .576, N = 20, \text{ and } F(2, 17) = 11.536$$

We can say the observed  $F$  of 11.536 falls at the

1. 97.5th percentile in the noncentral  $F(2, 17, 4.190)$  distribution; and the same observed  $F$  falls at the
2. 2.5th percentile in the noncentral  $F(2, 17, 52.047)$  distribution.

<sup>4</sup>[www.statpower.net/Software.html](http://www.statpower.net/Software.html)

For these data, the 95% confidence interval for  $\lambda$  is [4.190, 52.047]. Using Equation 3.5 to convert the lower and upper bounds of this interval to  $\rho^2$  units for  $N = 20$  gives us the noncentral 95% confidence interval based on  $R^2 = .576$ , which is [.173, .722]. (You should verify these results.) The interval just reported is not symmetrical about  $R^2 = .576$ , but this is expected in noncentrality interval estimation. Exercise 4 asks you to calculate the 95% noncentral confidence interval based on the same value of  $R^2$  but in a larger sample.

There are noncentral distributions for other test statistics, such as  $t$  and  $\chi^2$ , and they all assume that the null hypothesis is false by the degree indicated by the value of the noncentrality parameter. The latter equals zero in central test distributions, so central test distributions are just special cases of noncentral test distributions (i.e., they belong to the same distribution family). Noncentral test distributions play an important role in certain types of statistical analyses. Computer programs that estimate the power of significance tests as a function of study characteristics and the predicted effect size analyze noncentral distributions. This is because the concept of power assumes that the null hypothesis is false, and it is false by the degree indicated by a nonzero effect size. The latter generally corresponds to a value of the noncentrality parameter that is also not zero.

Another application is the estimation of confidence intervals based on sample statistics that measure effect size besides  $R^2$ . For example, distributions of standardized mean differences ( $d$ ), or the ratio of a mean contrast over the standard deviation, generally follow central  $t$  distributions when the corresponding parameter is zero; otherwise,  $d$  statistics are distributed as noncentral  $t$  distributions. There are special computer programs for noncentrality interval estimation based on  $d$  statistics (Cumming, 2012). Effect size estimation also assumes that the null hypothesis—especially when it is a nil hypothesis—is false.

Some measures of model fit in SEM are based on noncentral  $\chi^2$  distributions. These statistics measure the degree of **approximate (close) fit**, which allow for an “acceptable” amount of departure from **exact (perfect) fit**. What is considered “acceptable” departure from perfection is related to the value of the noncentrality parameter for the  $\chi^2$  that the computer calculates for the model and data. Other fit statistics in SEM measure the departure from exact fit, and these statistics are generally described by central  $\chi^2$  distributions, where the null hypothesis that the model has perfect fit in the population is assumed to be true. But the null hypothesis just stated is assumed to be false by statistics that measure approximate fit. Assessment of model fit against these two standards, approximate versus exact, is covered later in Chapter 12.

## BOOTSTRAPPING

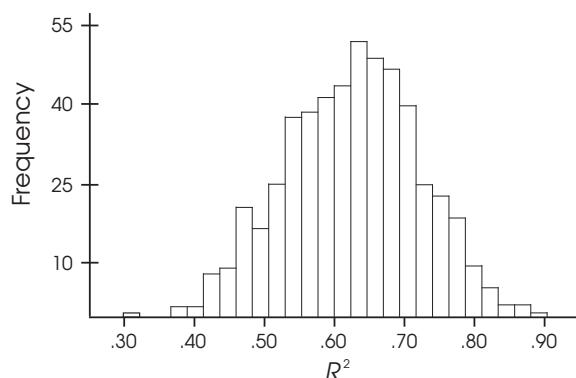
The technique of bootstrapping was developed by the statistician B. Efron in the 1970s (e.g., 1979). It is a computer-based method of **resampling** that combines the cases in a data set in different ways to estimate statistical precision. Perhaps the best known form is **nonparametric bootstrapping**, which generally makes no assumptions other than that the distribution in the sample reflects the basic shape of that in the population.

This method treats your sample (i.e., data file) as a pseudo-population in that cases are randomly selected *with replacement* to generate other data sets, usually of the same size as the original. Because of sampling with replacement, (1) the same case can be selected in more than one generated data set or at least twice in the same generated sample, and (2) the composition of cases will vary slightly across the generated samples.

When repeated many times (e.g., 500) by the computer, bootstrapping simulates random sampling with replacement. It also constructs an **empirical sampling distribution**, the frequency distribution of the values of a statistic across generated samples. **Nonparametric bootstrapped confidence intervals** are calculated in the empirical distribution. For example, the lower and upper bounds of a 95% bootstrapped confidence interval correspond to, respectively, the 2.5th and 97.5th percentiles in the empirical sampling distribution. These limits contain 95% of the bootstrapped values of the statistic. This method is potentially useful for statistics with complex distributions. An example follows.

I used the nonparametric Bootstrap procedure of SimStat for Windows (Version 2.6.1) (Provalis Research, 1995–2011) to resample from the data in Table 2.1 in order to generate a total of 500 bootstrapped samples each with 20 cases.<sup>5</sup> Presented in Figure 3.3 is the empirical sampling distribution of  $R^2$  across all generated samples. SimStat reported that the mean of this distribution is .626, the median is .630, and the standard deviation is .102. The first result (.626) is close to the observed value of  $R^2 = .576$  for these data, which is expected.

The nonparametric bootstrapped 95% confidence interval in the empirical sampling distribution for this example is [.425, .813]. This result from bootstrapping is quite different from the noncentral 95% confidence interval we calculated earlier for the same data, or [.173, .722], but bootstrapped results in small samples can be very inaccurate. This is because bootstrapping can magnify the effects of unusual features in a small data



**FIGURE 3.3.** Empirical sampling distribution for  $R^2_{Y,X,W}$  in 500 bootstrapped samples for the data in Table 2.1.

<sup>5</sup><http://provalisresearch.com>

set. Note that the computer will generate a different empirical sampling distribution for the same data, if each time it is given a different **seed**, or a long number (vector) used to initiate simulated random sampling. Consequently, any result in a single application of nonparametric bootstrapping is not generally unique.

A raw data file is needed for nonparametric bootstrapping. This is not true in **parametric bootstrapping**, where the computer randomly samples from a theoretical probability density function specified by the researcher. When repeated many times by the computer, values of statistics in synthesized samples vary randomly about the specified parameters, which simulates sampling error. Parametric bootstrapping is a kind of Monte Carlo method that is used in computer simulation studies of the properties of estimators. Distributional assumptions can be added incrementally in parametric bootstrapping or successively relaxed over the generation of synthetic data sets.

Several SEM computer tools, including Amos, EQS, LISREL, Mplus, Stata, and lavaan for R, feature bootstrap methods. Some of these methods can estimate standard errors or generate confidence intervals based on certain estimators, such as statistics that measure model–data correspondence or indirect causal effects (Hancock & Liu, 2012). Parametric bootstrapping methods are used in SEM to conduct simulation studies, such as for power analysis, sample size determination, and hypothesis testing (Bandalos & Gagné, 2012).

## SUMMARY

Statistical significance is not a gold standard, and thinking about data analysis as a search for whether results are “significant” or “not significant” may be fruitless. This is because the presence of statistical significance does not reliably signal that results are noteworthy or even of mild interest, just as the failure to find statistical significance does not indicate that nothing of interest was found. It does not help that many, and perhaps most, researchers do not understand what statistical significance really means. Researchers should instead think more about whether observed effect sizes are precise and large enough to be of substantive interest. Keeping a skeptical view of significance testing will help you in SEM—and in other kinds of complex multivariate analyses, too—to avoid getting lost in a blizzard of asterisks. Also reviewed in this chapter was the logic of noncentrality interval estimation and bootstrapping, both of which can be used to calculate confidence intervals for statistics with complex distributions, including some that are used in SEM. Preparation of your data for analysis in SEM is considered in the next chapter.

## LEARN MORE

Kline (2013a, chap. 4) describes additional cognitive errors about statistical significance, and Lambdin (2012) and Ziliak and McCloskey (2008) offer strong critiques of significance testing.

- Kline, R. B. (2013a). *Beyond significance testing: Statistics reform in the behavioral sciences*. Washington, DC: American Psychological Association.
- Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory and Psychology*, 22, 67–90.
- Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

## EXERCISES

Explain what is wrong versus right in each definition of statistical significance listed next. These quotes are not from academic sources, but you can easily find similar kinds of errors in academic works, too.

1. The statistical significance of a result is an estimated measure of the degree to which it is “true” (in the sense of “representative of the population”). More technically, the value of the  $p$  level represents a decreasing index of the reliability of a result. The higher the  $p$  level, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the  $p$ -level represents the probability of error that is involved in accepting our observed result as valid, that is, as “representative of the population.”<sup>6</sup>
2. This is a very important and common term in psychology, but one that many people have problems with. Technically, statistical significance is the probability of some result from a statistical test occurring by chance. . . . Most often, psychologists look for a probability of 5% or less that the results are due to chance, which means a 95% chance the results are “not” due to chance.<sup>7</sup>
3. The calculation of statistical significance . . . is subject to a certain degree of error. The researcher must define in advance the probability of a sampling error. . . . Sample size is an important component of statistical significance in that larger samples are less prone to flukes. Only random, representative samples should be used in significance testing.<sup>8</sup>
4. Calculate the 95% noncentral confidence interval for  $R^2_{Y.X,W} = .576$ ,  $F(2, 47) = 31.925$ , and  $N = 50$  using a computer tool for noncentrality interval estimation.

<sup>6</sup><http://dictionary.babylon.com/statistical%20significance>

<sup>7</sup>[www.alleydog.com/glossary/definition.php?term=Statistical%20Significance](http://www.alleydog.com/glossary/definition.php?term=Statistical%20Significance)

<sup>8</sup>[www.investopedia.com/terms/s/statistical-significance.asp](http://www.investopedia.com/terms/s/statistical-significance.asp)

# 4

## Data Preparation and Psychometrics Review

---

Just as in other types of statistical analyses, data preparation is critical in SEM for three reasons. First, it is easy to make a mistake entering data into computer files. Second, the most widely used estimation methods in SEM make specific distributional assumptions about the data. These assumptions must be taken seriously because violation of them could result in bias. Third, data-related problems can make SEM computer tools fail to yield a logical solution. A researcher who has not carefully prepared and screened the data could mistakenly believe that the model is at fault, and confusion ensues. How to select good measures is also considered along with review of basic psychometric issues, including the evaluation of score reliability and validity. It is not possible to cover all aspects of data screening or psychometrics in a single chapter, but more advanced works are cited throughout, and they should be consulted for more information. This adage attributed to Abraham Lincoln sets the tone for this chapter: If I had eight hours to chop a tree, I'd spend six sharpening my axe.

---

### **FORMS OF INPUT DATA**

Most primary researchers—those who conduct original studies—input raw data files for analysis with SEM computer programs. Just as in multiple regression, raw data themselves are not necessary for many—and perhaps most—types of SEM. For example, when analyzing continuous outcomes with default maximum likelihood estimation, a matrix of summary statistics instead can be the input to an SEM computer tool instead of a raw data file. In fact, you can replicate most of the analyses described in this book using the data matrix summaries that accompany them—see the website for this book. This is a great way to learn because you can make mistakes using someone else's data before analyzing your own. Many journal articles about the results of SEM contain

enough information, such as correlations and standard deviations, to create a matrix summary of the data, which can then be submitted to a computer program for analysis. Thus, readers of these works can, with no access to the raw data, replicate the original analyses or estimate alternative models not considered in the original work.

Basically, all SEM computer tools accept either a raw data file or a matrix summary of the data. If a raw data file is submitted, the program will create its own matrix, which is then analyzed. You should consider the following issues when choosing between a raw data file and a matrix summary as program input:

**1. Some special types of analyses require raw data files.** There are three basic kinds. One is when continuous outcomes have severely non-normal distributions and the data are analyzed with a method that assumes normality, but test statistics and standard errors are calculated that adjust for non-normality. A second situation concerns missing data. You should know that default maximum likelihood estimation does not handle incomplete raw data files. But special versions of the maximum likelihood method are available in many SEM computer tools that analyze incomplete data sets. The third case is when outcome variables are not continuous, that is, they are ordinal or nominal variables. Such outcomes can be analyzed in SEM, but raw data files are needed. For analyses that do not involve any of these applications, either the raw data or a matrix summary of them can be analyzed.

**2. Matrix input offers a potential economy over raw data files.** Suppose that 1,000 cases are measured on 10 continuous variables. The data file may be 1,000 lines (or more) in length, but a matrix summary for the same data might be only 10 lines long.

**3. Sometimes a researcher might “make up” a data matrix using theory or results from a meta-analysis, so there are no raw data, only a matrix summary.** A made-up data matrix can be submitted to an SEM computer tool for analysis. This is also a way to diagnose certain kinds of technical problems that can crop up in SEM. This point is elaborated in later chapters.

If means are not analyzed, there are two basic summaries of raw data—correlation matrices with standard deviations and covariance matrices. For example, presented in the top part of Table 4.1 are the correlation matrix with standard deviations (left) and the covariance matrix (right) for the raw data in Table 2.1 on three continuous variables. Whenever possible, at least four-decimal accuracy is recommended for matrix input. Precision at this level helps to minimize rounding error in the analysis. All summary matrices in Table 4.1 are in **lower diagonal form** where only the unique values of correlations or covariances are reported in the lower-left-hand side of the matrix. Most SEM computer tools accept lower diagonal matrices as alternatives to full ones, with redundant entries above and below the diagonal, and can “assemble” a covariance matrix given the correlations and standard deviations. Exercise 1 asks you to reproduce the covariance matrix in the upper right part of Table 4.1 from the correlations and standard deviations in the upper left part of the table.

**TABLE 4.1. Matrix Summaries of the Data in Table 2.1**

X	W	Y	X	W	Y
Summaries without means					
r, SD			cov		
1.0000			9.0421		
.2721	1.0000		2.3053	7.9368	
.6858	.4991	1.0000	22.4158	15.2842	118.1553
3.0070	2.8172	10.8699			
Summaries with means					
r, SD, M			cov, M		
1.0000			9.0421		
.2721	1.0000		2.3053	7.9368	
.6858	.4991	1.0000	22.4158	15.2842	118.1553
3.0070	2.8172	10.8699	16.9000	49.4000	102.9500
16.9000	49.4000	102.9500			

It may be problematic to submit for analysis just a correlation matrix without standard deviations, specify that all standard deviations equal 1.0 (which standardizes everything), or convert the raw scores to normal deviates (*z* scores) and then submit for analysis the data file of standardized scores. This is because most estimation methods in SEM, including default maximum likelihood estimation, assume that the variables are unstandardized. This implies that if a correlation matrix without the original standard deviations is analyzed, the results may not be correct. Potential problems include the derivation of incorrect standard errors for standardized estimates if special methods for standardized variables are not used. Some SEM computer programs give warning messages or terminate the run if the researcher requests the analysis of a correlation matrix only with default maximum likelihood estimation. Thus, it is generally safer to analyze a covariance matrix (or a correlation matrix with standard deviations). Accordingly, covariances are analyzed for most examples in this book. When a correlation matrix is analyzed, I use a special estimation method for standardized variables described in Chapter 11. The issues just discussed about the pitfalls of analyzing correlation matrices without standard deviations explain why you must clearly state in written reports the specific kind of data matrix analyzed and the estimation method used.

Matrix summaries of raw data must consist of the covariances and means whenever means are analyzed in SEM. Presented in the lower part of Table 4.1 are matrix summaries of the data in Table 2.1 that include the correlations, standard deviations, and means (left) and the covariances and means (right). Both matrices convey the same information. Even if your analyses do not concern means, you should nevertheless report the means of all variables. You may not be interested in analyzing means, but someone else

may be. Always report sufficient descriptive statistics (including the means) so that others can reproduce your results.

## POSITIVE DEFINITENESS

The data matrix that you submit—or the one calculated by the computer from your raw data—to an SEM computer program should be **positive definite**, which is required for most estimation methods. A matrix that lacks this characteristic is **nonpositive definite**; therefore, attempts to analyze such a data matrix would probably fail. A positive definite data matrix has the properties summarized next and then discussed:

1. The matrix is **nonsingular** or has an inverse. A matrix with no inverse is **singular**.
2. All eigenvalues of the matrix are positive ( $> 0$ ), which also says that the matrix determinant is positive.
3. There are no out-of-bounds correlations or covariances.

In most kinds of multivariate analyses (SEM included), the computer needs to derive the inverse of the data matrix as part of its linear algebra operations. If the matrix has no inverse, these operations fail. An **eigenvalue** is the variance of an **eigenvector**, and both are from a principal components analysis of the data matrix, or **eigendecomposition**, that creates a total of  $v$  orthogonal linear combinations, or eigenvectors, of the observed variables, where  $v$  is the total number of those variables. The maximum number of eigenvectors for a data matrix equals  $v$ , and the set of all possible eigenvectors explains all the variance of the original variables.

If any eigenvalue equals zero, then (1) the matrix is singular, and (2) there is some pattern of perfect collinearity that involves at least two variables (e.g.,  $r_{XY} = 1.0$ ) or three or more variables in a more complex configuration (e.g.,  $R_{YX,W} = 1.0$ ). Perfect collinearity means that some denominators in matrix calculations will be zero, which results in illegal (undefined) fractions (estimation fails). Near-perfect collinearity, such as  $r_{XY} = .95$ , manifested as near-zero eigenvalues, can also cause this problem.

Negative eigenvalues ( $< 0$ ) may indicate a data matrix element—a correlation or covariance—that is **out of bounds**. Such an element would be mathematically impossible to derive if all elements were calculated from the same cases with no missing data. For example, the value of the Pearson correlation between two variables  $X$  and  $Y$  is limited by the correlations between these variables and a third variable  $W$ . Specifically, the value of  $r_{XY}$  must fall within the range defined next:

$$(r_{XW} \times r_{WY}) \pm \sqrt{(1 - r_{XW}^2)(1 - r_{WY}^2)} \quad (4.1)$$

Given  $r_{XW} = .60$  and  $r_{WY} = .40$ , for example, the value of  $r_{XY}$  must fall within the range

.24 ± .73, or from -.49 to .97

Any other value of  $r_{XY}$  would be out of bounds. (You should verify this result using Equation 4.1.) Another way to view Equation 4.1 is that it specifies a **triangle inequality** for values of correlations among three variables measured in the same sample.<sup>1</sup>

In a positive definite data matrix, the maximum absolute value of  $\text{cov}_{XY}$ , the covariance between  $X$  and  $Y$ , must respect the limit defined next:

$$\max |\text{cov}_{XY}| \leq \sqrt{s_X^2 \times s_Y^2} \quad (4.2)$$

where  $s_X^2$  and  $s_Y^2$  are, respectively, the sample variances of  $X$  and  $Y$ . In words, the maximum absolute value for the covariance between two variables is less than or equal to the square root of the product of their variances; otherwise, the value of  $\text{cov}_{XY}$  is out of bounds. For example, given

$$\text{cov}_{XY} = 13.00, s_X^2 = 12.00, \text{ and } s_Y^2 = 10.00$$

the covariance between  $X$  and  $Y$  is out of bounds because

$$13.00 > \sqrt{12.00 \times 10.00} = 10.95$$

which violates Equation 4.2. The value of  $r_{XY}$  for this example is also out of bounds because it equals 1.19 (an impossible result), given these variances and covariance. Exercise 2 asks you to verify this fact.

The **determinant** of the data matrix is the serial product (the first times the second times the third, and so on) of the eigenvalues. Assuming that all eigenvalues are positive, the determinant is a kind of matrix variance. Specifically, it is the volume of the multivariate space “mapped” by the set of observed variables.<sup>2</sup> If any eigenvalue equals zero, then the determinant is zero; in this case, the matrix has no inverse (it is singular). A close-to-zero eigenvalue will probably make the determinant be close to zero, which signals the potential inability of the computer to derive the inverse. If some odd number of the eigenvalues (1 or 3 or 5, etc.) is negative, then the determinant will be negative, too. A data matrix with a negative determinant may have an inverse, but the whole matrix is still nonpositive definite, perhaps due to out-of-bounds correlations or covariances. See Topic Box 4.1 for more information about causes of nonpositive definiteness in the data matrix and possible solutions.

Before analyzing in SEM either a raw data file or a matrix summary, the original data file should be screened for the problems considered next. Some of these potential

---

<sup>1</sup>In a geometric triangle, the length of a given side must be less than the sum of the lengths of the other two sides but greater than the difference between the lengths of the two sides.

<sup>2</sup>For diagrams, see <http://en.wikipedia.org/wiki/Determinant>

**TOPIC BOX 4.1****Causes of Nonpositive Definiteness and Solutions**

Many points summarized here are from Wotheke (1993) and Rigdon (1997). Some causes of nonpositive definite data matrices are listed next. Most can be detected through data screening.

1. Extreme bivariate or multivariate collinearity among the observed variables.
2. The presence of outliers that force the values of correlations to be extremely high.
3. Pairwise deletion of cases with missing data.
4. Making a typing mistake when transcribing a data matrix from one source, such as a table in a journal article, to another, such as a command file for computer analysis, can result in a nonpositive definite data matrix. For example, if the value of a covariance in the original matrix is 15.00, then mistakenly typing 150.00 in the transcribed matrix could generate a non-positive definite matrix.
5. Plain old sampling error can generate nonpositive definite data matrices, especially in small or unrepresentative samples.
6. Sometimes matrices of estimated Pearson correlations, such as polyserial or polychoric correlations derived for noncontinuous observed variables, can be nonpositive definite.

Here are some tips about diagnosing whether a data matrix is positive definite before submitting it for analysis to an SEM computer tool: Copy the full matrix (with redundant entries above and below the diagonal) into a text (ASCII) file, such as Microsoft Windows Notepad. Next, point your Internet browser to a free, online matrix calculator and then copy the data matrix into the proper window on the calculating webpage. Finally, select options on the calculating webpage to derive the determinant and eigenvalues with the corresponding eigenvectors. Look for outcomes that indicate nonpositive definiteness, such as near-zero, zero, or negative eigenvalues. A handy matrix calculator is available at [www.bluebit.gr/matrix-calculator](http://www.bluebit.gr/matrix-calculator).

Suppose that the covariances among continuous variables  $X$ ,  $W$ , and  $Y$ , respectively, are

$$\begin{bmatrix} 1.00 & & \\ .30 & 2.00 & \\ .65 & 1.15 & .90 \end{bmatrix} \quad (\text{I})$$

The eigenvalues for this matrix (I) derived using the online matrix calculator just mentioned are

$$(2.918, .982, 0)$$

The third eigenvalue is zero, so let us inspect the weights for the third eigenvector, which for  $X$ ,  $W$ , and  $Y$ , respectively, are

$$(-.408, -.408, .816)$$

Some other online matrix calculators report the eigenvector weights as  $-1, -1, 2$ , but these values are proportional to the weights just reported. None of these weights equals zero, so all three variables are involved in perfect collinearity. The pattern for these data is

$$R_{Y \cdot X, W} = R_{W \cdot X, Y} = R_{X \cdot Y, W} = 1.0$$

To verify this pattern, I used the SPSS syntax listed next to automatically convert the covariance matrix for this example to a correlation matrix:

```
comment convert covariance matrix to correlation matrix.
matrix data variables=rowtype_ x w y/format=full.
begin data
cov 1.00 .30 .65
cov .30 2.00 1.15
cov .65 1.15 .90
end data.
mconvert.
```

The correlation matrix (II) for  $X$ ,  $W$ , and  $Y$ , respectively, in lower diagonal form is

$$\begin{bmatrix} 1.0 & & \\ .2121 & 1.0 & \\ .6852 & .8572 & 1.0 \end{bmatrix} \quad (\text{II})$$

Given these correlations, you should verify that  $R_{Y \cdot X, W} = R_{W \cdot X, Y} = R_{X \cdot Y, W} = 1.0$ .

The LISREL program offers an option for **ridge adjustment**, which multiplies the diagonal entries by a constant  $> 1.0$  until negative eigenvalues disappear (the matrix becomes positive definite). These adjustments increase the variances until they are large enough to exceed any out-of-bounds covariance entry in the off-

diagonal part of the matrix. This technique “fixes up” a data matrix so that necessary algebraic operations can be performed (Woithke, 1993), but parameter estimates, standard errors, and fit statistics are biased after ridge adjustment. A better solution is to try to solve the problem of nonpositive definiteness through data screening.

There are other contexts where you may encounter nonpositive definite matrices in SEM, but these generally concern (1) matrices of parameter estimates for your model or (2) matrices of correlations or covariances predicted from your model. A problem is indicated if any of these matrices is nonpositive definite. We will deal with these contexts in later chapters.

difficulties are causes of nonpositive definite data matrices, but others concern distributional assumptions for continuous outcomes.

## **EXTREME COLLINEARITY**

Extreme collinearity can occur because what appear to be separate variables actually measure the same thing. Suppose that  $X$  measures accuracy and  $Y$  measures speed. If  $r_{XY} = .95$ , for instance, then  $X$  and  $Y$  are redundant notwithstanding their different labels (speed is accuracy, and vice versa). Either one or the other could be included in the same analysis, but not both. Researchers can inadvertently cause extreme collinearity when composites and their constituents are analyzed together. Suppose that a questionnaire has 10 items and the total score is summed across the items. Although the bivariate correlation between the total score and each of the individual items may not be high, the multiple correlation between the total score and the items must equal 1.0, which is multivariate collinearity in its most extreme.

Methods to detect collinearity among three or more continuous variables are summarized next. Most of these methods are available in regression diagnostics procedures of computer programs for general statistical analyses:

1. Calculate  $R^2$  between each variable and all the rest. The observation that  $R^2 > .90$  for a particular variable analyzed as the criterion suggests extreme multivariate collinearity.
2. A related statistic is **tolerance**, or  $1 - R^2$ , which indicates the proportion of total standardized variance that is unique. Tolerance values  $< .10$  may indicate extreme multivariate collinearity.
3. Another statistic is the **variance inflation factor** (VIF), or  $1/(1 - R^2)$ , the ratio of the total standardized variance over the proportion of unique variance (tolerance). The variable in question may be redundant, if VIF  $> 10.0$ .

There are two basic options for dealing with extreme collinearity: eliminate variables or combine redundant ones into a composite. For example, if  $X$  and  $Y$  are highly correlated, one could be dropped or their scores could be averaged or summed to form a single new variable, but note that this new variable must replace both  $X$  and  $Y$  in the analysis. Extreme collinearity can also happen between latent variables when their estimated correlation is so high that they are not distinct. Later we will consider this type of extreme collinearity in the technique of confirmatory factor analysis (CFA).

## OUTLIERS

Outliers are scores that are very different from the rest. A **univariate outlier** is a score that is extreme on a single variable. There is no single definition of “extreme,” but one heuristic is that scores more than three standard deviations beyond the mean may be outliers. Univariate outliers are easy to find by inspecting frequency distributions of  $z$  scores (e.g.,  $|z| > 3.0$  indicates an outlier). But this method is susceptible to distortion by the very outliers that it is supposed to detect; that is, it is not robust. Suppose that scores for five cases are

19, 25, 28, 32, and 10,000

The last score (10,000) is obviously an outlier, but it so distorts the mean and standard deviation for all scores that the  $|z| > 3.0$  rule fails, also called **masking**:

$$M = 2,020.80 \quad SD = 4,460.51 \quad \text{and} \quad z = \frac{10,000 - 2,020.80}{4,460.51} = 1.79$$

A more robust decision rule for detecting univariate outliers is

$$\frac{|X - Mdn|}{1.483 (\text{MAD})} > 2.24 \tag{4.3}$$

where  $Mdn$  designates the sample median—which is more robust against outliers than the mean—and MAD is the **median absolute deviation** (MAD) of all scores from the sample median. The quantity MAD does not estimate the population standard deviation  $\sigma$ , but the product of MAD and the scale factor 1.43 is an unbiased estimator of  $\sigma$  in a normal distribution. The value of the ratio in Equation 4.3 is the distance between a score and the median expressed in robust standard deviation units. The constant 2.24 in Equation 4.3 is the square root of the approximate 97.5th percentile in a central  $\chi^2$  distribution with a single degree of freedom. A potential outlier thus has a score on the ratio in Equation 4.3 that exceeds 2.24. For the five scores in the example,  $Mdn = 28.00$ , and the absolute values of median deviations are, respectively,

9.00, 3.00, 0, 4.00, and 9,972.00

The median of the deviations just listed is  $MAD = 4.00$ , and so for  $X = 10,000$  we calculate

$$\frac{9,972.00}{1.483 (4.00)} = 1,681.05 > 2.24$$

which obviously detects the score of 10,000 as an outlier. Wilcox (2012) describes additional robust outlier detection methods.

A **multivariate outlier** has extreme scores on two or more variables, or a pattern of scores that is atypical. For example, a case may have scores between two and three standard deviations above the mean on all variables. Although no individual score might be considered extreme, the case could be a multivariate outlier if this pattern is unusual. Here are some options for detecting multivariate outliers without extreme individual scores:

1. Some SEM computer programs, such as Amos and EQS, identify cases that contribute the most to multivariate non-normality, and such cases may be multivariate outliers.
2. Calculate for each case its squared **Mahalanobis distance**,  $D_M^2$ , which indicates the distance in variance units between the profile of scores for that case and the vector of sample means, or **centroid**, correcting for intercorrelations.

In large samples with normal distributions,  $D_M^2$  is distributed as central  $\chi^2$  with degrees of freedom equal to the number of variables, or  $v$ . A relatively high  $D_M^2$  with a low  $p$  value in the corresponding  $\chi^2(v)$  distribution may lead to the rejection of the null hypothesis that the case comes from the same population as the rest. A conservative level of statistical significance is usually recommended for this test, such as .001. Some standard computer procedures for multiple regression can automatically calculate and save  $D_M^2$  values to the raw data file.

Let us assume that an outlier is not due to a data entry error (e.g., 99 was entered instead of 9) or the failure to specify a missing data code (e.g., -9) in the data editor of a statistics computer program; that is, the outlier is a valid score. Now, what to do with the outlier? One possibility is that the case does not belong to the population from which the researcher intended to sample. Suppose that a senior graduate student audits an undergraduate class in which a questionnaire is administered. The auditing student is from a different population, and his or her questionnaire responses may be extreme compared with those of classmates. If it is determined that a case with outliers is not from the same population, then it is best to delete that case; otherwise, there are ways to reduce the influence of extreme-but-valid scores if they are retained. One option is to convert extreme scores to a value that equals the next most extreme score that is within three standard deviations of the mean. Another is to apply a mathematical transformation to a variable with outliers. Transformations are considered later in this chapter.

## NORMALITY

The default estimation method in SEM, maximum likelihood, assumes **multivariate normality (multinormality)** for continuous outcome variables. This means that

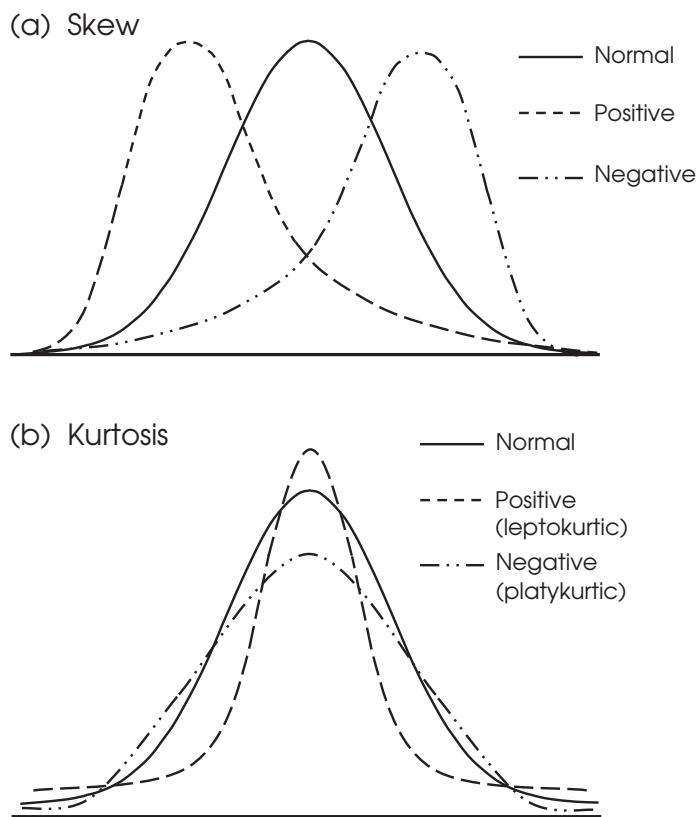
1. all the individual univariate distributions are normal;
2. all joint distributions of any pair of variables is bivariate normal; that is, each variable is normally distributed for each value of every other variable; and
3. all bivariate scatterplots are linear with homoscedastic residuals.

Because it is often impractical to examine all joint frequency distributions, it can be difficult to assess all aspects of multivariate normality. There are significance tests intended to detect violation of multivariate normality, including Mardia's (1985) test, but all such tests have limited usefulness. One reason is that slight departures from multivariate normality could be significant in large samples, and power in small samples may be low, so larger departures could be missed. Fortunately, many instances of multivariate non-normality are detectable through inspection of univariate frequency distributions.

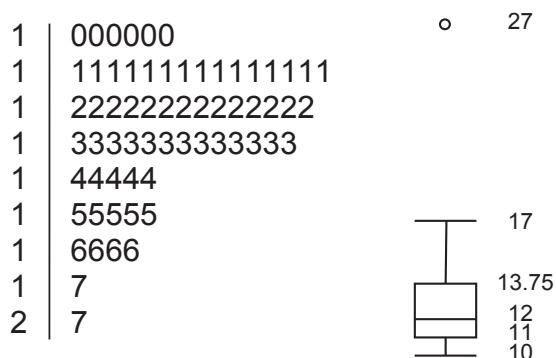
Skew and kurtosis are two ways that a univariate distribution can be non-normal, and they can occur either separately or together in the same variable. Skew implies that the shape of a unimodal distribution is asymmetrical about its mean. **Positive skew** indicates that most of the scores are below the mean, and **negative skew** indicates just the opposite. Presented in Figure 4.1(a) are examples of distributions with either positive skew or negative skew compared with a normal curve. For a unimodal, symmetric distribution, **positive kurtosis** indicates heavier tails and a higher peak and **negative kurtosis** indicates just the opposite, both relative to a normal curve with the same variance. A distribution with positive kurtosis is described as **leptokurtic**, and a distribution with negative kurtosis is described as **platykurtic**. Presented in Figure 4.1(b) are examples of distributions with either positive kurtosis or negative kurtosis compared with a normal curve. Skewed distributions are generally leptokurtic, which means that remedies for skew also may fix kurtosis. Blest (2003) describes a kurtosis measure that adjusts for skewness.

Extreme skew is easy to detect through inspection of frequency distributions or histograms. Two other types of displays helpful for spotting skew are **stem-and-leaf plots** and **box plots (box-and-whisker plots)**. For example, presented in the left side of Figure 4.2 is a stem-and-leaf plot for  $N = 64$  scores. The lowest and highest scores are, respectively, 10 and 27. The latter score is an outlier ( $z > 5.0$ ). In the stem-and-leaf plot, the numbers to the left side of the vertical line ("stems") represent the "tens" digit of each score, and each number to the right ("leaf") represents the "ones" digit. The shape of the stem-and-leaf plot indicates positive skew.

Presented in the right side of Figure 4.2 is a box plot for the same scores. The bottom and top borders of the rectangle in a box plot correspond to, respectively, the 25th percentile (1st quartile) and the 75th percentile (3rd quartile). The line inside the rect-



**FIGURE 4.1.** Distributions with (a) positive skew or negative skew and with (b) positive kurtosis or negative kurtosis relative to a normal curve.



**FIGURE 4.2.** A stem-and-leaf plot (left) and a box plot (right) for the same distribution ( $N = 64$ ).

angle of a box plot represents the median (50th percentile, or 2nd quartile). The “whiskers” are the vertical lines that connect the first and third quartiles with, respectively, the lowest and highest scores that are not extremes, or outliers. The length of the whiskers shows how far nonextreme scores spread away from the median. Skew is indicated in a box plot if the median line does not fall within the center of the rectangle or if the “whiskers” have unequal lengths. In the box plot of Figure 4.2, the high score of 27 is extreme and thus is represented in the box plot as a single open circle above the upper “whisker.” The box plot in the figure indicates positive skew because there is greater spread of scores above the median than below the median.

Kurtosis is harder to spot by eye when inspecting frequency distributions, stem-and-leaf plots, or box plots, especially in distributions that are more or less symmetrical. Departures from normality due to skew or kurtosis may be apparent in **normal probability plots**, in which data are plotted against a theoretical normal distribution in such a way that points should approximate a straight line. The distribution is otherwise non-normal, but it is hard to discern the degree of non-normality due to skew or kurtosis apparent in normal probability plots. An example of a normal probability plot is presented later in this chapter.

Fortunately, there are more precise measures of skew and kurtosis. Perhaps the best known standardized measures of these characteristics that permit comparison of different distributions to the normal curve are the **skew index** ( $\hat{\gamma}_1$ ) and **kurtosis index** ( $\hat{\gamma}_2$ ), which are calculated as follows:

$$\hat{\gamma}_1 = \frac{S^3}{(S^2)^{3/2}} \quad \text{and} \quad \hat{\gamma}_2 = \frac{S^4}{(S^2)^2} - 3.0 \quad (4.4)$$

where  $S^2$ ,  $S^3$ , and  $S^4$  are, respectively, the second through fourth **moments about the mean**:

$$S^2 = \frac{\sum(X - M)^2}{N}, \quad S^3 = \frac{\sum(X - M)^3}{N}, \quad \text{and} \quad S^4 = \frac{\sum(X - M)^4}{N} \quad (4.5)$$

The sign of  $\hat{\gamma}_1$  indicates the direction of the skew, positive or negative, and a value of zero indicates a symmetrical distribution. The value of  $\hat{\gamma}_2$  in a normal distribution equals zero, and its sign indicates the type of kurtosis, positive or negative.

The ratio of either  $\hat{\gamma}_1$  or  $\hat{\gamma}_2$  over its standard error is interpreted in large samples as a  $z$  test of the null hypothesis that there is no population skew or kurtosis. These tests may not be helpful in large samples because even slight departures from normality could be statistically significant, and low power in small samples means that appreciable skew or kurtosis can go undetected. Significance testing with  $\hat{\gamma}_1$  or  $\hat{\gamma}_2$  is not generally helpful in data screening. An alternative is to interpret absolute values of  $\hat{\gamma}_1$  or  $\hat{\gamma}_2$ , but there are few clear-cut standards for doing so. Some guidelines can be offered based on computation simulation studies of estimation methods used in SEM (e.g., Nevitt & Hancock, 2000). Variables where  $|\hat{\gamma}_1| > 3.0$  are described as “severely” skewed by some authors of these studies. There is less consensus about  $\hat{\gamma}_2$ , for which absolute values from about 8.0 to 20.0 have been described as indicating “severe” kurtosis. A conservative rule of

thumb, then, seems to be that  $|\hat{\gamma}_2| > 10.0$  suggests a problem and  $|\hat{\gamma}_2| > 20.0$  indicates a more serious one. For the data in Figure 4.2,  $\hat{\gamma}_1 = 3.10$  and  $\hat{\gamma}_2 = 15.73$ , so the distribution is severely non-normal by the standards just suggested. Do not conclude that a distribution is normal, if  $|\hat{\gamma}_1| \leq 3.0$  and  $|\hat{\gamma}_2| \leq 10.0$ . This is because  $\hat{\gamma}_1 = \hat{\gamma}_2 = 0$  in a true normal distribution; otherwise, the only thing that can be reasonably said is that the shape of the distribution may not be severely non-normal.

## TRANSFORMATIONS

In a **normalizing transformation**, the original scores are converted with a mathematical operation to new ones that may be more normally distributed. The effect of applying such a transformation is to compress one part of a distribution more than another, thereby changing its shape but not the rank of the scores. This describes a **monotonic transformation**. There are basically two situations in SEM when normalizing transformations might be considered:

1. The researcher plans to use a **normal theory method**, such as default maximum likelihood, that requires normal distributions, but distributions of continuous outcomes are severely non-normal.
2. There are multiple observed measures, or indicators, of the same theoretical construct, but some of their relations with each other are curvilinear. Transformations that normalize distributions also tend to linearize relations among multiple indicators.

Before applying a normalizing transformation, you should think about the variables of interest and whether the expectation of normality is reasonable. Some variables are expected to have non-normal distributions, such as reports of alcohol or drug use and certain personality characteristics (Bentler, 1987). If so, then transforming an inherently non-normal variable to force a normal distribution may fundamentally alter it (the target variable is not actually studied). In this case, it would be better to use a different estimation method for continuous outcomes in SEM, one that does not assume normality, such as robust maximum likelihood. Another consideration is whether the metric of outcome variables is meaningful, such as athletic performance in seconds or postoperative survival time in years. Applying a transformation means that the original meaningful metric is lost, which could be a sacrifice.

Normalizing transformations may be more useful when there is no expectation of normality or metrics of outcome variables are arbitrary. An example is the total score for a set of true-false items. Because responses can be coded using any two different numbers, the total score is arbitrary. Standard scores such as percentiles and normal deviates are arbitrary, too, because one standardized metric can be substituted for another. Described in Topic Box 4.2 are types of normalizing transformations that may work in

different situations with practical suggestions for using them—see Osborne (2002) for more information. Exercise 3 asks you to find a normalizing transformation for the data in Figure 4.2.

Sometimes normalizing transformations can linearize relations between indicators of the same construct. For example, Budtz-Jørgensen, Keiding, Grandjean, and Weihe (2002) studied the effect of prenatal methylmercury exposure, through maternal consumption of contaminated pilot whale meat, on child neurobehavioral status among

### **TOPIC BOX 4.2**

#### **Normalizing Transformations**

Three kinds of normalizing transformation are described next with suggestions for their use:

1. *Positive skew.* Before applying these transformations, add a constant to the scores so that the lowest value is 1.0. A basic transformation is the square root transformation, or  $X^{1/2}$ . It works by compressing the differences between scores in the upper end of the distribution more than the differences between lower scores. Logarithmic transformations are another option. A logarithm is the power (exponent) to which a base number must be raised in order to get the original number, such as  $10^2 = 100$ , so the logarithm of 100 in base 10 is 2.0. Distributions with extremely high scores may require a transformation with a higher base, such as  $\log_{10} X$ , but a lower base may suffice for less extreme cases, such as the natural base e (approximately 2.7183) for the transformation  $\log_e X = \ln X$ . The inverse function  $1/X$  is an option for even more severe positive skew. Because inverting the scores reverses their order, (1) reflect (reverse) the original scores (multiply them by -1.0) and (2) add a constant to the reflected scores so that the maximum score is at least 1.0 before taking the inverse.

2. *Negative skew.* All the transformations just mentioned also work for negative skew when they are applied as follows: First, reflect the scores, and then add a constant so that the lowest score equals 1.0. Next, apply the transformation, and reflect the scores again to restore their original ordering.

3. *Other types of non-normality.* Odd-root functions (e.g.,  $X^{1/3}$ ) and sine functions tend to bring in outliers from both tails of the distribution toward the mean. Odd-powered polynomial transformations, such as  $X^3$ , may help for negative kurtosis. If the scores are proportions, the arcsine square root transformation function, or  $\arcsin X^{1/2}$ , may normalize the distribution.

There are other kinds of normalizing transformations, and this is one of their potential problems: It can be difficult to find a transformation that works with a

particular set of scores. The **Box-Cox transformations** (Box & Cox, 1964) may require less trial and error. The most basic form is defined next only for positive scores:

$$X^{(\lambda)} = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log X, & \text{if } \lambda = 0. \end{cases}$$

where the exponent  $\lambda$  is a constant that normalizes the scores. There are computer algorithms for finding the value of  $\lambda$  that maximizes the correlation between original and transformed scores (Friendly, 2006). There are other variations of the Box-Cox transformation (Osborne, 2010), some of which can be applied in regression analyses to deal with heteroscedasticity.

residents in the Faroe Islands. Two biological markers were mercury concentration in cord blood and maternal hair, and the third measure was the amount of self-reported monthly consumption of whale meat. Blood or hair concentration scores can be so high that they have curvilinear relations with questionnaire data, so Budtz-Jørgensen et al. (2002) applied logarithmic transformations to the blood and hair concentrations before analyzing them.

Some distributions can be so severely non-normal that no transformation will work. Count variables are an example. A **count variable** is the number of times a discrete event happens over a period of time such as the number of serious automobile accidents over the past 5 years. Distributions of such variables tend to be positively skewed, and many cases may have scores of zero. Count variables generally follow non-normal distributions known as **Poisson distributions**, where the mean and variance are approximately equal. Some SEM computer tools, such as Mplus, offer special methods for analyzing count variables. These methods are related to the technique of **Poisson regression**, which also analyzes log linear models for count data (Agresti, 2007).

Little (2013) described **percentage or proportion of maximum scoring** (POMS) transformations for rescaling questionnaire items that measure a common domain but where responses are recorded on different Likert scales. After transformation, all items will have the same metric. Suppose that items with a 5-point Likert scale are administered to participants at time 1 but at time 2 the same items have a 7-point Likert scale. To compare responses over time, a transformation is needed. One option is to convert the narrower scale in this example (1-5) to the wider scale (1-7), as follows:

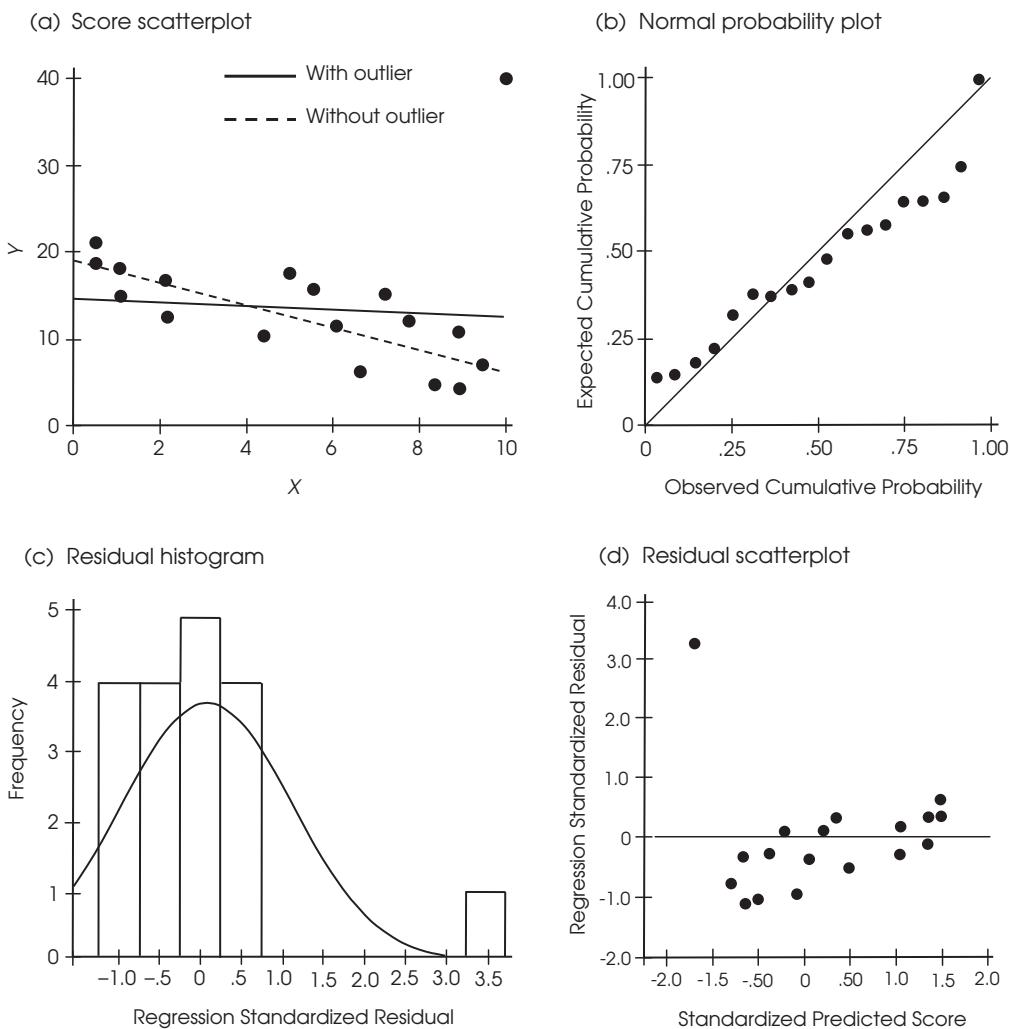
$$R7 = \left( \frac{O5 - 1}{4} \right) \times 6 + 1 \quad (4.6)$$

where  $R7$  is the rescaled item with a 1-7 response format and  $O5$  is the original item with a 1-5 scale. In Equation 4.6, the term  $O5 - 1$  transforms the original 1-5 scale to a 0-4

scale; dividing by 4 then converts the scale to 0–1; then multiplying by 6 yields a 0–6 scale; and finally adding 1 generates the final 0–7 Likert scale.

### Linearity and Homoscedasticity

Linear relations between continuous outcomes and homoscedastic residuals are part of multivariate normality. Curvilinear relations are easy to detect by looking at bivariate scatterplots. Heteroscedastic residuals may be caused by non-normality in either variable, greater measurement error at some levels of either variable than others, or outliers.



**FIGURE 4.3.** (a) Score scatterplot with outlier ( $N = 18$ ) and the linear regression lines with and without the outlier ( $N = 17$ ). (b) A normal probability plot of the regression standardized residuals. (c) A histogram of the regression standardized residuals. (d) Residual scatterplot.

ers. For example, presented in Figure 4.3(a) is a scatterplot for  $N = 18$  scores. One score (40) on Y is  $> 3$  standard deviations above the mean. For these data,  $r_{XY} = -.074$ , and the bivariate regression line is nearly horizontal, but these results are affected by the outlier. After removing the outlier ( $N = 17$ ), then  $r_{XY} = -.772$ , and the new regression line better fits the remaining data—see Figure 4.3(a).

The regression standardized residuals for all the data ( $N = 18$ ) in Figure 4.3(a) are plotted in various ways over Figures 4.3(b)–4.3(d). Figure 4.3(b) is a normal probability plot of the expected versus observed cumulative probabilities, which do not fall among a diagonal line. The histogram of the residuals with a superimposed normal curve is presented in Figure 4.3(c). The residuals are not normally distributed. A scatterplot of the residuals and standardized predicted scores is presented in Figure 4.3(d). The residuals are heteroscedastic because they are not evenly distributed about zero throughout the entire length of this scatterplot.

## RELATIVE VARIANCES

In an **ill-scaled covariance matrix**, the ratio of the largest to smallest variance is greater than say, 100.0. Analysis of such a matrix in SEM can cause problems. Most estimation methods in SEM are iterative, which means that initial estimates are derived by the computer and then modified through subsequent cycles of calculation. The goal is to derive better estimates at each stage, estimates that improve the overall fit of the model to the data. When improvements from step to step become sufficiently small—that is, they fall below the **convergence criterion**—iterative estimation stops because the solution is stable. But if the estimates do not settle down to stable values, the process may fail. One cause is variances of observed variables that are very different in magnitude. When the computer adjusts the estimates from one step to the next in an iterative method for an ill-scaled matrix, the sizes of these changes may be huge for variables with small variances but trivial for others with large variances. Consequently, the entire set of estimates may head toward worse rather than better fit.

To prevent this problem, variables with extremely low or high variances can be rescaled by multiplying their scores by a constant, which changes the variance by a factor that equals the squared constant. For example,

$$s_X^2 = 12.0 \quad \text{and} \quad s_Y^2 = .12$$

so their variances differ by a factor of 100.0. Using the constant .10, we can rescale X as follows:

$$s_{X \times .10}^2 = .10^2 \times 12.0 = .12$$

so now variables  $X \times .10$  and Y have the same variance, or .12. Now we rescale Y so that it has the same variance as X, or 12.0, by applying the constant 10.0, or

$$s_{Y \times 10.0}^2 = 10^2 \times .12 = 12.0$$

Rescaling a variable in this way changes its average and variance but not its correlation with other variables. This is because multiplying a variable by a constant is just a linear transformation that does not affect relative differences among the scores. An example with real data follows.

Roth, Wiebe, Fillingham, and Shay (1989) administered measures of exercise, hardness (resiliency, tough mindedness), fitness, stress, and level of illness in a sample of university students. Reported in Table 4.2 is a summary matrix of these data (correlations, means, and variances). The largest and smallest variances in this matrix (see the table) differ by a factor of more than 27,000, so the covariance matrix is ill-scaled. I have seen some SEM computer programs fail to analyze this matrix due to this characteristic. To prevent this problem, I multiplied the original variables by the constants listed in the table in order to make their variances more homogeneous (the constant 1.0 means no change). Among the rescaled variables, the largest variance is only about 13 times greater than the smallest variance. The rescaled matrix is not ill-scaled.

## MISSING DATA

The topic of how to analyze data sets with missing observations is complicated. Entire books are devoted to it (Enders, 2010; McKnight, McKnight, Sidani, & Figueredo, 2007); there are also articles or chapters about methods for dealing with missing data in SEM (Allison, 2003; Graham & Coffman, 2012; Peters & Enders, 2002). This is fortunate

**TABLE 4.2. Example of an Ill-Scaled Data Matrix**

Variable	1	2	3	4	5
1. Exercise	—				
2. Hardiness	-.03	—			
3. Fitness	.39	.07	—		
4. Stress	-.05	-.23	-.13	—	
5. Illness	-.08	-.16	-.29	.34	—
<i>M</i>	40.90	0.0	67.10	4.80	716.70
Original $s^2$	4,422.25	14.44	338.56	44.89	390,375.04
Constant	1.00	10.00	1.00	5.00	.10
Rescaled $s^2$	4,422.25	1,440.00	338.56	1,122.25	3,903.75
Rescaled SD	66.50	38.00	18.40	33.50	62.48

*Note.* These data (correlations, means, and variances) are from Roth et al. (1989);  $N = 373$ . Note that low scores on the hardiness measure used by these authors indicate greater hardiness. In order to avoid confusion due to negative correlations, the signs of the correlations that involve the hardiness measure were reversed before they were recorded in this table.

because it is not possible here to give a comprehensive account of the topic. The goal instead is to give you a sense of basic analysis options and to explain the relevance of these options for SEM.

Ideally, researchers would always work with complete data sets, ones with no missing values; otherwise, prevention is the best strategy. For example, questionnaire items that are clear and unambiguous may prevent missing responses, and completed forms should be reviewed for missing responses before participants submit a computer-administered survey or leave the laboratory. In the real world, missing values occur in many, if not most, data sets, despite the best efforts at prevention. Missing data occur for many reasons, including hardware failure, missed appointments, and item nonresponse. A few missing values, such as < 5% in the total data set, may be of little concern. This is because selection among methods to deal with missing data is arbitrary in that the method used tends not to make much difference. Higher rates of data loss present more challenges, especially if the **data loss mechanism** is not truly random (or at least predictable). In this case, the choice of method can appreciably affect the results. This is why researchers should always explain how missing data were handled in the analysis.

## Data Loss Mechanisms

There are basically three data loss mechanisms. All can operate within the same data set because each can affect different subsets of variables. Also, it is not always clear which pattern holds for a particular variable with missing values. The most optimistic case—and probably the most unrealistic in actual data—is when data are **missing completely at random** (MCAR). For variable Y, this means that (1) missing observations differ from the observed scores only by chance; that is, whether scores on Y are missing or not missing is unrelated to Y itself. (2) The presence versus absence of data on Y is unrelated to all other variables in the data set. In this case, the observed (nonmissing) data are just a random sample of scores that the researcher would have analyzed had the data been complete (Enders, 2010). Results based on the complete cases only should not be biased, although power may be reduced due to a smaller effective sample size. An example of haphazard missing data is when questionnaire responses to items about mental health are lost due to sporadic computer problems that have nothing to do with either respondents' true mental health status or their responses to questions about other topics.

A second data loss mechanism is indicated when the property of missingness on Y is unrelated to Y itself but is correlated with other variables in the data set; that is, missing data arise from a process that is both measured and predictable in a particular sample (Little, 2013). This process is called **missing at random** (MAR), which is an odd term because the data loss mechanism depends on other variables, and thus is not random. An example of an MAR process would be when men are less likely to respond to questions about mental health than women, but among men the probability of responding is unrelated to their true mental health status.

Information lost due to an MAR process is potentially recoverable through imputation, where missing scores are replaced by predicted scores. The predicted scores

are generated from other variables in the data set that predict missingness on Y. If the strength of that prediction is reasonably strong, then results on Y after imputation may be relatively unbiased. In this sense, the MAR pattern is described as **ignorable** concerning potential bias. Note that both the MAR and MCAR patterns of data loss can affect the same variable.

A strategy that anticipates the MAR pattern is to measure **auxiliary variables**. Such variables may not be of substantive interest, but they predict missingness on other variables in the data set. For example, gender, socioeconomic status, and parental involvement are potential auxiliary variables in longitudinal studies of children, and inclusion of these predictors when imputing scores on other variables may decrease bias (Little, 2013). Auxiliary variables require care in their selection. This is because the inclusion of too many auxiliary variables in smaller samples can increase imprecision by so much that more sophisticated methods for imputation can fail. This is especially true if less than about 10% or so of the variance in missingness on Y is explained by auxiliary variables ( $R^2 < .10$ ) (Hardt, Herke, & Leonhart, 2012).

When data are **missing not at random** (MNAR), the data loss mechanism is **non-ignorable**, which means that the presence versus absence of scores on Y depends on Y itself. An example from medicine occurs when patients drop out of a study when a particular treatment causes unpleasant side effects. Because that discomfort is not measured, however, the data are missing due to a process that is unknown in a particular data set. Results based on the complete cases only can be severely biased when the data loss pattern is MNAR. For example, a treatment may look more beneficial than it really is if data from patients who were unable to tolerate the treatment are lost. Some bias may be reduced if other measured variables happen to covary with unmeasured causes of data loss, but whether this is true in a particular sample is usually unknown. The choice of method to deal with the incomplete records can make a difference in the results when the MNAR pattern holds.

## Diagnosing Missing Data

It is not easy in practice to determine whether the data loss mechanism is random or systematic, especially when each variable is measured only once. Specifically, there are ways to determine whether the assumption of MCAR is reasonable, but there is no definitive test that provides direct evidence of either MAR or MNAR if the former hypothesis is rejected. Little and Rubin (2002) describe a multivariate statistical test of the MCAR assumption that simultaneously compares complete versus incomplete cases on Y across all other variables. If this comparison is significant, then the MCAR hypothesis is rejected. Plausibility of the MCAR assumption can also be examined through a series of univariate comparisons of the *t* test of cases that have missing scores on Y with cases that have complete records on other variables. The problems with these significance tests are that they can have low power in smaller samples and they can flag trivial differences as significant in larger samples.

A related tactic involves creating a dummy-coded variable that indicates whether a score is missing and then examining cross tabulations with other categorical variables, such as gender or treatment condition. Some computer programs for general statistical analysis have special commands or procedures for analyzing missing data patterns. An example is the Missing Values procedure of SPSS, which can conduct all the diagnostic tests just mentioned. The PRELIS program of LISREL also has extensive capabilities for analyzing missing data patterns.

If the assumption of MCAR is rejected, then we cannot ever really be sure whether the data loss mechanism is MAR or MNAR. This is because variables may be omitted that account for data loss on  $Y$  that are related to  $Y$  itself. Because these variables are unmeasured, the true extent of nonrandom, systematic data loss will not be known. It helps if other, measured variables in the data set predict missingness on  $Y$ , but only some of the information on  $Y$  may be recovered in an imputation process based on these predictors. For this reason it is prudent to ascertain potential auxiliary variables when planning a study.

There is no magical statistical “fix” that will eliminate bias due to systematic data loss. About the best that can be done is to understand the nature of the underlying data loss pattern and accordingly modify your interpretation of the results. If the selection of one option for dealing with missing data instead of another makes a difference in the results and it is unclear which option is best, then you should report both sets of findings. This approach makes it plain that the results depend on how missing observations were handled. This tactic is a kind of **sensitivity analysis** in which data are reanalyzed under different assumptions—here, using alternative missing data techniques—and the results are compared with the original findings.

## Classical Methods

Classical techniques for handling incomplete cases have been around for a long time and are available as options in many computer programs for general statistical analysis, but they are increasingly considered obsolete. One reason is that such methods generally assume that the missing value mechanism is MCAR, which is often unrealistic. Such methods tend to yield biased estimates under the less strict assumption of MAR, and even more so when the data loss mechanism is MNAR. They also take relatively little advantage of the structure in the data. Classical methods are briefly reviewed next, but there are better, more modern methods, which will also be discussed in this chapter.

There are two general kinds of classical methods: **available case methods**, which analyze data available through deletion of incomplete cases, and **single-imputation methods**, which replace each missing score with a single calculated (imputed) score. Available case methods include **listwise deletion** in which cases with missing scores on any variable are excluded from all analyses. The effective sample size with listwise deletion includes only cases with complete records. This number can be much smaller than the original sample size if missing observations are scattered across many records.

In regression analyses, listwise deletion of incomplete cases generates reasonably good estimates when the data loss mechanism depends on the predictors but not on the criterion (Little & Rubin, 2002).

An advantage of listwise deletion is that all analyses are conducted with the same cases. This is not so with **pairwise deletion**, in which cases are excluded only if they have missing data on variables involved in a particular analysis. Suppose that  $N = 300$  for an incomplete data set. If 250 cases have no missing scores on variables  $X$  and  $Y$ , then the effective sample size for  $\text{cov}_{XY}$  is this number. If fewer or more cases have valid scores on  $X$  and  $W$ , however, the effective sample size for  $\text{cov}_{XW}$  will not be 250. It is this property of the method that can give rise to out-of-bounds correlations or covariances. Presented in Table 4.3 is a small data set with missing scores on all three variables. The covariance matrix generated by pairwise deletion for these data is nonpositive definite. Exercise 4 asks you to verify this statement.

The most basic single-imputation method is **mean substitution**, which involves replacing a missing score with the overall sample mean. A variation is **group-mean substitution**, in which a missing score in a particular group (e.g., women) is replaced by the group mean. This variation may be preferred when group membership is a predictor in the analysis or when a model in SEM is analyzed over multiple groups. Both methods are simple, but they can distort the distribution of the data by reducing variability. Suppose in a data set where  $N = 75$  that 15 cases have missing values on some variable. Substituting the mean of the 60 valid cases does not change the mean for  $N = 75$  after imputation compared with the mean for  $N = 60$  before imputation. But the variance for the  $N = 60$  scores before substitution will be greater than the variance for the  $N = 75$  scores after substitution. Mean substitution also tends to make distributions more peaked at the mean, too, which further distorts the underlying distribution of the data (Vriens & Melton, 2002).

**Regression substitution** is somewhat more sophisticated. In this method, each missing score is replaced by a predicted score using multiple regression based on non-missing scores on other variables. Regression substitution uses more information than mean substitution, but it assumes that variables with missing observations can be predicted reasonably well from other variables in the same data set; otherwise, there is little

**TABLE 4.3. Example of an Incomplete Data Set**

Case	X	W	Y
A	42	13	8
B	34	12	10
C	22	—	12
D	—	8	14
E	24	7	16
F	16	10	—
G	30	10	—

point in imputing missing scores with predicted scores. A variation is **stochastic regression imputation**, in which the computer adds a randomly sampled error term from the normal distribution or other user-specified distribution to each predicted score, which reflects uncertainty in the score. This capability is implemented in the Missing Values procedure in SPSS.

A more sophisticated single-imputation method is **pattern matching**, in which the computer replaces a missing observation with a score from a case with the most similar profile on other variables. Pattern matching is available in the PRELIS program of LISREL. Another option is **random hot-deck imputation**, which separates complete from incomplete cases; sorts both sets of records so that cases with similar profiles on background variables are grouped together; randomly interleaves the incomplete cases and complete ones; and replaces missing scores with those on the same variable from the nearest complete record. Myers (2011) describes a macro that performs random hot-deck imputation in SPSS. All single-imputation methods tend to underestimate error variance, especially if the proportion of missing observations is relatively high (Vriens & Melton, 2002).

## Modern Methods

Contemporary methods generally assume a data loss pattern that is MAR, not MCAR. When the pattern is not random (MNAR), these more sophisticated techniques will also yield biased estimates, but probably less so compared with classical techniques (Peters & Enders, 2002).

There are two basic kinds of modern methods for analyses with missing data. A **model-based method** takes the researcher's model as the starting point. Next, the procedure partitions the cases in a raw data file into subsets, each with the same pattern of missing observations, including none (complete cases). Relevant statistical information, including the means and the variances, is extracted from each subset, so all cases are retained in the analysis. The parameters of the researcher's model are then estimated after combining all available information over the subsets of cases. Thus, parameter estimates and their standard errors are calculated directly from the available data without deletion or imputation of missing values. Some SEM computer tools, including Amos, LISREL, and Mplus, offer a special version of the maximum likelihood method—sometimes called **full information maximum likelihood** (FIML)—for incomplete data files that works in the way just described. Some FIML procedures for incomplete data allow the specification of auxiliary variables (see Graham & Coffman, 2012).

Multiple imputation is a **data-based method** that typically works with the whole raw data file, not just with the observed variables that comprise the researcher's model. As the name suggests, **multiple imputation** can generally replace a missing score with multiple estimated (imputed) values from a predictive distribution that models the data loss mechanism. In nontechnical terms, a model for both the complete and incomplete data is defined under these methods. The computer then estimates means and variances in the whole sample that satisfy a statistical criterion. The process of imputation

is repeated so that the analysis is actually conducted with multiple versions of imputed data sets. In large data sets, a relatively high number of imputed data sets may need to be generated (e.g., 100) in order for the results to have reasonable precision (Little, 2013). The final set of estimates comes after the computer synthesizes the results from all replications.

Some methods for multiple imputation are based on the **expectation-maximization** (EM) **algorithm**, which has two steps. In the E (expectation) step, missing observations are imputed by predicted scores in a series of regressions in which each incomplete variable is regressed on the remaining variables for a particular case. In the M (maximization) step, the whole imputed data set is submitted for maximum likelihood estimation. These two steps are repeated until a stable solution is reached across the M steps. The EM algorithm for multiple imputation is available in EQS and LISREL, among other SEM computer tools. In addition, some methods are based on the **Markov Chain Monte Carlo** (MCMC) approach, which is a class of methods for random sampling from a theoretical probability distribution. The MCMC method is used to draw from a predictive distribution for the missing data, and these draws become the imputed scores. Multiple imputation in Mplus is based on the MCMC method.

There may be times in SEM when multiple imputation is favored over the FIML method (Graham & Coffman, 2012). Not all SEM computer programs feature FIML estimation for incomplete data files. In this case, the researcher could use procedures for multiple imputation in computer tools for general statistical analyses. For example, the MI procedure in SAS/STAT could be used to impute the data, and later the MIANALYZE procedure could combine results from the imputed data sets after they have been analyzed with a computer tool for SEM. It is also generally easier to incorporate auxiliary variables in multiple imputation than in the FIML method. But if the FIML method is available in your SEM computer tool, it is a reasonable option for conducting the analysis with an incomplete data set.

## **SELECTING GOOD MEASURES AND REPORTING ABOUT THEM**

It is just as critical in SEM as in other types of analyses to (1) select measures with strong psychometric properties and (2) report these characteristics in written summaries. This is because the product of measures, or scores, is what is analyzed. If the scores do not have good psychometrics, then the results can be meaningless.

Presented in Table 4.4 is a checklist of descriptive, practical, and technical information that should be considered before selecting a measure. Not all of these points may be relevant in a particular study, and some types of research have special measurement needs that may not be represented in the table. If so, just modify the checklist to better reflect a particular situation. The *Mental Measurements Yearbook* (Carlson, Geisinger, & Jonson, 2014) is a good source of information about commercial tests. It is also available as a searchable electronic database in many university libraries. Maddox (2008) describes measures in psychology, education, and business. A directory of noncommer-

**TABLE 4.4. Checklist for Evaluating Potential Measures**General

- Stated purpose of the measure
- Attribute(s) claimed to be measured
- Characteristics of samples in which measure was developed (e.g., normative sample)
- Language of test materials
- Costs (manuals, forms, software, etc.)
- Limitations of the measure
- Academic or professional affiliation(s) of author(s) consistent with test development
- Publication date and publisher

Administration

- Test length and testing time
- Measurement method (e.g., self-report, interview, unobtrusive)
- Response format (e.g., multiple choice, free response)
- Availability of alternative forms (versions)
- Individual or group administration
- Paper-and-pencil or computer administration
- Scoring method, requirements, and options
- Materials or testing facilities needed (e.g., computer, quiet testing room)
- Training requirements for test administrators or scorers (e.g., test user qualifications)
- Accommodations for test takers with physical or sensory disabilities

Test documentation

- Test manual available
- Manual's description of how to correctly derive and interpret scores
- Evidence for score reliability and characteristics of samples (e.g., reliability induction)
- Evidence for score validity and characteristics of samples
- Evidence for test fairness (e.g., lack of gender, race, or age bias)
- Results of independent reviews of the measure

cial measures from articles in psychology, sociology, or education journals is available in Goldman and Mitchell (2007). These measures are not protected by copyright, but as a professional courtesy you should ask the author's permission before using or adapting a particular test. There is also the freely accessible Measurement Instrument Database for the Social Sciences, an online test database.<sup>3</sup>

Readers who have already taken a measurement course are at some advantage when it comes to selecting a test because they can critically evaluate candidate measures. They should also know how to evaluate whether those scores in their own samples are reliable and valid. Readers without this background are encouraged to fill in this gap. Formal coursework is not the only way to learn more about measurement. Just like learning about SEM, more informal ways to learn measurement theory include partici-

---

<sup>3</sup>[www.midss.org](http://www.midss.org)

pation in seminars or workshops and self-study. A good undergraduate-level book that emphasizes classical measurement theory in psychology and education is Thorndike and Thorndike-Christ (2010), and the graduate-level work by Raykov (2011) deals with modern measurement theory.

Unfortunately, the state of practice about reporting on the psychometric characteristics of scores analyzed is too often poor. For example, Vacha-Haase and Thompson (2011) found that 55% of authors did not even mention score reliability in over 13,000 primary studies from a total of 47 meta-analyses of reliability generalization in the behavioral sciences. Authors mentioned reliability in about 16% of the studies, but they merely inducted values reported in other sources, such as test manuals. Such **reliability induction**, or inferring from particular coefficients calculated in other samples to a different population, requires explicit justification. But researchers rarely compare characteristics of their sample with those from cited studies of score reliability. For example, scores from a computer-based task of reaction time developed in samples of young adults may not be as precise for elderly adults. A better practice is for researchers to report estimates of score reliability from their own samples. They should also cite reliability coefficients reported in published sources (reliability induction) but with comment on similarities between samples described in those other sources and the researcher's sample.

Thompson and Vacha-Haase (2000) speculated that another cause of poor reporting practices is the apparently widespread but false belief that it is *tests* that are reliable or unreliable, not *scores* in a particular sample. In other words, if researchers believe that reliability, once established, is an immutable property of tests, then they may put little effort into estimating reliability in their own samples. They may also adopt a "black box" mentality that assumes that reliability can be established by others, such as a select few academics who conduct measurement-related research. The truth is that reliability and validity are attributes of scores in particular samples where the intended uses of those scores must also be considered.

Measurement is a broad topic, so it is impossible to succinctly cover all its aspects, but familiarity with the issues considered next should help you to select good measures and report necessary information about scores generated from them. This presentation will also help you to better understand certain analysis options in CFA, the factor-analytic technique in SEM.

## SCORE RELIABILITY

**Score reliability** is the degree to which scores in a particular sample are precise. It is estimated as one minus the proportion of total observed variance due to random error. These estimates are reliability coefficients, which for measure  $X$  are often designated with the symbol  $r_{XX}$ . Because  $r_{XX}$  is a proportion of variance, its theoretical range is 0–1.0. For example, if  $r_{XX} = .80$ , then  $1 - .80 = .20$ , or 20% of total variance is unsystematic. But the remaining standardized variance, or 80%, may not all be systematic.

This is because a particular type of reliability coefficient may estimate a single source of random error, and scores can be affected by multiple sources of error. As  $r_{xx}$  approaches zero, the scores are increasingly more like random numbers, and random numbers measure nothing. It can happen that an empirical reliability coefficient is less than zero. A negative reliability coefficient is interpreted as though its value were zero, but such a result ( $r_{xx} < 0$ ) indicates a serious problem with the scores.

The type of reliability coefficient reported most often in the literature is **coefficient alpha**, also called **Cronbach's alpha**. It measures **internal consistency reliability**, or the degree to which responses are consistent across the items of a measure. If internal consistency is low, then the content of the items may be so heterogeneous that the total score is not the best possible unit of analysis. A conceptual equation is

$$\alpha_C = \frac{n_i \bar{r}_{ij}}{1 + (n_i - 1)\bar{r}_{ij}} \quad (4.7)$$

where  $n_i$  is the number of items, not cases, and  $\bar{r}_{ij}$  is the average Pearson correlation between all pairs of items. For example, given  $n_i = 20$  with a mean interitem correlation of .30, then

$$\alpha_C = \frac{20 (.30)}{1 + (20 - 1) (.30)} = .90$$

Internal consistency reliability is higher as there are more items or the average interitem correlation increases. In observed variable analyses, it is best to analyze scores from measures that are internally consistent. This is also generally good advice for latent variable analyses, including SEM, but see Little, Lindenberger, and Nesselroade (1999) about exceptions. Exercise 5 asks you to calculate and interpret  $\alpha_C$  for a small set of items.

An older method of estimating internal consistency is that of **split-half reliability**, where a single test is split into two parts, such as an odd–even item split, and scores from two halves are correlated. The observed correlation is then corrected for test length, and the corrected result is the split-half reliability coefficient. For a particular set of items, the value of  $\alpha_C$  is the average of all possible split-half coefficients (e.g., odd vs. even items, first-half vs. second-half items, etc.), so in this sense  $\alpha_C$  is a more general estimate of internal consistency than any split-half coefficient.

A drawback of  $\alpha_C$  is that it is actually not a very good indicator of whether a set of items measures a single factor. This is because lower values of  $\bar{r}_{ij}$  can be offset by greater numbers of items,  $n_i$ . Suppose that  $\bar{r}_{ij} = .01$  for 1,000 items. The average correlation across the items is practically zero, so they clearly do not measure a common domain. But with so many items in this example,  $\alpha_C = .91$ . (You should verify this statement.) In this example, the large number of items overwhelms the near-zero value of  $\bar{r}_{ij}$ . This means that a high value of  $\alpha_C$  does not guarantee internal consistency because long, multidimensional scales will also have high values of  $\alpha_C$  (Streiner, 2003). At the other extreme, very high values of  $\alpha_C$  can suggest redundancy in a small item set. For example, given  $\alpha_C = .95$  for  $n_i = 2$  items, then  $\bar{r}_{ij} = .90$ , which indicates that the two items

are not distinct (they are extremely collinear). Better ways to estimate the reliability of construct measurement in SEM are described in Chapter 13.

Estimation of other kinds of score reliability may require multiple occasions, test forms, or examiners. **Test-retest reliability** involves the readministration of a measure to the same group on a second occasion. If the two sets of scores are highly correlated, error due to temporal factors may be minimal. **Alternate- (parallel-) forms reliability** involves the evaluation of score precision across different versions of a test. This method estimates whether variation in items drawn from the same domain leads to changes in rank order between the two forms. If so, then scores are unstable across different versions, which raises doubts that a common domain is measured. **Interrater reliability** is relevant for subjectively scored tests: If independent raters do not agree in scoring, then examiner-specific factors may contribute unduly to score variability.

In observed variable analyses, there is no gold standard as to how high coefficients should be in order to conclude that score reliability is satisfactory, but here are some guidelines: Generally, coefficients around .90 are considered “excellent,” values around .80 as “very good,” and values about .70 as “adequate.” Note that somewhat lower levels of score reliability can be tolerated in latent variable methods compared with observed variable methods, if the sample size is sufficiently large (Little et al., 1999).

Low score reliability has detrimental effects in observed variable analyses. Poor reliability reduces statistical power; it also generally reduces effect sizes below their true (population) values. Unreliability in scores of two different variables,  $X$  or  $Y$ , attenuates their observed correlation. This formula from classical measurement theory shows the exact relation:

$$\max |\hat{r}_{XY}| = \sqrt{r_{XX} \times r_{YY}} \quad (4.8)$$

where  $\max |\hat{r}_{XY}|$  is the theoretical (estimated) maximum absolute value of the correlation. In other words, the absolute correlation between  $X$  and  $Y$  can equal 1.0 only if scores on both variables are perfectly reliable. Suppose that  $r_{XX} = .10$  and  $r_{YY} = .90$ . Given this information, the theoretical maximum absolute value of  $r_{XY}$  can be no higher than

$$\max |\hat{r}_{XY}| = \sqrt{.10 \times .90} = .30$$

A variation of Equation 4.8 is the **correction for attenuation**:

$$\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX} \times r_{YY}}} \quad (4.9)$$

where  $\hat{r}_{XY}$  is the *estimated* validity coefficient if scores on both measures were perfectly reliable. In general,  $\hat{r}_{XY}$  is greater in absolute value than  $r_{XY}$ , the observed correlation. For example, given

$$r_{XY} = .30, r_{XX} = .90, \text{ and } r_{YY} = .40$$

then  $\hat{r}_{XY} = .50$ ; that is, we expect that the “true” correlation between  $X$  and  $Y$  would be  $.50$ , controlling for measurement error. Because disattenuated correlations are only estimates, it can happen that their absolute values exceed  $1.0$ . A better way to control for measurement error is to use SEM where constructs are specified as latent variables, each measured by multiple indicators. In fact, SEM is more accurate at estimating correlations between factors or between indicators and factors than observed-variable methods (Little et al., 1999).

## SCORE VALIDITY

**Score validity** concerns the soundness of inferences based on the scores, and information about score validity conveys to the researcher whether applying a test is capable of achieving certain aims. Kane (2013) elaborated on this theme by describing **interpretation-use arguments**, an approach to validity that concerns the plausibility and appropriateness of both the interpretation and the proposed uses of scores. In this view, validity is not a fixed property of tests; rather, it involves the proposed interpretation and intended uses of the scores. As the range of potential generalizations from test scores increases, such as from an observed sample of performances (test data) to predicted performances in other settings, more evidence is needed. Messick (1995) emphasized the qualities of relevance, utility, value implications, and social consequences of test use and interpretation in validation. An example of the social consequences of testing includes the fair and accurate assessment of cognitive abilities among minority children.

**Construct validity** involves whether scores measure a target hypothetical construct, which is latent and thus can be measured only indirectly through its indicators. There is no single, definitive test of construct validity, nor is it established in a single study. Instead, measurement-based research usually concerns a particular aspect of construct validity. For instance, **criterion-related validity** concerns whether test scores ( $X$ ) relate to a criterion ( $Y$ ) against which the scores can be evaluated. Specifically, are sample values of  $r_{XY}$  large enough to support the claim that a test explains an appreciable amount of the variability in the criterion? Whether an admissions test for graduate school predicts eventual program completion is a question of criterion-related validity.

Convergent validity and discriminant validity involve the evaluation of measures against each other instead of against an external standard. Variables presumed to measure the *same* construct show **convergent validity** if their intercorrelations are appreciable in magnitude. But if measures that supposedly reflect the same construct also share the same measurement method, their intercorrelations could be inflated by **common method variance**. Thus, the best case for convergent validity occurs when measures of the same presumed trait are each based on a different measurement method (Campbell & Fiske, 1959). Likewise, **discriminant validity** is supported if the intercorrelations among a set of variables presumed to measure *different* constructs are not too high, but this evidence is stronger when the measures are not based on the same method. If  $r_{XY}$

= .90 and these two variables are each based on a different measurement method, one cannot claim that X and Y assess distinct constructs. Hypotheses about convergent and discriminant validity are routinely tested in CFA.

**Content validity** deals with whether test items are representative of the domain(s) they are supposed to measure. Content validity is often critical for scholastic achievement measures, such as tests that should assess specific skills at a particular grade level (e.g., Grade 4 math). It is important for other kinds of tests, too, such as symptom rating scales. The items of a depression rating scale, for example, should represent the symptom areas thought to reflect clinical depression. Expert opinion is the basis for establishing content validity, not statistical analysis.

As in other kinds of statistical methods, SEM requires the analysis of scores with good evidence for validity. Because score reliability is generally required for score validity—but does not guarantee it—this requirement includes good score reliability, too (see Little et al., 1999, for exceptions). Otherwise, the accuracy of the interpretation of the results is doubtful. So using SEM does not free researchers from having to think about measurement (just the opposite is true).

## ITEM RESPONSE THEORY AND ITEM CHARACTERISTIC CURVES

For two reasons, it is worthwhile to know about **item response theory** (IRT), also known as **latent trait theory**. First, techniques in IRT permit more sophisticated estimation of item psychometrics than is possible in classical measurement theory. Methods in IRT can be used to equate scores from one test to another, evaluate the extent of item bias over different populations, and construct individualized tests for examinees of different ability levels, or **tailored testing**, among other possibilities. Second, it is an alternative to CFA for analyzing ordinal data. In the past, researchers who analyzed IRT models used specialized software, but now some SEM computer programs such as LISREL and Mplus can analyze at least basic kinds of IRT models. How to analyze ordinal data in CFA is considered later in the book, but part of the logic for doing so is related to that of IRT.

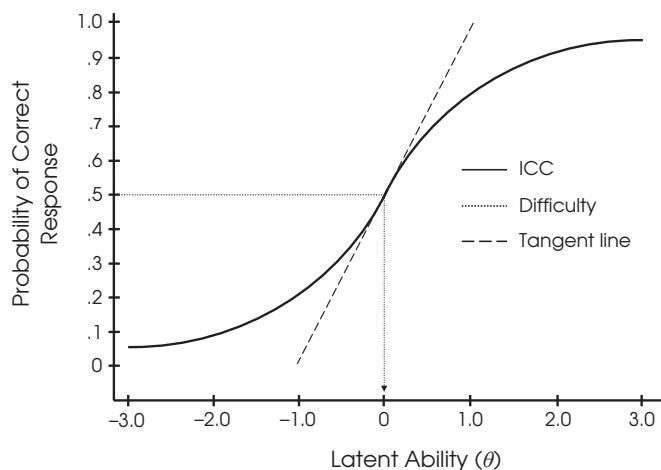
The body of IRT consists of mathematical models that relate responses on individual items to a continuous latent variable  $\theta$ . Assume for this discussion that items are dichotomously scored (0 = incorrect, 1 = correct) and that  $\theta$  is an ability dimension with a normal deviate ( $z$ ) metric. Presented in Figure 4.4 is an **item characteristic curve** (ICC), or a sigmoid function that relates  $\theta$  to the probability of a correct answer. This ICC depicts a **two-parameter IRT model**, where the parameters are item difficulty and item discrimination. Difficulty is the level of ability that corresponds to a 50% chance of getting the item correct, and discrimination is the slope of the tangent line to the ICC at that point. In the figure, difficulty is  $\theta = 0$  (i.e., the mean) because this level of ability predicts that 50% of examinees will pass the item, and discrimination is the slope of the tangent line at this point. The steeper the slope, the more discriminating the item, and the stronger its relation with  $\theta$ . **Three-parameter IRT models** also include a guessing

parameter, and it indicates the probability that an examinee of low ability would correctly guess the answer. A **Rasch model** has a single parameter, item difficulty. Uniform discrimination for all items implies a constant construct, one that can be measured in the same way for all examinees regardless of ability level. In this way, evaluation of Rasch models can be viewed as more confirmatory than fitting more complex IRT models to the data.

Figure 4.4 might look familiar. This is because the shape of an ICC and the sigmoid functions analyzed in logistic regression and probit regression are similar (see Figure 2.4). Shared among all these techniques is the analysis of a continuous latent variable that underlies responses to categorical observed variables. Parameter estimates in IRT can be scaled in either logistic units or probit units, and we will see later in the book that estimates in CFA can be mathematically transformed to estimates of the type generated in IRT. Baylor et al. (2011) gives a clear introduction to IRT.

## SUMMARY

The most widely used methods for continuous outcomes in SEM require screening the data for multivariate normality. It is critical to select appropriate methods for handling missing data. Such methods generally assume that the data loss mechanism is random or at least predictable. Modern options, such as multiple imputation or a special maximum likelihood method for incomplete data files, are generally better choices than classical methods, such as case deletion or single imputation. Computer tools for SEM can analyze either raw data files or matrix summaries. Most estimation methods in SEM assume unstandardized variables, so a covariance matrix is preferred over a correlation



**FIGURE 4.4.** Item characteristic curve (ICC) for the predicted probability of a correct response for a dichotomously scored item in a two-parameter item response theory model. Item difficulty is  $\theta = 0$ , and item discrimination is the slope of the tangent line at  $\theta = 0$ .

matrix without standard deviations when a matrix summary is the input and means are not analyzed. In written reports, researchers should provide information about the psychometrics of their scores. Analysis of scores with poor reliability or validity can jeopardize the results. Computer tools for SEM are described in the next chapter.

## LEARN MORE

The description of modern methods for analyses with missing values by Enders (2010) is exceptionally clear, and Graham and Coffman (2012) discuss specific options in SEM. Malone and Lubansky (2012) consider data screening in SEM with several examples.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277–295). New York: Guilford Press.

Malone, P. S., & Lubansky, J. B. (2012). Preparing data for structural equation modeling: Doing your homework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 263–276). New York: Guilford Press.

## EXERCISES

1. Reproduce the covariance matrix in the upper right part of Table 4.1 from the correlations and standard deviations in the upper left part of the table.
2. Given  $\text{cov}_{XY} = 13.00$ ,  $s_X^2 = 12.00$ , and  $s_Y^2 = 10.00$ , show that the corresponding correlation is out of bounds.
3. Find a normalizing transformation for the data in Figure 4.2.
4. Calculate the covariance matrix for the incomplete data in Table 4.3 using pairwise deletion. Show that the corresponding correlation matrix has an element that is out of bounds.
5. Presented next are scores on five dichotomously-scored items (0 = wrong, 1 = correct) for eight cases (A–H). Calculate the internal consistency reliability  $\alpha_C$  for these data using Equation 4.7. If a reliability procedure is available in your computer program for general statistical analyses, use it to verify your calculations:

A: 1, 1, 0, 1, 1	B: 0, 0, 0, 0, 0
C: 1, 1, 1, 1, 0	D: 1, 1, 1, 0, 1
E: 1, 0, 1, 1, 1	F: 0, 1, 1, 1, 1
G: 1, 1, 1, 1, 1	H: 1, 1, 0, 1, 1

# 5

## Computer Tools

---

Described in this chapter are two categories of SEM computer tools, (1) stand-alone programs that do not require a larger software environment and (2) procedures, packages, or commands that are part of larger computing environments. Some of these computer programs are commercial products, but others are freely available. Options for interacting with SEM computer tools are outlined, and pros and cons of different methods are considered. Computer tools for analyzing directed causal graphs are also described. The relative ease of use of modern computer tools should not lull researchers into thinking that SEM is easy or requires minimal conceptual understanding. This adage attributed to the Canadian communications theorist Herbert Marshall McLuhan is appropriate: We shape our tools and afterwards our tools shape us. I hope that computer use sharpens, rather than dulls, your ability to think critically about SEM.

---

### **EASE OF USE, NOT SUSPENSION OF JUDGMENT**

Computers are necessary for SEM analyses. Just over 40 years ago, LISREL was the only widely available SEM program. At that time, LISREL and related programs were rather difficult to use because they required the generation of rather arcane code for the analysis, and were available only on mainframe computers with stark command-line user interfaces. The abundance of relatively inexpensive, yet capable, personal computers has changed things. Statistical software for personal computers with a graphical user interface is easier to use than their character-based predecessors. User friendliness in modern SEM computer tools—and others for general statistical analyses—is a near-revolution compared with older programs.

Most SEM computer tools still permit the user to write code in that application's native syntax. Some programs offer the alternative to use a graphical editor to specify

the model by drawing it on the screen with geometric symbols such as boxes, circles, and arrows. The program then translates the model graphic into lines of code, which are then used to generate the output. Thus, (1) the user need not know very much (if anything) about how to write syntax in order to run a sophisticated type of multivariate analysis, and (2) the role for technical programming skills is likely to diminish even further. For researchers who understand the fundamental concepts of SEM, this development can only be a bonus—anything that reduces the drudgery and gets one to the results quicker is a plus.

But there are potential drawbacks to push-button modeling. For example, no- or low-effort programming could encourage use of SEM in uninformed or careless ways. This is why it is more important than ever to be familiar with the conceptual and statistical bases of SEM. Computer programs, however easy to use, should be only the tools of your knowledge and not its master. Steiger (2001) notes that emphasis on ease of use of statistical computer tools can give beginners the false impression that SEM is easy. To be sure, some SEM computer tools with drawing editors are advertised with the tagline “SEM made easy!” This message implies that all one has to do is draw the model on the screen and let the computer do the rest.

The reality is that things can and do go wrong in SEM. Beginners often quickly discover that analyses fail because of technical problems, including a terminated program run with cryptic error messages or uninterpretable output (Steiger, 2001). These things happen because actual research problems can be technical, and the availability of user-friendly software does not change this fact. This is why this book places so much emphasis on conceptual knowledge instead of teaching you how to use a particular computer tool: *In order to deal with problems in the analysis when—not if—they occur, you must understand what went wrong and why.*

## HUMAN-COMPUTER INTERACTION

There are three basic ways to interact with SEM computer tools:

1. Batch mode processing where the user writes syntax that specifies the model, data, analysis, and output. Next, the syntax is executed through a “run” command.
2. Through interaction with “wizards” (templates) that automatically build the model and analysis as the user clicks with the mouse cursor on elements in graphical dialog boxes such as text fields, check boxes, or radio buttons. Next, the corresponding syntax is automatically generated by the computer, which is then executed.
3. By drawing the model on screen in a graphical editor. When the diagram is finished, the analysis is run in the graphical user interface.

Batch mode is for users who know the programming language for an SEM computer tool. Syntax files are usually saved as plain-text (ASCII) files, which can later be opened with any basic text editor. Knowledge of program syntax may be unnecessary when using a “wizard” or a drawing editor. Some drawing editors automatically write the corresponding syntax, which can be saved as a text file, but others analyze the model and data without generating a syntax file. Although drawing editors are popular with beginners, there are potential drawbacks—see Topic Box 5.1. The hassles described in the box explain why some researchers switch from a drawing editor when first learning about SEM to working in batch mode as they gain experience.

#### TOPIC BOX 5.1

#### **Graphical Isn't Always Better**

The potential disadvantages of graphical editors in SEM computer tools are outlined next:

1. It can be tedious to draw onscreen a complex model with many variables, such as numerous repeated measures variables in a longitudinal design. This is because the screen quickly fills up with graphical objects. The resulting visual clutter can make it difficult to keep track of what you are doing.
2. Specifying analyses where models are simultaneously fitted to data from two or more samples can be difficult. This is because it may be necessary to look through several different screens or windows in order to find the information about data or model specification for each sample.
3. Standard graphical symbols for model diagrams in SEM do not “translate” well for doing a multilevel analysis. In fact, there is sometimes more than a single way to represent the same multilevel analysis using symbols for model diagrams from SEM. But some computer tools, such as Stata, allow for basic types of multilevel SEM analyses in a drawing editor.
4. It is easier to annotate the analysis by putting comments in a syntax file compared with working in a drawing editor, which may not support user-supplied comments. It is so easy to lose track of what you have done without detailed comments. Thus, using a drawing editor that does not allow annotations can engender carelessness in record keeping (Little, 2013).
5. It seems that it would be easy to produce a publication-quality diagram in a drawing editor, but this is not exactly true. Drawing editors may use a fixed symbol set that does not include special symbols that you want to appear in the diagram. There may be limited options for adjusting the appearance of the diagram (e.g.,

changing line widths). Graphs generated by drawing editors may be rendered in a relatively low resolution that is fine for the computer monitor but not for display in a printed document. To tell the truth, it takes a lot of time to make a publication-quality diagram in *any* graphical computer tool. But once you make a few examples, you can reuse graphical elements, such as those for error terms, in future diagrams.

Here is a trade secret I'll share with you: All model diagrams in this book were created using nothing other than Microsoft Word shapes (rectangles, ovals, arrows, etc.) that are grouped together. Maybe I am a little biased, but I think these diagrams are not too bad. Sometimes you can do a lot with a simple but flexible tool. In this case, you do not need a professional-grade drawing program to create publication-quality model diagrams.

As many researchers become more experienced using SEM computer tools, they tend to stop using a drawing editor to specify their models. For example, they may discover that it can be easier to specify a complex model through a wizard that presents a series of templates. Other researchers eventually learn the syntax of their SEM computer tool and start working in batch mode. There are advantages to this approach. The capability to document the analysis through the insertion of comments in the syntax file was mentioned. Another is that it is often possible to work faster in batch mode than by using a drawing editor. All of the syntax for a complex analysis may fit within a single screen of a text editor. Yes, working in syntax is tedious because every character and line must be correct, but the same is true about drawing editors: Every graphical detail must be correct, or the analysis might fail. So don't fear the prospect of working in batch mode for your SEM analyses. Instead, batch mode is probably in your future, too.

## TIPS FOR SEM PROGRAMMING

Listed next are suggestions for using SEM computer tools; see also Little (2013, pp. 27–29):

1. Readers can download from the book's website syntax, data, and output files for Amos, EQS, LISREL, Mplus, Stata, and the lavaan package in R for each detailed example. Readers can open these files on their own computers without installing any of these programs. Learn from these examples.
2. Annotate your syntax files with comments, which are usually designated in code by a special symbol, such as \*, !, or /. The computer ignores code so designated when the syntax file is executed. Use comments to document the history of the analysis, including the exact form of the model specified, the data, and requested output. Such information is useful for research collaborators who did not conduct the analysis, but who should

later understand the analysis. Comments also help researchers to remember just what they did in a particular analysis days, weeks, months, or even years later. Without sufficient comments one quickly forgets.

3. In analyses where models are progressively simplified (trimmed), the researcher can “comment out” part of the syntax, or deactivate it, by designating those commands as comments in the next analysis. This method preserves the original code as a record of the changes.

4. Beginners sometimes try to analyze models that are overly complicated, and analyses of complex models are more likely to fail. Because the syntax is longer, there are more possibilities for making a mistake. Another reason is that as a model gets bigger, it may be harder to tell whether it is identified. If the researcher does not know that a complex model is actually not identified, then the failure of the analysis may be falsely attributed to a syntax error.

5. Begin instead with a simpler model that you know is identified. Try to get the analysis of that initial model to successfully run. Then build up the model by adding parameters that reflect your hypotheses until the whole target model is eventually specified. If the analysis fails after adding a particular parameter, the reason may be identification.

6. Sometimes iterative estimation fails because the computer needs better **start values**, or initial estimates of the model’s parameters, in order to find a converged solution. (Iterative estimation can also fail due to an ill-scaled data matrix or extreme collinearity.) Some SEM computer tools automatically generate their own start values, but computer-derived values do not always lead to converged solutions. Although the computer’s “guesses” about start values are usually pretty good, sometimes it is necessary for you to provide better ones in order for the solution to converge, especially for more complex models. Suggestions for specifying start values for different types of models are offered later in the book.

7. Analysis of measurement models in CFA may converge better if all indicators of each factor have roughly the same metric. Little’s (2013) method of POMS transformation may be useful if the indicators are Likert-scale items. Variances of continuous indicators in an ill-scaled matrix can be made more similar by multiplying their scores by the appropriate constants.

## SEM COMPUTER TOOLS

Listed in Table 5.1 are the computer programs or procedures described next. They are classified in the table by whether the tool is free, whether it operates as a stand-alone software package, and available modes of interacting with the program. Among computer tools available at no cost are a stand-alone graphical program ( $\Omega$ nyx) and pack-

**TABLE 5.1. Characteristics of Computer Tools for Structural Equation Modeling**

Computer tool	Free	Environment needed	Interaction modes		
			Batch (syntax)	Wizard (template)	Drawing editor
<u>Stand-alone programs</u>					
Amos			✓	✓	✓
EQS			✓	✓	✓
LISREL			✓	✓	✓
Mplus			✓	✓	✓
Ωnyx	✓				✓
<u>Packages, procedures, or commands in larger environments</u>					
sem, lavaan, lava, systemfit	✓	R	✓		
OpenMx	✓	R	✓		
CALIS		SAS/STAT	✓		
Builder, sem, gsem		Stata	✓	✓	✓
SEPATH		STATISTICA	✓	✓	
RAMONA		SYSTAT	✓		

ages for SEM in R, which is itself free. The rest of the computer programs listed in the table are commercial products.

The computer tools in Table 5.1 can analyze all of the structural equation models covered in Parts II and III of this book. Most of these programs can also analyze means or models across multiple samples, and several support **sampling (probability, frequency) weights**. These weights correct for departures in samplings of groups that depart appreciably from populations base rates. That is, such weights correct for potential bias due to imperfect sampling. The descriptions that follow emphasize major features of each program; see the websites listed next for the most current information. These links are also available on this book's website.

## Amos

The IBM SPSS Amos (Analysis of Moment Structures) program (Amos Development Corporation, 1983–2013) is for Windows platform computers.<sup>1</sup> It does not need the SPSS environment to run. It has two main parts, Amos Graphics and a separate Program Editor for working in Amos syntax. Using Amos Graphics does not require knowledge of Amos syntax. The model is specified by drawing it onscreen, and the analysis is controlled in the drawing editor, too. Amos Graphics does not translate

<sup>1</sup>[www.ibm.com/software/products/spss-amos](http://www.ibm.com/software/products/spss-amos)

the diagram into syntax, so there is no plain-text archive of the analysis. A set of templates can automatically draw a latent growth model, among other predefined model elements. Through the Specification Search toolbar, individual paths in the model can be designated as optional. Next, the computer tests models with all possible subsets of the designated paths. Values of fit statistics for all tested models appear in a summary table, and the corresponding diagram can be viewed by clicking with the mouse cursor in the table.

The Amos Program Editor works in batch mode. Its syntax does not use a fixed set of keywords for variable names; instead, such names are specified by the user. The Program Editor is also a language interpreter and debugger for Microsoft Visual Studio VB.NET or C# (“C-sharp”), and Amos syntax is based on object-oriented programming constructs in these languages. Users with programming experience can write VB.NET or C# scripts that modify the functionality of Amos Graphics. Other utilities that are part of Amos include a file manager, a seed manager for recording seed values in simulations of random sampling (e.g., bootstrapping), a data file viewer, and an output file viewer.

Special features of Amos include the capability to generate bootstrapped estimates of standard errors and confidence intervals for all parameter estimates. A version of maximum likelihood estimation for incomplete raw data files is available, and there are other options for analyzing censored or ordinal data. Amos has extensive capabilities for Bayesian estimation, including the generation of graphical posterior distributions, but their correct use requires knowledge of Bayesian statistics. Amos can also analyze mixture models with latent categorical variables either with training data, where some cases are already classified but not the rest, or without training data. Books by Blunch (2013) and Byrne (2010) are resources for Amos users.

## **EQS**

The EQS (Equations) program (Bentler, 2006) is for Windows platform computers.<sup>2</sup> It can be used for all stages of the analysis from data entry and screening to exploratory analyses to SEM. The EQS data editor has many capabilities of a general statistical program, including case selection, variable transformation, and analyses that include regression, ANOVA, and exploratory factor analysis. There are options for analyzing missing data patterns and multiple imputation with the EM algorithm. The user can interact with EQS in three different ways: through batch mode, templates that collect information about the model and data and automatically write syntax, or a drawing editor. The last two ways do not require knowledge of EQS syntax. Available tools in the drawing editor, Diagrammer, can automatically draw onscreen an entire path, factor, or latent growth curve model after the user completes a few templates about variables and effects. It automatically writes EQS syntax into a background window that can be saved as a plain-text file.

---

<sup>2</sup>[www.mvsoft.com](http://www.mvsoft.com)

The syntax of EQS is based on the **Bentler–Weeks representational system**, in which the parameters are regression coefficients for effects on dependent variables and the variances and covariances of independent variables when means are not analyzed. All types of models are thus set up in a consistent way. When specifying a model in syntax, the researcher can use either the original, equations-based EQS syntax where the user must explicitly specify error terms or a paragraph-based syntax where error terms are automatically specified by the computer. Special strengths of EQS include the availability of various estimation methods for non-normal outcome variables, including methods that estimate the degree of skew or kurtosis in the data and then accordingly adjust the estimates. Other features include model-based bootstrapping, the ability to correctly analyze a correlation matrix without standard deviations (raw data are required), and special syntax and estimation methods for multilevel analyses. The current version of EQS is 6.2, but a version 7 is planned.<sup>3</sup> Byrne (2006) gives examples of EQS analyses. Mair, Wu, and Bentler (2010) describe the REQS package, which reads EQS syntax and data files and then imports the results into R after computation in EQS.<sup>4</sup>

## LISREL

The senior SEM computer tool, LISREL (Linear Structural Relations) for Windows, has capabilities that range from data entry and management to exploratory data analysis to SEM.<sup>5</sup> Included with LISREL is PRELIS, which prepares raw data files for analysis. It also has capabilities for diagnosing missing data patterns, pattern matching, bootstrapping, and simulation. As of Version 9.1 (Scientific Software International, 2013), there is closer integration of PRELIS and LISREL. For example, it is no longer necessary to estimate Pearson correlations in PRELIS and then read the estimates into LISREL. These estimations can now be computed directly in LISREL. Multiple imputation can now be performed directly in LISREL, too.

Interactive LISREL consists of a series of templates that prompt the user for information about the model and data and then automatically write command syntax in a separate window. It also allows the user to specify the model by drawing it onscreen through the Path Diagram functionality. When the diagram is run, LISREL automatically writes the corresponding syntax which is then executed. Users already familiar with one of two different LISREL programming languages can, as an alternative, directly enter code into the LISREL syntax editor. If the command “Path Diagram” is placed near the end of the syntax file, LISREL will automatically draw the diagram for the analysis. This allows the user to verify whether the model specified in syntax is actually the one that he or she intended to analyze.

The original LISREL syntax is based on matrix algebra. It is not easy to use until after one has memorized the whole system. An advantage is efficiency: One can often

---

<sup>3</sup>Eric Wu, Multivariate Software (personal communication, January 14, 2015).

<sup>4</sup><http://cran.r-project.org/web/packages/REQS>

<sup>5</sup>[www.ssicentral.com](http://www.ssicentral.com)

specify a complex model in relatively few lines of code. The other LISREL programming language is SIMPLIS (“simple LISREL”), which is not based on matrix algebra, nor does it require familiarity with the classic syntax. Programming in SIMPLIS requires little more than naming the observed and latent variables (but not error terms) and specifying paths with equation-type statements. Certain types of analyses cannot be performed in SIMPLIS, such as those that require imposing nonlinear constraints on parameter estimates. This type of constraint is required by some methods that estimate the curvilinear or interactive effects of latent variables.

The LISREL program has extensive features for analyzing ordinal data, including the option to specify various link functions (e.g., logit, probit, log linear). A special, full-information version of maximum likelihood estimation for ordinal data is also available. Multilevel analyses for up to five levels are supported for models with continuous outcome variables and for up to three levels for models with dichotomous, ordinal, nominal, or count variables. A free student edition of LISREL is available.<sup>6</sup> It makes for a good learning tool because it can analyze many of the models and data sets described in this book. Books by Diamantopoulos and Siguaw (2000) and Vieira (2011) are for LISREL users.

## Mplus

The Mplus program (Muthén & Muthén, 1998–2014) for Windows, Macintosh, and Linux platform computers is divided into a base program and three optional add-on modules for analyzing additional kinds of latent variable models.<sup>7</sup> The Base Program for SEM can analyze models with outcomes that are any combination of continuous, dichotomous, nominal, ordinal, censored, or count variables. It can also analyze discrete- and continuous-time survival models. There are also capabilities for conducting exploratory factor analysis, multiple imputation, and Monte Carlo simulation studies. The Mplus program can also estimate interactive effects of latent variables, generate bootstrapped standard errors or confidence intervals, and perform Bayesian estimation. Special syntax supports the specification of probability weights in complex sampling designs, latent growth models, and CFA models tested over multiple samples (invariance testing). Capabilities for additional kinds of advanced SEM analyses are described in later chapters.

The user interacts with the Mplus Base Program in one of three different ways, in batch mode by writing programs in the Mplus language that specify the model and data; this is done through a language generator (wizard) that prepares files for batch analysis, or through Diagrammer, the Mplus drawing editor. Through the Mplus language generator, the user completes templates about analysis details, such as where the data file is to be found and variable names. The user’s responses are then automatically converted to Mplus syntax that is written to an editor window, but the user must write the syntax

---

<sup>6</sup>[www.ssicentral.com/lisrel/student.html](http://www.ssicentral.com/lisrel/student.html)

<sup>7</sup>[www.statmodel.com](http://www.statmodel.com)

that specifies the model. As the model is drawn onscreen in the Diagrammer, the corresponding Mplus command syntax is automatically written to a syntax editor window. After the whole model is drawn, the user executes (runs) the syntax. Another option is to specify the model in syntax, and Mplus will automatically generate the model diagram with parameter estimates after the user runs the syntax.

Mplus syntax includes keywords for associating indicators with underlying factors, relating predictor variables to outcome variables (observed or latent), and analyzing means. There is also special, compact syntax for specifying latent growth models and interactive or curvilinear effects of latent variables. The Multilevel Add-On to the base program estimates multilevel versions of models for regression analysis, path analysis, factor analysis, SEM, and continuous-time survival analysis. The Mixture Model Add-On analyzes models with categorical latent variables (classes), including regression mixture models, path analysis mixture models, survival mixture models, latent class analysis with multiple categorical latent variables, and finite mixture models, among others. The third optional module is the Combination Add-On, which contains all the features of the multilevel and mixture model add-ons. The *MplusAutomation* package (Hallquist & Wiley, 2015) exports data and results from Mplus to R.<sup>8</sup> Books by Byrne (2012b), Geiser (2013), and Wang and Wang (2012) support Mplus users.

## **Ωnyx**

The **Ωnyx** (pronounced “onix”) program (von Oertzen, Brandmaier, & Tsang, 2015) runs under the Java Runtime Environment (version 1.6 or later) on Windows, Macintosh, or Linux platform computers. It is a graphical software environment for creating and analyzing structural equation models that can be freely downloaded over the Internet.<sup>9</sup> There is no option for specifying the model in native **Ωnyx** syntax. Instead, the researcher draws the model onscreen. After the diagram is complete, the **Ωnyx** program can automatically generate a script that specifies the model in Mplus syntax or in the native syntax of SEM packages for R (*lavaan*, *OpenMx*, or *sem*). These scripts can be saved as text files. The **Ωnyx** program can also read syntax written for *OpenMx* and then automatically draw the model onscreen.

The **Ωnyx** program can also automatically generate the representation of the researcher’s model in the matrix notation of LISREL or the McArdle and McDonald (1984) **reticular action model** (RAM). Structural equation models are represented in RAM notation with three different matrices: **S** (symmetric) for covariances, **A** (asymmetric) for effects of one variable on another, and **F** (filter) for specifying the observed variables. The RAM notational system also includes a set of graphical symbols for model diagrams. This symbolism for model diagrams is used in this book and is introduced in the next chapter.

---

<sup>8</sup><http://cran.r-project.org/web/packages/MplusAutomation>

<sup>9</sup><http://onyx.brandmaier.de>

Another option is to analyze the model in **Ωnyx** using its maximum likelihood method. The program reads raw data files in tab-separated (.dat), comma-separated (.csv), or SPSS (.sav) formats; a summary covariance matrix can also be submitted to **Ωnyx** as the data file. Data are associated with diagrams in **Ωnyx** by dragging variable names from the window for the data file and dropping them in their proper places in the window for the model. After observed variables are linked to the diagram, the estimation of model parameters automatically begins. If a converged solution is found, the results are displayed in their own window. Any changes to the model diagram are reflected by near-simultaneous changes in the corresponding results window.

### **R (sem, lavaan, lava, systemfit, OpenMx)**

The R programming language and environment for statistical computing, data mining, and graphics is part of the GNU Project, a free software collaborative.<sup>10</sup> It runs on Unix, Windows, and Macintosh platform computers and can be freely downloaded.<sup>11</sup> A basic R installation has about the same capabilities as commercial programs for general statistical analyses, but there are now thousands of freely available packages that further extend R's statistical repertoire.<sup>12</sup> The SEM'n'R group supports training in R and SEM.<sup>13</sup>

The SEM packages described next work only in batch mode processing (see Table 5.1). This means that the researcher writes R syntax that specifies the data and then writes the native syntax for a particular SEM package that specifies the model and analysis. Both the R environment in general and SEM packages for R in particular support object-oriented programming. This means that data, models, and analyses results can all be defined as classes with attributes (properties) and methods (functions) for manipulating class content. Researchers with no programming experience whatsoever may find working in R to be austere, but others should be able to adapt without great difficulty.

One of the first SEM packages for R is Fox's (2006) **sem**, which has been updated (Fox 2012).<sup>14</sup> This package can analyze most of the models described in this book, including those with mean structures. Models are specified using McArdle–McDonald RAM notation. The **sem** package has capabilities for calculating robust standard errors and bootstrapping. A version of maximum likelihood estimation for incomplete raw data files is also available. Rosseel's (2012) **lavaan** (*latent variable analysis*) package can analyze models with ordinal or continuous outcomes.<sup>15</sup> A website devoted to sup-

---

<sup>10</sup> [www.gnu.org/gnu/thegnuproject.html](http://www.gnu.org/gnu/thegnuproject.html)

<sup>11</sup> [www.r-project.org](http://www.r-project.org)

<sup>12</sup> <http://cran.r-project.org>

<sup>13</sup> [www.sem-n-r.com](http://www.sem-n-r.com)

<sup>14</sup> <http://cran.r-project.org/web/packages/sem>

<sup>15</sup> <http://cran.r-project.org/web/packages/lavaan>

porting lavaan includes analysis examples and links to related Internet resources.<sup>16</sup> Models are specified in a text- and equations-based language for defining regression models and measurement models. The lavaan package can analyze ordinal data, continuous outcomes with severely non-normal distributions, and incomplete data files. There are also options for bootstrapping. Beaujean (2014) gives examples of latent variable analyses in lavaan.

The lava (linear latent variables) package by Holsta and Budtz-Jørgensena (2012) analyzes structural equation models for both continuous outcomes and censored or binary outcomes.<sup>17</sup> Its syntax is designed to separate definition of the data from specification of the model. This makes it easier to add or remove parts of the model (i.e., respecify it) without changing data definitions. There are also capabilities for analyzing incomplete raw data files with a special version of maximum likelihood estimation, imposing nonlinear constraints in hypothesis testing, fitting models to data from multiple samples, and conducting Monte Carlo-type simulations. The systemfit package (Henningsen & Hamann, 2007) estimates systems of simultaneous linear equations for observed variables.<sup>18</sup> In contrast to standard multiple regression, which analyzes a single equation at a time and assumes that the error variances of multiple criteria are pairwise uncorrelated, equations for multiple criteria are simultaneously analyzed by systemfit while allowing for overlapping error variance. This type of analysis is called **seemingly uncorrelated regressions**, which is actually a misnomer because the regressions are correlated due to overlapping error terms. There are also options for testing whether regression coefficients are equal or not equal.

The OpenMx package (Boker et al., 2011) is a rewrite of an older program known as Mx (Neale, Boker, Xie, & Maes, 2004), which is a matrix processor and a numerical optimizer that can analyze structural equation models. An installation of Mx with a visual editor for specifying a model by drawing it onscreen is called MxGraph, but this program may not install on 64-bit computers or run under modern operating systems. The OpenMx package implements the capabilities of Mx in R using object-oriented programming.<sup>19</sup> It can be used with multicore computers or across large grids of networked computers when analyzing extremely large data sets. The OpenMx package analyzes the full range of structural equation models plus factor mixture models, latent class models, multivariate ordinal models, and genetic epidemiological models, among others. Models are specified in syntax using equations-based notation, matrix algebra notation, or a combination of the two in the same analysis. A special feature is MxModel objects, which can represent an entire model, the data, postestimation hypothesis tests, and analysis output all in a single programming construct. These objects permit greater flexibility in the analysis when comparing alternative models all fitted to the same data.

---

<sup>16</sup><http://lavaan.org>

<sup>17</sup><http://cran.r-project.org/web/packages/lava>

<sup>18</sup><http://cran.r-project.org/web/packages/systemfit>

<sup>19</sup><http://openmx.psyc.virginia.edu>

Multiple models can also be analyzed together in a single application of bootstrapping or Monte Carlo simulation.

Other R packages offer statistical tools for SEM analyses conducted in `sem`, `lavaan` or `OpenMx`, among others. For example, `semTools` (Pornprasertmanit et al., 2014) supports testing for measurement invariance in CFA.<sup>20</sup> It can also estimate the power of certain types of significance tests in SEM. The `simsem` package (Pornprasertmanit, Miller, Schoemann, Quick, & Jorgensen, 2014) is for creating simulated data sets in the SEM framework.<sup>21</sup> The `semPlot` package (Epskamp, 2014) is for generating path diagrams; it can also create model syntax for one program, such as `lavaan`, based on the output of another program, such as `sem`.<sup>22</sup> There are other SEM-related packages for R, and more are being developed all the time. This is an active area of computer program development, and I expect R-based analysis to play an ever larger role in SEM.

## **SAS/STAT (CALIS)**

A brief history of the CALIS (Covariance Analysis of Linear Structural Equations) procedure for SEM in SAS/STAT (SAS Institute, 2014) for Windows and Unix platform computers is needed.<sup>23</sup> This procedure was added to Version 6 of SAS statistical software in 1989. Through the next few versions, CALIS was a syntax-based procedure for analyzing observed or latent-variable models that lacked the capability to directly test models across multiple samples or to estimate models with mean structures. In 2008, the TCALIS module appeared in SAS/STAT Version 9.2 as an experimental procedure with the functionalities just mentioned. The TCALIS procedure also allowed users to specify models using a few different representational systems, including an EQS-type equations syntax, a LISREL-type matrix syntax, or a RAM-type matrix notation, among others. The new capabilities of TCALIS were subsequently incorporated into CALIS for SAS/STAT Version 9.22, and the updated CALIS procedure replaced the experimental TCALIS procedure.

In the most recent version, users still specify their models in CALIS using a syntax model of choice, but the procedure now has the capability to draw the analyzed model onscreen. Graphical features of the diagram are specified in syntax. A model diagram created in CALIS can be further edited in the ODS (output delivery system) Graphics Editor of SAS/STAT. Other noteworthy features of CALIS include the capability to analyze severely non-normal continuous outcomes or incomplete raw data files with a special version of maximum likelihood estimation. The method that the computer will use to generate start values can also be selected. Results from an analysis in CALIS can be automatically saved for a new analysis with either CALIS or a different procedure. O'Rourke and Hatcher (2013) describe SEM analyses in CALIS.

<sup>20</sup><http://semtools.r-forge.r-project.org>

<sup>21</sup><http://simsem.org>

<sup>22</sup><http://sachaepskamp.com/semPlot>

<sup>23</sup>[www.sas.com/en\\_us/software/analytics/stat.html](http://www.sas.com/en_us/software/analytics/stat.html)

### **Stata (Builder, sem, gsem)**

Stata (StataCorp, 1985–2015) is an integrated computing environment for data management, statistical analysis, graphics, programming, and simulation for Windows, Macintosh, and Unix platform computers.<sup>24</sup> It also has extensive capabilities for SEM. There are two main commands for specifying the model in syntax, `sem` and `gsem` (generalized SEM). The `sem` command analyzes models with continuous outcomes that are observed or latent variables, and the `gsem` command analyzes observed outcomes that are continuous, binary, ordinal, multinomial, count, or censored variables. Just the `sem` command can analyze models across multiple samples. Statistical models estimated using the `gsem` command include logistic regression, probit regression, and Poisson regression. It also has capabilities for multilevel modeling in a SEM framework and for analyzing models based on item response theory. Both commands have special syntax for hypothesis testing, residuals diagnostics, output control, calculation of robust standard errors, and bootstrapping.

The Builder does not require knowledge of Stata syntax. It allows the user to specify the model by drawing it onscreen. Stata automatically generates for Builder diagrams the corresponding syntax, which can be annotated and saved as a text file. Special drawing tools in the Builder automatically generate a measurement component (a factor with its indicators) or a regression component (an outcome variable with multiple predictors). Special symbols are used for observed variables in generalized SEM analyses that indicate the underlying distribution (e.g., Gaussian for continuous variables, Poisson for count variables) and the corresponding link function (e.g., logit, probit). There is also a special symbol in the Builder for specifying a basic multilevel analysis. Acock (2013) describes SEM analyses in Stata.

### **STATISTICA (SEPATH)**

J. Steiger's SEPATH (Structural Equation Modeling and Path Analysis) is the SEM procedure in STATISTICA Advanced (StatSoft, 2013), an integrated environment for data visualization, simulation, and statistical analysis in Windows platform computers.<sup>25</sup> Models are specified using the PATH1 programming language that mimics the appearance of diagrams in McArdle–McDonald RAM symbolism. Users who already know the PATH1 language can enter syntax directly into a dialog box. Two other methods do not require PATH1 knowledge. One is a graphical path template in which the user specifies variables and direct effects or covariance by clicking with the mouse cursor in graphical dialogs. The other method is a graphical template for specifying measurement models. Both methods automatically write syntax to a separate window.

Special features of SEPATH include the capabilities to correctly analyze a correlation matrix without standard deviations in summary form (i.e., the raw data are not

---

<sup>24</sup> [www.stata.com](http://www.stata.com)

<sup>25</sup> [www.statsoft.com](http://www.statsoft.com)

required) and generate simulated random samples in Monte Carlo studies. The SEPATH procedure offers options to precisely control parameter estimation, but their effective use requires technical knowledge of nonlinear optimization. In addition, a power analysis procedure in STATISTICA Advanced (also by J. Steiger) estimates the power of various significance tests in SEM. A researcher can also use this procedure to estimate minimum sample sizes needed in order to obtain a target level of power, such as  $\geq .90$ . Power analysis in SEM is described in Chapter 12.

## **SYSTAT (RAMONA)**

M. Browne's RAMONA (Reticular Action Model or Near Approximation) is the SEM procedure in SYSTAT (Systat Software, 2009), a comprehensive program for general statistical analysis in Windows computers.<sup>26</sup> The user interacts with RAMONA in the general SYSTAT environment by submitting a batch file for interactive sessions. An alternative method is to use a wizard with graphical dialogs for naming variables and defining the type of data to be analyzed, but syntax that specifies the model must be typed directly in a text window by the user.

Syntax for RAMONA is straightforward and involves only two parameter matrices, one for direct effects and the other for covariances between independent variables. Special features of RAMONA include the ability to correctly fit a model to a correlation matrix only. There is also a "Restart" command that automatically takes parameter estimates from a prior analysis as start values in a new analysis. This capability is handy when evaluating whether a complex model is actually identified. The RAMONA procedure cannot analyze a model across multiple samples, and there is no direct way to analyze models with a mean structure.

## **OTHER COMPUTER RESOURCES FOR SEM**

The MATLAB program (MathWorks, 2013) is a commercial computing environment and programming language for data analysis, visualization, and simulation.<sup>27</sup> It has many built-in functionalities for linear algebra, curve fitting, and optimization and numerical integration, among others. There are also optional add-ons that support more specialized kinds of analyses, including those for multivariate statistical techniques. Widely known in engineering and the natural sciences, MATLAB is increasingly used by behavioral science researchers, too. There are some MATLAB routines for SEM. For example, MATLAB code for SEM analyses in functional magnetic resonance imaging (fMRI) studies is described by Choi, Song, Chun, Lee, and Song (2013). Goldstein, Bonnet, and Rocher (2007) describe a MATLAB routine for multilevel SEM analyses of comparative data on educational performance across different campuses.

<sup>26</sup> [www.systat.com](http://www.systat.com)

<sup>27</sup> [www.mathworks.com](http://www.mathworks.com)

## COMPUTER TOOLS FOR THE SCM

There are also computer tools for Pearl's structural causal model (SCM). They do not analyze data. Instead, these programs analyze causal models that correspond to directed acyclic graphs, which assume unidirectional causal effects. Textor, Hardt, and Knüppel (2011) describe DAGitty, a freely accessible, Internet browser-based graphical environment for drawing, editing, and analyzing directed acyclic graphs.<sup>28</sup> As the graph is drawn onscreen, DAGitty automatically lists in text windows the testable implications of the graph. It can generate a list of covariates that will minimize bias in estimating causal effects after the researcher, using the mouse cursor, designates an exposure variable and an outcome variable. The program automatically writes syntax that describes the diagram, and the syntax file can be saved. The user can also specify the model in DAGitty's native syntax before it automatically draws the corresponding graph onscreen.

The Belief and Decision Network Tool (Porter et al., 1999–2009) is a freely available Java applet for learning about directed acyclic graphs.<sup>29</sup> After drawing a graph onscreen, this program can then be optionally run in quiz mode, where it poses true–false questions about whether pairs of variables are conditionally independent, controlling for other variables in the graph. The correct answer is shown after the user enters his or her response. In “ask the applet” mode, the user clicks on two focal variables and a set of covariates, and the program automatically indicates whether the focal variables are independent, given those covariates. These capabilities provide tutorials for learning about concepts in graph theory.

The freely available DAG (directed acyclic graph) Program by Knüppel and Stang (2010) is for Windows platform computers.<sup>30</sup> The user specifies the graph by entering syntax that describes the variables and presumed causal effects among them in a series of templates that make up the user interface. Model specifications are summarized in text fields that enumerate covariates as well as exposure, outcome, and unmeasured (latent) confounders, but the DAG Program does not draw the corresponding diagram. Sets of covariates that minimize bias in estimating causal effects are also automatically listed.

The dagR package for R (Breitling, 2010) provides a set of functions for drawing, manipulating, and analyzing directed acyclic graphs and also simulating data consistent with the corresponding diagram.<sup>31</sup> It is intended for epidemiological studies but can be used by researchers in other disciplines, too. The dagR package also allows researchers to evaluate the effects of analyzing different subsets of covariates when analyzing the presumed causal effects of exposure variables on outcome variables and also for finding spurious associations. Models are specified in syntax, but the corresponding graph of

---

<sup>28</sup> [www.dagitty.net](http://www.dagitty.net)

<sup>29</sup> <http://aispace.org/bayes>

<sup>30</sup> <http://epi.dife.de/dag>

<sup>31</sup> <http://cran.r-project.org/web/packages/dagR>

the model can be manipulated in the R environment. A model specified in `dagR` can be saved as an R programming object that can be transferred to another researcher for analysis in R.

## SUMMARY

Most contemporary SEM computer tools are no more difficult to use than other computer programs for general statistical analysis. Ideally, this situation should allow you to be more concerned with the logic and rationale of the analysis than with the mechanics of carrying it out. The capability to specify a structural equation model by drawing it onscreen helps beginners to be productive right away. But with experience you may find that specifying models in syntax and working in batch mode are actually faster and more efficient, and thus easier. Problems can be expected in the analysis of complex models, and no amount of user friendliness in the interface of a computer tool can negate this fact. When (not if) things in the analysis go wrong, you need, first, to have a good conceptual understanding of the nature of the problem and, second, basic computer skills in order to correct the problem. You should also not let ease of computer tool use lead you to carry out unnecessary analyses or select analytical methods that you do not really understand. The concepts and tools considered in Part I of this book set the stage for considering the specification and identification phases of SEM in Part II.

## LEARN MORE

Byrne (2012a) and Narayanan (2012) describe a total of eight different SEM computer tools and also more general considerations in choosing what type of software package to use.

Byrne, B. M. (2012a). Choosing structural equation modeling computer software: Snapshots of LISREL, EQS, Amos, and Mplus. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–324). New York: Guilford Press.

Narayanan, A. (2012). A review of eight software packages for structural equation modeling. *American Statistician*, 66, 129–138.



## **Part II**

# Specification and Identification



# 6

## Specification of Observed Variable (Path) Models

---

The specification of structural models of observed variables—path models—is the topic of this chapter. Outlined first are the basic steps of SEM and graphical symbols used in model diagrams. Some straightforward rules are suggested for counting the number of observations (which is not the sample size) and the number of model parameters. Both of these quantities are needed for checking model identification (see the next chapter). Causal inference in research is also considered. Central to this discussion is the idea that causal inference depends on assumptions regardless of study design, and some assumptions—including directionality specifications—are not verifiable with the data. Thus, clear articulation of assumptions and communication of their rationale are essential in written reports of SEM analyses.

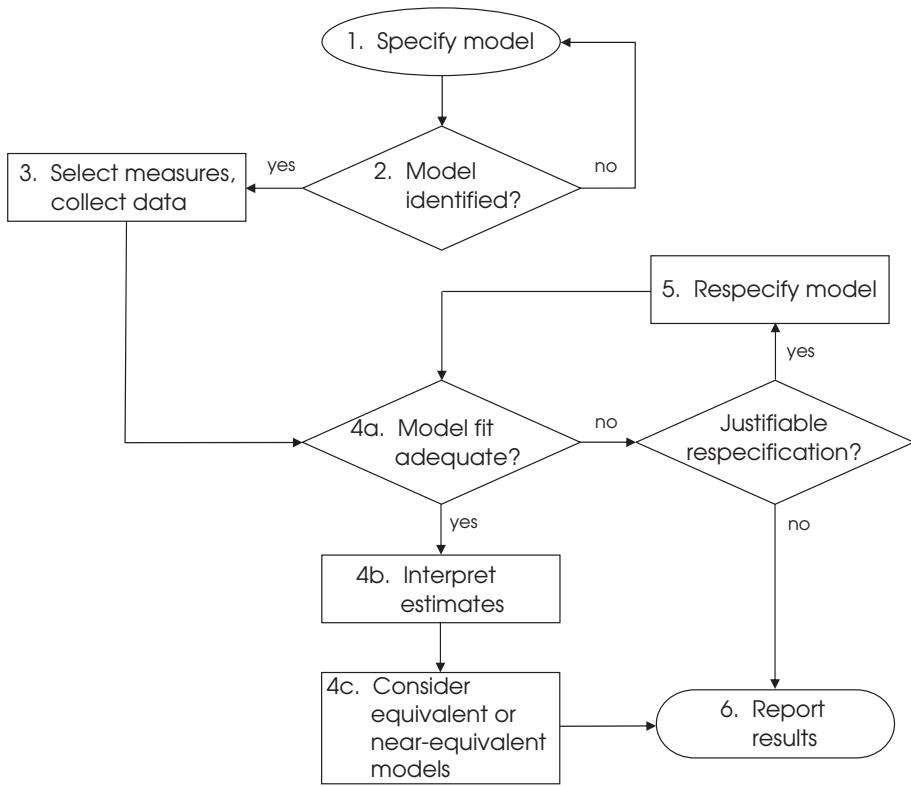
---

### STEPS OF SEM

Six basic steps in most SEM analyses, and two additional optional steps in a perfect world, would be carried out in every analysis. Review of the basic steps will help you to understand the relation of specification to later steps of SEM and to recognize the utmost importance of specification.

#### Basic Steps

The basic steps (see the flowchart in Figure 6.1) are actually iterative because problems at a later step may require a return to an earlier step. Also note in the figure that retaining no model is a possible—and perfectly acceptable—outcome, if there is no justifiable



**FIGURE 6.1.** Flowchart of the basic steps of structural equation modeling. A justifiable respecification has a basis in theory or prior empirical results. Step 5 assumes that the respecified model is identified.

respecification that fits the data. Later chapters elaborate specific issues at each step beyond specification for particular SEM techniques:

1. Specify the model.
2. Evaluate model identification (if not identified, go back to step 1).
3. Select the measures (operationalize the constructs) and collect, prepare, and screen the data.
4. Estimate the model:
  - a. Evaluate model fit; if poor, respecify the model, but only if doing so is justifiable (skip to step 5); otherwise, retain no model (skip to step 6).
  - b. Assuming a model is retained, interpret the parameter estimates.
  - c. Consider equivalent or near-equivalent models (skip to step 6).
5. Respecify the model, which is assumed to be identified (return to step 4).
6. Report the results.

### Specification

Researchers often express their hypotheses with graphical conceptual models, which provide a visual representation of theoretical variables of interest and expected relations among them. Outcome (dependent) variables in SEM are referred to as **endogenous variables**. The word “endogenous” means “from within,” and every endogenous variable has at least one cause, which are usually placed in the left side of the diagram. Some of these causes are independent variables, which in SEM are called **exogenous variables**. The word “exogenous” means “from without (the outside),” and whatever causes exogenous variables are not represented in the model; that is, their causes are unknown as far as the model is concerned.

*Specification is the most important step.* This statement is true because results from later steps assume that the model—the researcher’s hypotheses—are basically correct. I also suggest that researchers make a list of possible changes to the initial model that would be justified according to theory or empirical results; that is, prioritize the hypotheses, and represent just the most critical ones in the initial model. This is because it is often necessary to respecify models (step 5), and respecification should respect the same principles as specification.

### Identification

Although graphical conceptual models are useful heuristics for organizing knowledge and representing hypotheses, they must eventually be translated into a statistical model that can actually be analyzed. A statistical model can be described by a series of equations. These equations define the model parameters, which correspond to presumed relations among variables that the computer eventually analyzes with the sample data.

Statistical models must generally respect certain rules or restrictions. One requirement is that of **identification**. A model is identified if it is *theoretically possible* for the computer to derive a unique estimate of every model parameter. A model is not identified if this property does not hold. The word “theoretically” emphasizes identification as a property of the model and not the data. For example, if a model is not identified, it remains so regardless of the sample size ( $N = 100, 1,000$ , etc.). Therefore, models that are not identified should be respecified (return to step 1); otherwise, attempts to analyze them may be fruitless.

Suppose that a researcher specifies a model that is true to a particular theory, but the resulting model is not identified. In this case, there is little choice in SEM but to respecify the model so that it is identified, but respecifying the original model can be akin to making an intentional specification error from the perspective of the theory. We will see in a later chapter that Pearl’s structural causal model (SCM) better supports the analysis of models where some, but not all, causal parameters are identified. This is because the SCM provides some relatively straightforward ways to graphically determine whether a particular causal effect is identified or not identified. Determining

whether particular causal effects are identified can be harder in SEM. In this way, the SCM is more flexible concerning identification than SEM.

### *Measure Selection and Data Collection*

The various activities for these steps—select good measures, collect the data, and screen them—were discussed in Chapter 4.

### *Estimation*

This step involves using an SEM computer tool to conduct the analysis. Several things take place at this step: (1) Evaluate model fit, which means determine how well the model explains the data. Perhaps more often than not, an initial structural equation model does not fit the data very well. When (not if) this happens to you, skip the rest of this step and consider the question, Can a respecification of the original model be justified, given relevant theory and results of prior empirical studies?

Assuming that the answer to the question is “yes” and given satisfactory fit of the respecified model, then (2) interpret the parameter estimates. Next, (3) consider equivalent or near-equivalent models. Recall that equivalent models explain the same data just as well as the researcher’s preferred model but with a contradictory pattern of causal effects among the same variables. For a given model, there may be many—and in some cases infinitely many—equivalent versions. Thus, the researcher should explain why his or her favorite model should not be rejected in favor of equivalent ones. There may also be near-equivalent models that fit the same data just about as well as the researcher’s preferred model, but not exactly so. Near-equivalent models are often just as critical a validity threat as equivalent models, if not even more so.

### *Respecification*

A researcher usually arrives at this step because the fit of his or her initial model is poor. In the context of model generation, now is the time to refer to that list of theoretically justifiable possible changes I suggested when you specified the initial model. We will deal with respecification in more detail later in the book, but a bottom line of that discussion is that a model’s respecification should be guided more by rational considerations than by purely statistical ones. Any respecified model must be identified; otherwise, you will be “stuck” at this step until you have an estimable model.

### *Reporting the Results*

The final step is to accurately and completely describe the analysis in written reports. The fact that so many published articles that concern SEM are seriously flawed in this regard was discussed earlier. These blatant shortcomings are surprising considering that there are published guidelines for reporting results in SEM (Boomsma, Hoyle, &

Panter, 2012). Best practice recommendations for SEM are outlined in the last chapter of this book.

## Optional Steps

Two optional steps in SEM could be added to the basic ones just described:

7. Replicate the results.
8. Apply the results.

It was mentioned earlier that most of the empirical SEM literature is made up of one-shot studies that are never replicated and that there are few examples of the application of results from SEM analyses. Neglecting to properly carry out the basic steps (1–6) may be part of the problem.

## MODEL DIAGRAM SYMBOLS

Model diagrams are represented in this book by using symbols from the McArdle–McDonald reticular action model (RAM). The RAM symbolism explicitly represents every type of model parameter with its own graphical symbol. This property has pedagogical value for learning about SEM. It also helps you to avoid mistakes when you are translating a diagram to the syntax of a particular computer tool. Part of RAM symbolism is universal in SEM. This includes the representation in diagrams of

1. Observed variables with squares or rectangles.
2. Latent variables (including error terms) with circles or ellipses.
3. Hypothesized directional causal effects, or **direct effects**, on endogenous variables with a line with a single arrowhead (e.g.,  $\rightarrow$ ).
4. Covariances (in the unstandardized solution) or correlations (in the standardized solution) between exogenous variables with a curved line with two arrowheads ( $\curvearrowright$ ).

The symbol described in (4) designates an **unanalyzed association** (covariance) between two exogenous variables. Although such associations are estimated by the computer, they are unanalyzed in the sense that no prediction is put forward about *why* the two exogenous variables covary (e.g., does one cause another?—do they have a common cause?). In RAM symbolism (this next symbol is not universal), two-headed curved arrows that exit and reenter the same variable ( $\bigcirclearrowright$ ) represent the variance of an exogenous variable. Because causes of exogenous variables are not represented in model diagrams, the exogenous variables are considered free to both vary and covary. The symbols  $\bigcirclearrowright$  and  $\curvearrowright$ , respectively, reflect these assumptions. Specifically, the symbol

$\curvearrowleft$  will usually connect every pair of observed exogenous variables in path models, and the symbol  $\Omega$  will connect every observed or latent exogenous variable to itself in RAM notation.

This is not so for endogenous variables in model diagrams. Unlike exogenous variables, the presumed causes of endogenous variables are explicitly represented in the model. Accordingly, endogenous variables are *not* free to vary or covary. This means in model diagrams that the symbol for an unanalyzed association,  $\curvearrowleft$ , does not directly connect any pair of endogenous variables, and the symbol for a variance  $\Omega$  will never originate from and end with any endogenous variable in RAM symbolism. Instead, the model as a whole represents the researcher's account about *why* endogenous variables vary and covary with each other and also with the exogenous variables. During the analysis, this "explanation" is compared with the sample covariances (the data). If the observed covariances and those predicted by the model are similar, the model is said to fit the data; otherwise, the "explanation" is rejected.

Model parameters in RAM symbolism are represented with only three symbols:  $\rightarrow$ ,  $\Omega$ , and  $\curvearrowleft$ . The following rule for defining parameters in words parallels these symbols and is consistent with the Bentler–Weeks representational system for SEM:

---

Parameters of structural equation models when means are not analyzed include (Rule 6.1)

1. direct effects on endogenous variables from other variables; and
  2. variances and covariances of the exogenous variables.
- 

That's it. The simple rule just stated applies to all core structural equation models described in Parts II and III of this book. These models have covariance structures only. An advantage of RAM symbolism is that you can quickly determine the number of model parameters by counting the number of  $\rightarrow$ ,  $\Omega$ , and  $\curvearrowleft$  symbols in the diagram, so what you see is what you get. Outlined in Appendix 6.A is LISREL matrix notation for path models.

## CAUSAL INFERENCE

Causal inference in research depends on design, assumptions, and, to a lesser extent, statistical analysis. This is because analysis by itself is rarely sufficient to establish causation. With a strong research design that supports causal inference, fewer assumptions may be required but, as Pearl (2000) reminds us, "causal assumptions are prerequisite for validating any causal conclusion" (p. 136). This statement echoes that of Wright (1923) from nearly 80 years earlier when he said that "prior knowledge of the causal relations is assumed as prerequisite in the theory of path coefficients" (p. 240). Some causal assumptions can be checked against the data, but others are not empirically test-

able. Whether unverifiable assumptions are tenable is thus a matter of argument, not statistics. Accordingly, the researcher should explicitly articulate such assumptions while reassuring his or her audience that untestable assumptions are reasonable.

Experimental designs in which cases are randomly assigned to conditions are a gold standard for causal inference in the behavioral sciences. Such studies have design elements that bolster internal validity, or the approximate truth about causality discernible in the results.<sup>1</sup> These design elements (E) are summarized next along with corresponding logical requirements for inferring causation from covariation (Mulaik, 2009b, chap. 3):

- E-1.** Random assignment satisfies **temporal precedence**, or the requirement that presumed causes must occur before presumed effects.<sup>2</sup> In this case, the experimental manipulation, or independent variable, happens before the outcome, or dependent variable, is measured.
- E-2.** The control group serves as a counterfactual for the experimental (treatment) group. This counterfactual is imperfect because the same case cannot be simultaneously exposed to treatment and not exposed to treatment. But randomization over replications guarantees that the mean difference between experimental and control groups correctly approximates the true causal effect.
- E-3.** Randomization over replications also ensures that the independent variable is uncorrelated with all other potential causes of the outcome. This property concerns **isolation**, or the absence of other plausible explanations (confounders) that explain the observed covariation between the independent and dependent variables.

Some necessary assumptions (A) for causal inference in experimental designs are summarized next:

- A-1.** The **stable unit treatment value assumption** (Rubin, 2005) has two components: (a) the treatment status of any case does not affect the potential outcomes of any other case, and (b) treatment for all cases is equal; that is, there is no hidden variation in treatment. This assumption is stronger than the usual requirement for independent scores.
- A-2.** The treatment effect is **causally homogeneous**, which means that the causal effect is of the same functional relation for every case (there are no interac-

---

<sup>1</sup>External validity concerns whether causal inferences hold over variations in cases, treatments, settings, or measures.

<sup>2</sup>See Rosenberg (1998) for a discussion of Immanuel Kant's arguments about the possibility of simultaneous causation. The idea of entanglement in quantum mechanics also seems to allow for simultaneous causation over great distances at the quantum level.

tions). It is also assumed that the treatment is correctly administered and that treated cases are all compliant to the same degree.

- A-3. The distributional form and parameters of the corresponding probability distributions (including those for error variances) are correctly specified (e.g., Figure 1.1).
- A-4. Any unmeasured variables that mediate the treatment effect are free from interference; that is, there are no interruptions in mediating pathways between the independent and dependent variables. For measured variables presumed to mediate treatment effects, all such hypotheses are assumed to be correct (Mulaik, 2009b, pp. 95–100).
- A-5. The hypothesis of mediation requires the assumption of **modularity**. This means that the causal process consists of a number of components that are potentially isolatable and thus can be analyzed as separate entities; that is, the causal process is not organic or holistic (Knight & Winship, 2013), and thus inseparable into parts.
- A-6. The outcome is measured at the correct time delay, or lag, among treated cases. This means that treatment effects have had time to appear but also have not yet dissipated (Little, 2013, p. 47). A related assumption is that of **equilibrium**, where the causal effects have settled down after manipulation of the independent variable.
- A-7. There are no obvious threats to internal validity, such as differential effects of maturation, history, or regression to the mean, that confound treatment effects (Shadish, Cook, & Campbell, 2001).
- A-8. All hypothetical constructs are properly operationalized and measured, the scores are reliable, and the appropriate statistical technique is used (respectively, construct validity, conclusion validity).

It should be apparent that even the gold standard—an experimental design—requires many assumptions. Dependence on assumptions in causal inference is even greater in quasi-experimental designs where either (1) cases are not randomly assigned to treatment versus control groups or (2) there is no control group but there is a treatment group. This is because the absence of randomization or counterfactuals in such studies makes it more difficult to reject alternative explanations of the results compared with experimental designs. One of the greatest internal validity threats in such designs is selection bias, which happens when treatment and control groups differ systematically before the treatment is administered. The estimate of the treatment effect can be very biased, if not corrected for initial group differences.

In nonexperimental (passive observational) designs, few, if any, design elements may support causal inference. This is especially true if all variables are concurrently measured, such as when a set of questionnaires is completed during a single test session. Concurrent measurement provides no temporal precedence; thus, study design cannot establish

which of two variables, a presumed cause and a presumed effect, occurred first. Therefore, the sole basis for causal inference in such designs is assumption, one supported by a convincing, substantive rationale for specifying that  $X$  causes  $Y$  instead of the reverse or that  $X$  and  $Y$  mutually cause each other when all variables are measured at once. This process relies heavily on the researcher to rule out alternative explanations of the association between  $X$  and  $Y$  and also to measure other presumed causes of  $Y$ . Both require strong knowledge about the phenomenon under study. If the researcher cannot give a cogent account of directionality specifications, then causal inference in studies with concurrent measurement may be unwarranted. This explains why many researchers are skeptical about inferring causation in nonexperimental designs. An example follows.

Lynam, Moffitt, and Stouthamer-Loeber (1993) hypothesized that poor verbal skills cause delinquency among boys, but both variables just mentioned were simultaneously measured in their sample. This hypothesis raises some questions: Why this particular direction of causation? Is it not also plausible that certain behaviors associated with delinquency, such as truancy or neurological compromise due to drug use, could impair verbal skills? What about other causes of delinquency? In Lynam et al. (1993), the study participants were about 12 years old, which may preclude delinquent careers long enough to affect verbal skills. They cited results of prospective studies which indicated that low verbal ability precedes antisocial behavior. Lynam et al. (1993) measured other presumed causes of delinquency, such as social class, and controlled for them in the analysis. These arguments are not beyond criticism (Block, 1995), but they exemplify the type of rationale needed to justify directionality specifications. Regrettably, too few authors of nonexperimental studies give such detailed explanations.

A researcher has basically three options in SEM if he or she is uncertain about directionality in designs without temporal precedence:

1. Specify a model but without directionality specifications between key variables.
2. Specify and test alternative models, each with different directionality specifications.
3. Include reciprocal causal effects in the model as a way to cover both possibilities.

The specification of unanalyzed associations between exogenous variables is consistent with the first option just mentioned. A problem with the second option is that in SEM different models, such as model 1 with  $Y_1 \rightarrow Y_2$  and model 2 with  $Y_2 \rightarrow Y_1$ , may fit the same data equally well, or nearly so. When this happens, there is no statistical basis for choosing one model over another. The third option concerns causal loops, such as  $Y_1 \leftrightarrow Y_2$ , but the specification of such effects is *not* a simple matter, a point elaborated later in this chapter. So there are potential costs to the specification of causal loops as a hedge against uncertainty about directionality.

Measurement of variables at different times in longitudinal designs provides time precedence: The hypothesis that  $X$  causes  $Y$  would be bolstered if  $X$  were actually mea-

sured before Y. But time precedence is no guarantee. This is because the covariance between X and Y could still be relatively large even if Y causes X and the effect (X) is measured before the cause (Y) (Bollen, 1989, pp. 61–65). This could happen because X would have been affected by Y before either variable was actually measured in a longitudinal study. Even if X actually causes Y, the magnitude of their association may be low if the interval between their measurements is either too short (effects take time to materialize) or too long (temporary effects have dissipated). Longitudinal designs pose other challenges, such as case attrition. This is probably why most SEM studies feature concurrent rather than longitudinal measurement.

Some researchers take the view that causal inference is infeasible unless the design is experimental; that is, there is no causation without manipulation (Holland, 1986). This strong position poses some problems (Bollen & Pearl, 2013). It would preclude causal inference with demographic variables because they are not manipulable. But we know that variables such as age and gender have effects in many different areas, and it makes sense to view some of these effects as causal. Manipulation is not required to infer causation. The moon causes tides, for example, and we know so based on pure observation. Many questions in human studies are not amenable to experimental manipulation, but questions of causality are still relevant (e.g., what is the earnings return for university graduates? Does marriage lower the poverty rate?).

It is possible to correctly infer causation in nonexperimental designs, but the hurdles are certainly much greater. For example, a few decades ago it was an open question whether cigarettes cause lung cancer in humans. It was only with the accumulation of evidence across multiple samples and settings, which corroborated evidence from empirical studies where variables are manipulable (e.g., animal studies about the health effects of nicotine) and with the accurate prediction of the effects of interventions (e.g., bans on smoking in public places) that the modern consensus was reached that, yes, smoking does cause lung cancer in humans.

## SPECIFICATION CONCEPTS

Considered next are key issues in specification.

### What to Include

A basic specification issue revolves around what variables cause a target outcome. Because the literature for newer research areas can be limited, decisions about what to include in the model must sometimes be guided more by the researcher's expertise than by published reports. Consulting with experts in the field about possible specifications may also help. In more established areas, sometimes there is too much information; that is, so many potential causal variables may be mentioned in the literature that it is virtually impossible to include them all. To cope, the researcher must again rely on his or her judgment about the most crucial variables.

It is unrealistic to expect all causal variables to be measured. Given that most structural equation models may be misspecified in this regard, the best way to minimize potential bias is preventive: Make an omitted variable an included one through careful review of extant theory and research. If the sample is archival, then mention should be made of possible specification errors due to the omission of causal variables because they were not measured in the previous study for which the data were collected.

## How to Measure the Hypothetical Construct

Selection of measures is a recurrent research problem. Score reliability is especially important in path analysis, which is characterized by **single-indicator measurement**. This means that there is only one observed measure of each hypothetical construct. Therefore, it is critical that each measure have good psychometrics; otherwise, problems can arise, including over- or underestimation of causal effects and reduction of statistical power that prevents rejection of false models, among other issues described by Cole and Preacher (2014). These problems are magnified as path models become more complex. Disattenuating correlations is one way to control for score reliability (Equation 4.9), but it is not a standard part of path analysis.

Another approach is **multiple-indicator measurement**, where two or more observed variables are used to measure the same construct. Each indicator may reflect a somewhat different facet of the construct, and if multiple indicators are not all based on the same measurement method, there is less concern about common method variance. Having multiple indicators also tends to increase the reliability of factor measurement compared with single-indicator measurement. Path analysis does not directly accommodate multiple-indicator measurement, but the latter is a cardinal characteristic of latent-variable models in SEM, such as CFA models.

## Model Complexity

The potential complexity of any kind of structural equation model is limited by the total number of parameters that can be estimated. This total is limited by the number of **observations**, which is *not* the sample size ( $N$ ). Instead, it is literally the number of entries in the sample covariance matrix in lower diagonal form.<sup>3</sup> This number can be calculated with a simple rule:

---

If  $v$  is the number of observed variables in the model, the number of observations equals  $v(v + 1)/2$  when means are not analyzed. (Rule 6.2)

---

If  $v = 4$  observed variables are represented in a model, then the number of observations is  $4(5)/2$ , or 10. This count (10) equals the total number of variances (4) and unique

---

<sup>3</sup>The term *number of observations* is confusingly used in some SEM computer tools to refer to  $N$ .

covariances (below the main diagonal, or 6) in the data matrix. With  $v = 4$ , the greatest number of parameters that could be estimated by the computer is 10. Fewer parameters can be estimated in a simpler model, but not more than 10. The number of observations has nothing to do with sample size. If four variables are measured for 100 or 1,000 cases, the number of observations is still 10. Adding cases does not increase the number of observations; only adding *variables* can do so.

The difference between the number of observations and the number of parameters is the **model degrees of freedom**, or

$$df_M = p - q \quad (6.1)$$

where  $p$  is the number of observations (Rule 6.2) and  $q$  is the number of estimated parameters (Rule 6.1). The condition that there be at least as many observations as parameters can be expressed as the requirement that  $df_M \geq 0$ .

A model with more estimated parameters than observations ( $df_M < 0$ ) is not amenable to empirical analysis in SEM because such a model is not identified. If you tried to estimate a model with negative degrees of freedom, an SEM computer tool would likely terminate its run with error messages. Most models with zero degrees of freedom ( $df_M = 0$ ) perfectly fit the data and thus test no particular hypothesis.<sup>4</sup> Models with positive degrees of freedom generally do not have perfect fit. This is because  $df_M > 0$  allows for the possibility of model–data discrepancies. Raykov and Marcoulides (2000) describe each degree of freedom as a dimension along which the model can potentially be rejected. Thus, retained models with greater degrees of freedom have withstood a greater potential for rejection. This idea underlies the **parsimony principle**: Given two models with similar fit to the data, the simpler model is preferred, assuming that the simpler model is theoretically plausible.

## Parameter Status

Each model parameter can be free, fixed, or constrained depending on its specification. A **free parameter** is to be estimated by the computer with the data. In contrast, a **fixed parameter** is specified to equal a constant. The computer “accepts” this constant as the estimate regardless of the data. For example, the hypothesis that  $X$  has no direct effect on  $Y$  corresponds to the specification that the coefficient for the path  $X \rightarrow Y$  is fixed to zero. It is common in SEM to test hypotheses by specifying that a previously fixed-to-zero parameter becomes a free parameter, or vice versa. Results of such analyses may indicate whether to respecify a model by making it more complex (an effect is added—a fixed parameter becomes a free parameter) or more parsimonious (an effect is dropped—a free parameter becomes a fixed parameter).

---

<sup>4</sup>Pearl (2012) noted that estimates of particular effects in models where  $df_M = 0$  can be of substantive value. For example, it may be worthwhile to know that the magnitude of the direct effect of  $X_1$  on  $Y$  is three times that of the direct effect of  $X_2$  on the same endogenous variable.

A **constrained parameter** is estimated by the computer within some restriction, but it is not fixed to equal a constant. The restriction typically concerns the *relative* values of other constrained parameters. An **equality constraint** means that the estimates of two or more parameters are forced to be equal. Suppose that an equality constraint is imposed on two direct effects,  $X_1 \rightarrow Y$  and  $X_2 \rightarrow Y$ . This constraint simplifies the analysis because only one coefficient is needed rather than two. In a multiple-samples SEM analysis, a **cross-group equality constraint** forces the computer to derive equal estimates of the same parameter across all groups. This specification corresponds to the null hypothesis that the parameter is equal in all populations from which the samples are drawn.

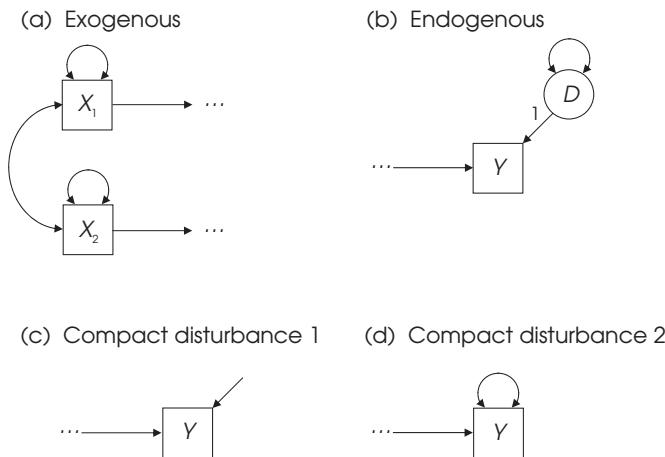
A **proportionality constraint** forces one parameter estimate to be some fraction of the other. For instance, the coefficient for one direct effect in a reciprocal relation may be forced to be three times the value of the other coefficient. An **inequality constraint** forces an estimate to be either less than or greater than the value of a specified constant. The requirement that the value of an unstandardized coefficient must be over 5.0 is an example of an inequality constraint. The imposition of proportionality or inequality constraints generally requires knowledge about the relative magnitudes of effects, but such knowledge is rare. A **nonlinear constraint** imposes a nonlinear relation between two parameter estimates. For example, the value of one estimate may be forced to equal the square of another. Some methods for estimating curvilinear or interactive effects of latent variables described in Chapter 17 rely on nonlinear constraints.

## PATH ANALYSIS MODELS

Path analysis is the oldest member of the SEM family, but it is not obsolete. About 25% of articles reviewed by MacCallum and Austin (2000) concern path models, so path analysis is still widely used. There are also times when there is just a single observed measure of each construct, and path analysis is a single-indicator technique. Finally, *if you master the principles of path analysis, you will be better able to understand and critique a wider variety of structural equation models*. So read this section carefully even if you are more interested in latent variable models.

### Building Blocks

Presented in Figure 6.2(a) is RAM symbolism for two observed continuous variables,  $X_1$  and  $X_2$ , specified as exogenous. Each variable is designated as free to vary, and they are assumed to covary, too. Whatever causes  $X_1$  or  $X_2$  is outside of the model, a property called **exogeneity**. Exogenous variables are presumed to cause endogenous variables, which are not shown in the figure. Exercise 1 asks you how a nominal exogenous variable with  $k \geq 2$  categories, such as membership in one of three different groups ( $k = 3$ ), would be represented in a model. Hint: Think of dummy, effect, or contrast codes in multiple regression.



**FIGURE 6.2.** McArdle–McDonald reticular action model (RAM) symbolism for (a) exogenous variables and (b) an endogenous variable in a path model. Compact, non-RAM symbolism for disturbances depicting (c) only the direct effect of omitted causes or (d) only error (unexplained) variance.

Presented in Figure 6.2(b) is a continuous endogenous variable  $Y$ . It is assumed to be caused by at least one other measured variable, exogenous or endogenous (not shown). Every endogenous variable has a **disturbance**, which is designated as  $D$  in the figure. A disturbance represents residual (unexplained) variation. Disturbances signal the assumption of probabilistic causality. They are also considered as latent variables, specifically, as unmeasured exogenous variables. This is because a disturbance represents all omitted causes of the endogenous variable plus measurement error. This is why disturbances are represented in RAM symbolism with circles and the symbol for the variance of an exogenous variable. Error variances must be estimated by the computer, so they count as a free model parameter (Rule 6.1).

The path  $D \rightarrow Y$  in Figure 6.2(b) represents the direct effect of all unmeasured causes. The numeral value (1) that appears in the figure next to the path is a **scaling constant** that assigns a metric to the disturbance. This is necessary because error variance is latent, and latent variables need scales before the computer can estimate anything about them. The scaling constant is also called an **unstandardized residual path coefficient**. It forces the computer to estimate the disturbance variance as some fraction of the total observed variance of the corresponding endogenous variable; that is,  $s_D^2 \leq s_Y^2$ .

Two different (non-RAM), abbreviated ways to represent disturbances in diagrams are shown in the bottom part of Figure 6.2. Both compact versions omit the scaling constant and the symbol for the disturbance as a latent variable. The representation in Figure 6.2(c) also omits the symbol for the variance. It shows only the direct effect of omitted causes on  $Y$ . The compact representation in Figure 6.2(d) includes only the symbol for the error variance attached directly to the symbol for  $Y$ . This representation does *not* imply that endogenous variable  $Y$  is free to vary. Either compact representation

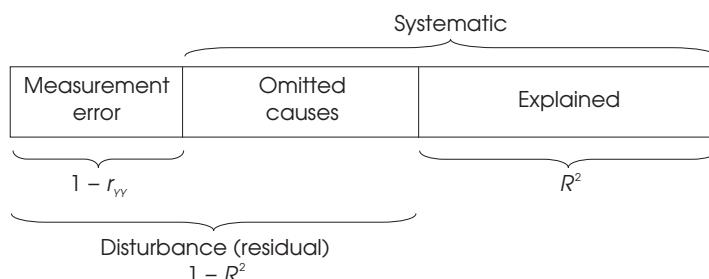
for disturbances just described is acceptable—and they save space in the diagram—but remember that disturbance variances are free parameters.

Do not think that disturbances in path models are equivalent to residuals in multiple regression. To do so would be to confuse a causal model (path analysis) with a statistical model (regression analysis). Regression residuals are artifacts of least squares estimation such that those residuals, by definition, are uncorrelated with the predictors. But disturbances are not analysis artifacts; instead, they are determined by physical reality, including social or genetic factors that affect the corresponding endogenous variable but are unmeasured (Pearl, 2012).

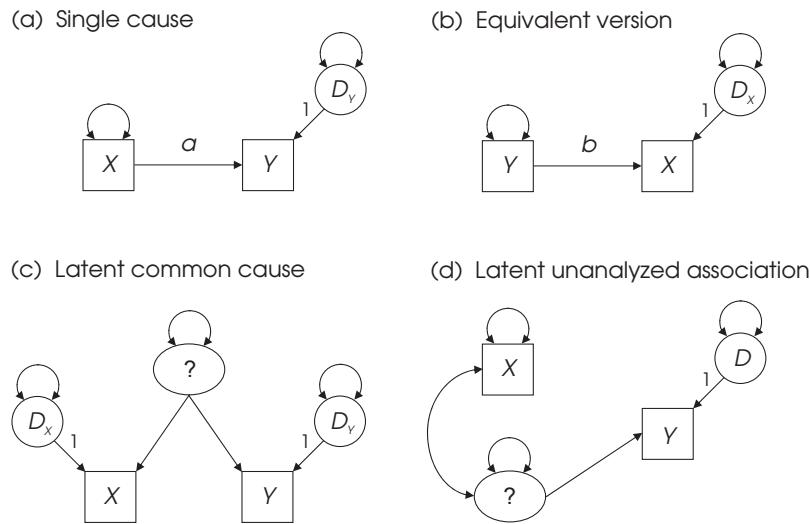
A conceptual partition of the standardized variance for a continuous endogenous variable is presented in Figure 6.3. The proportion of explained variance is  $R^2$ , the squared multiple correlation between Y and all the observed variables specified to cause Y. The proportion of unexplained variance is  $1 - R^2$ , which can be decomposed into measurement error and systematic variance due to omitted causes. The measurement error is estimated by  $1 - r_{YY}$ , where  $r_{YY}$  is a reliability coefficient. Exercise 2 asks you to state the expected consequences of decreasing score reliability for the endogenous variable depicted in the figure.

### Elemental Models and Assumptions

Next we consider elemental path models from which all of the more complex models can be constructed. The path model in Figure 6.4(a) assumes that (1) X is a cause of Y, and (2) all unmeasured causes of Y are uncorrelated with X; that is, D and X are independent. The hypothesis that X causes Y is a **weak causal assumption** that excludes all but one value for the direct effect  $X \rightarrow Y$ . The excluded value is zero (i.e., no causal effect), but without further restriction, the estimate of this effect could theoretically assume an infinite number of values  $\neq 0$ . A **strong causal assumption** reflects the prediction that the direct effect is exactly zero, which is represented in a model diagram by the absence of the symbol for a direct effect between two variables.



**FIGURE 6.3.** Partition of standardized variance for a continuous endogenous variable Y in a path model.  $r_{YY}$ , score reliability coefficient;  $R^2$ , proportion of variance explained by all measured variables with direct effects on the corresponding endogenous variable.



**FIGURE 6.4.** Model of a single direct effect (a) and an equivalent version (b). Noncausal models that predict an association between  $X$  and  $Y$  due to an omitted common cause (c) or to an unanalyzed association between  $X$  and an omitted cause of  $Y$  (d).

The assumption that  $X$  and  $D$  are independent is necessary for Figure 6.4(a) because the **path coefficient**—the statistical estimate of  $X \rightarrow Y$  that is represented by  $a$  in the figure that could be in either unstandardized or standardized form—is calculated holding constant all omitted causes (pseudo-isolation). Path models in SEM are parametric, so this direct effect is assumed to be linear, if both  $X$  and  $Y$  are continuous. Because exogenous variables have no error terms (see the figure), scores on  $X$  are assumed to be perfectly reliable ( $r_{XX} = 1.00$ ). Measurement error in  $Y$  would be manifested in its disturbance, so it is *not* assumed that  $r_{YY} = 1.00$ .

The situation in which a putative exogenous variable  $X$  actually covaries with  $D$  is **endogeneity**. This means that exogeneity does not hold and  $X$  is not really exogenous. Misspecification of directionality can lead to endogeneity. For example, if  $Y$  causes  $X$  instead of the reverse, then  $X$  cannot be exogenous. This situation ( $Y \rightarrow X$ ) is illustrated in Figure 6.4(b), which is an equivalent version of Figure 6.4(a) where  $X \rightarrow Y$  is specified. Although the two models make contradictory assumptions about directionality, they would have exactly the same fit to the data. Exercise 3 asks you to prove this assertion by comparing path coefficient  $a$  in Figure 6.4(a) with path coefficient  $b$  in Figure 6.4(b), assuming that both  $X$  and  $Y$  are continuous, their relation is linear, and the coefficient is estimated in bivariate regression. Endogeneity can also occur if  $X$  and  $Y$  reciprocally cause each other, or  $X \leftrightarrow Y$ . Antonakis, Bendahan, Jacquart, and Lalivé (2010) give an example: Suppose that more police are hired in order to reduce crime. But if an increase in crime leads to the decision to hire more police, the hiring of more police is not exogenous because the two variables reciprocally affect each other.

Variables  $X$  and  $Y$  can also covary for reasons that have nothing to do with causation. In Figure 6.4(c), the association between  $X$  and  $Y$  is spurious due to a common but unmeasured cause. Consequently, both variables are endogenous. In Figure 6.4(d), variables  $X$  and  $Y$  are expected to covary due to the relation between  $X$  and an unmeasured cause of  $Y$ , but  $X$  itself has no causal effect on  $Y$ . Looking at all four models in Figure 6.4, we find that the whole “trick” in path analysis (and SEM) is deciding which model (among others not shown in the figure) best explains the observed covariance. *But this decision is a matter of specification, not analysis.* This is because different sets of structural equations can be equally consistent with the same data (e.g., Figures 6.4(a) and 6.4(b)). In this sense, directional specifications are not generally “tested” in SEM; instead, they are assumed, and the overall fit of the model to the data is evaluated under such assumptions. But there is usually no direct “test” of whether a particular directionality specification is correct or incorrect. This fact highlights the importance of untestable assumptions in causal modeling.

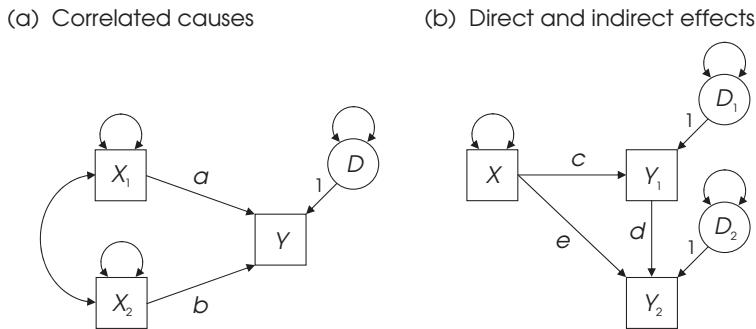
Figure 6.5(a) represents correlated causes ( $X_1, X_2$ ) of the same outcome,  $Y$ . Path coefficients for the direct effects ( $a, b$ ) are each estimated controlling for the covariation  $X_1$  and  $X_2$ , just as in regression analysis. This model assumes that (1) there is no measurement error in both  $X_1$  and  $X_2$ ; and (2) all unmeasured causes of  $Y$  are independent of both  $X_1$  and  $X_2$ . Also, (3) there is no moderation between  $X_1$  and  $X_2$ . This means that the causal effect of  $X_1$  on  $Y$  does not depend on  $X_2$ , and vice versa. The same assumption also says that coefficients  $a$  and  $b$  in the figure each remains constant over the levels of the other cause. The term **moderation** describes effects that are **causally heterogeneous**, or conditional. An example would be if the causal effect of  $X_1$  on  $Y$  were positive for cases with relatively low scores on  $X_2$  but the same causal effect was negative for cases with relatively high scores on  $X_2$ . Because moderation is always symmetrical, it would also be true that the causal effect of  $X_2$  on  $Y$  depends on the level of  $X_1$ . In this example, moderation is also a joint effect that requires the measurement of both  $X_1$  and  $X_2$  in order to detect their interaction.<sup>5</sup> It is possible to estimate moderation in path analysis, but models of conditional causation must be specified in a very particular way (i.e., not Figure 6.5(a)). We will return to this topic in Chapter 17.

In Figure 6.5(b) there are two direct effects on  $Y_2$  from other measured variables, the exogenous variable  $X$  and the endogenous variable  $Y_1$ . The former ( $X_1$ ) is also specified as cause of the latter  $Y_1$ . These specifications give  $Y_1$  a dual role as both a cause (of  $Y_2$ ) and an outcome (of  $X$ ). The pathway

$$X \rightarrow Y_1 \rightarrow Y_2$$

represents the **indirect effect** of  $X$  on  $Y_2$  through the intervening variable  $Y_1$ . In words,  $X$  causes some change in  $Y_1$ , which in turn leads to change in  $Y_2$ . The model in Figure 6.5(b) also assumes that (1)  $r_{XX} = 1.00$ ; (2) there is no interaction between the causes

<sup>5</sup>Some authors use the term *interaction* to describe joint effects that are not necessarily causal and reserve use of the term *moderation* for conditional causal effects, but this practice is not universal.



**FIGURE 6.5.** Models with (a) correlated causes and (b) direct effects and an indirect effect. The latter is estimated as the product  $cd$  for continuous variables.

of  $Y_2$ ,  $X$  and  $Y_1$ ; (3) omitted causes of both  $Y_1$  and  $Y_2$  are unrelated to  $X$ ; and (4) omitted causes of  $Y_1$  and omitted causes of  $Y_2$  are unrelated.

Assuming continuous variables in a linear model and no interaction, path coefficients  $c–e$  in Figure 6.5(b) estimate, respectively, the direct effects

$$X \rightarrow Y_1 \quad Y_1 \rightarrow Y_2 \quad \text{and} \quad X \rightarrow Y_2$$

Coefficients  $d$  and  $e$  control for the effects of, respectively,  $X$  and  $Y_1$  on  $Y_2$ . The term  $cd$ , or the product of the coefficients for the paths  $X \rightarrow Y_1$  and  $Y_1 \rightarrow Y_2$ , estimates the indirect effect of  $X$  on  $Y_2$  controlling for the direct effect of  $X$  on  $Y_2$ . The idea behind **product estimators**, such as  $cd$ , of indirect effects among continuous variables is elaborated in later chapters, but for now we can say that indirect effects are routinely estimated in path analysis (and in other SEM techniques, too). The **total effect** of  $X$  on  $Y_2$  is defined as  $e + cd$ , which is the sum of the direct and indirect effects of  $X$  in Figure 6.5(b).

### Indirect Effects versus Mediation

Indirect effects are always part of mediation, but they are not synonymous. This is because **mediation** refers to the *causal hypothesis* that one variable causes *changes* in another variable, which in turn leads to *changes* in the outcome variable (Little, 2013). The intervening variable in this definition is the **mediator**, and it transmits part of the effect of a causally prior variable to a third variable affected by the mediator. Thus, mediation refers to a causal pathway through which effects are conducted from a source through a conduit—the mediator—to the final outcome.

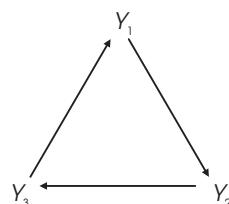
The emphasis on *changes* in the above definition of mediation is because without evidence for change, the only effects that could be supported are indirect effects, not mediation. In other words, mediation always involves indirect effects, but not all indirect effects automatically signal mediation. This is especially true in nonexperimental designs with no time precedence. Such designs are also called cross-sectional designs. Look back at

Figure 6.5(b). If variables  $X$ ,  $Y_1$ , and  $Y_2$  are simultaneously measured, then it is impossible to estimate changes among these components of the indirect pathway from  $X$  to  $Y_2$ . In their arguments for designs with time precedence, Maxwell and Cole (2007) note that conditions under which mediation would be correctly analyzed with cross-sectional data are very rare and that parameter estimates are almost always biased. A minimal design that corresponds to Figure 6.5(b) is one where  $X$  is an experimental variable with two conditions (e.g., drug vs. placebo),  $Y_1$  is a presumed mediator (e.g., compliance) measured on a second occasion, and  $Y_2$  is the outcome (e.g., health status) measured on a third occasion. Longitudinal designs for estimating mediation are described in a later section of this chapter, and experimental mediation designs are described in Chapter 8. In general, use of the term *mediation* should be reserved for designs that feature time precedence; otherwise, use of the term *indirect effect* is more realistic.

## RECURSIVE AND NONRECURSIVE MODELS

There are two kinds of path models. **Recursive models** are the most straightforward and have two basic features: their disturbances are uncorrelated, and all causal effects are strictly unidirectional. The models in Figures 6.4–6.5 are all recursive. **Nonrecursive models** have causal (feedback) loops or may have correlated disturbances. The model in Figure 6.6(a) is nonrecursive because it has a **direct feedback loop** in which  $Y_1$  and  $Y_2$  are specified as direct causes of each other ( $Y_1 \leftrightarrow Y_2$ ). Each of these variables is measured only once and also simultaneously; that is, feedback is estimated with data from a cross-sectional design, not a longitudinal design.

**Indirect feedback loops** involve at least three variables connected by direct effects that eventually point back to earlier variables. In a path diagram, an indirect feedback loop with  $Y_1$ ,  $Y_2$ , and  $Y_3$  would be represented as a “triangle” with paths that connect them in the order specified by the researcher. Presented without disturbances or other variables in the model, an example of an indirect feedback loop with three variables is shown next:



Because each variable in the feedback loop just illustrated is involved in an indirect effect, such as  $Y_2$  in the indirect pathway

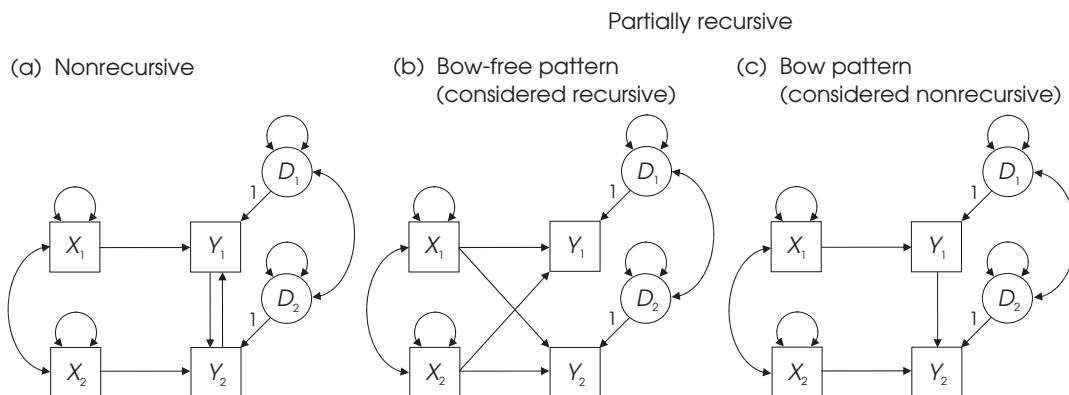
$$Y_1 \rightarrow Y_2 \rightarrow Y_3$$

feedback is thus indirect.

The model of Figure 6.6(a) also has a **disturbance covariance** (for unstandardized variables) or a **disturbance correlation** (for standardized variables). The term *disturbance correlation* is used from this point on regardless of whether or not the variables are standardized. A disturbance correlation, such as  $D_1 \curvearrowleft D_2$ , reflects the assumption that the corresponding endogenous variables ( $Y_1, Y_2$ ) share at least one unmeasured cause. Unlike unanalyzed associations between measured exogenous variables, such as  $X_1 \curvearrowleft X_2$ , the specification of correlated disturbances in the model is not routine. Why this is true is explained momentarily.

There is another type of path model, one that has unidirectional effects and correlated disturbances; two examples of this type are presented in Figures 6.6(b) and 6.6(c). The classification of such models is not consistent. Some authors call these models nonrecursive, whereas others use the term **partially recursive**. But more important than the label for these models is the distinction made in the figure: Partially recursive models with a **bow-free pattern** of disturbance correlations can be treated in the analysis just like recursive models. A bow-free pattern means that correlated disturbances are restricted to pairs of endogenous variables *without* direct effects between them—see Figure 6.6(b). In contrast, partially recursive models with a **bow pattern** of disturbance correlations must be treated in the analysis as nonrecursive models. A bow pattern means that a disturbance correlation occurs *with* a direct effect between that pair of endogenous variables—see Figure 6.6(c) (Brito & Pearl, 2003). All ensuing references to nonrecursive and recursive path models include, respectively, partially recursive models with and without a bow pattern of disturbances.

Before we continue, let's apply the rules for counting observations, parameters, and degrees of freedom to the nonrecursive model in Figure 6.6(a). Because there are  $v = 4$  observed variables in this model, the number of observations is  $4(5)/2 = 10$ . There are a total of four exogenous variables, two measured ( $X_1, X_2$ ) and two unmeasured ( $D_1, D_2$ ). The variances (4) and covariances (2) (i.e.,  $X_1 \curvearrowleft X_2, D_1 \curvearrowleft D_2$ ) of these exogenous vari-



**FIGURE 6.6.** Examples of nonrecursive and partially recursive path models.

ables are free parameters. There are a total of four direct effects on endogenous variables from other measured variables, including

$$X_1 \rightarrow Y_1 \quad X_2 \rightarrow Y_2 \quad Y_1 \rightarrow Y_2 \quad \text{and} \quad Y_2 \rightarrow Y_1$$

Because the total number of observations and free parameters for this model are equal (10), the model degrees of freedom are zero ( $df_M = 0$ ). Exercise 4 asks you to count the number of free parameters in Figures 6.6(b) and 6.6(c), and Exercise 5 asks you to describe in words the hypotheses about the relations between  $Y_1$  and  $Y_2$  that are represented in Figure 6.6(c).

### **Implications of the Distinction between Recursive and Nonrecursive Models**

The assumptions of recursive models that all causal effects are unidirectional and that the disturbances are independent simplify the statistical demands for the analysis. For example, multiple regression can be used to estimate path coefficients and disturbance variances in recursive path models. The occurrence of a technical problem in the analysis is less likely for recursive models. It is also true that recursive structural models are identified, given that necessary requirements for identification are satisfied (Chapter 7). The same assumptions of recursive models that ease the analytical burden are also restrictive. For example, neither causal loops nor correlated disturbances in a bow pattern can be represented in a recursive model.

The kinds of effects just mentioned can be represented in nonrecursive models, but such models require special assumptions. Estimation of causal loops with cross-sectional data assumes equilibrium, or that changes in the system underlying reciprocal causation have already manifested their effects and that the system is in a steady state. Kaplan, Harik, and Hotchkiss (2001) remind us that there is generally no statistical way to directly evaluate the equilibrium assumption; that is, it must be substantively argued. But this assumption is rarely acknowledged in studies where feedback effects are estimated with cross-sectional data. This is unfortunate because the results of computer simulation studies by Kaplan et al. (2001) indicate that violation of the equilibrium assumption can lead to severely biased estimates. Another assumption is that of **stationarity**, the requirement that the basic causal structure does not change over time.

Controversy has arisen over estimating reciprocal causation in designs with concurrent measurement (Wong & Law, 1999). The main reason for the controversy is the absence of temporal precedence in such designs. This implies that the two-way paths that make up a direct feedback loop, such as  $Y_1 \leftrightarrow Y_2$  in Figure 6.6(a), represent an instantaneous cycling process, but in reality there may be no such causal mechanism (Hunter & Gerbing, 1982). In this sense, the measurement occasions in designs without temporal precedence are always wrong.

But the absence of temporal precedence may not always be a liability when estimating reciprocal causation. Finkel (1995) argued that the lag for some causal effects

is so short that it would be impractical to measure them over time. Examples are the reciprocal effects of the moods of spouses on each other. Although the causal lags in this example are not zero, they may be so short as to be virtually synchronous. If so, the assumption of instantaneous cycling for feedback loops in nonrecursive models would be more defensible. In this case, it may be more appropriate to estimate reciprocal effects with very short lags in cross-sectional designs even when panel data are available (Wong & Law, 1999). This is because the true length of causal lags is not always known. In this case, longitudinal data collected according to some particular temporal measurement schedule are not automatically superior to cross-sectional data.

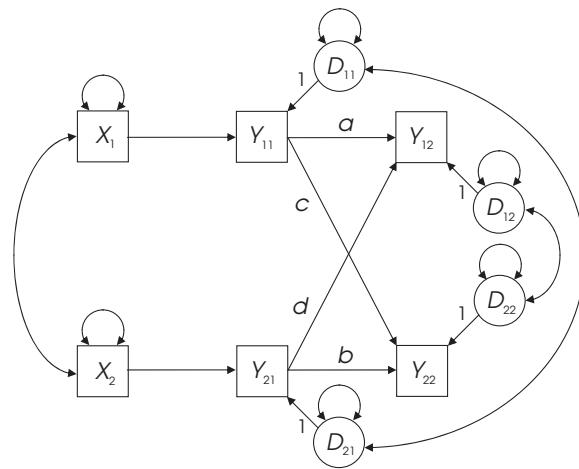
Disturbances of variables that make up a feedback loop are often assumed to covary. This specification makes sense because if variables mutually cause each other, then it is plausible that they may share unmeasured causes. Some of the error in predicting one variable in a causal loop, such as  $Y_1$  in Figure 6.6(a), may be due to the other variable in that loop, or  $Y_2$  in the figure, and vice versa. Studies with repeated measures are another context for specifying correlated disturbances. This is because the error variances of such variables may overlap, which is described as **autocorrelated errors**. The degree of shared error variance may be greater if the test-retest interval is relatively short (e.g., back-to-back learning trials). The capability to test hypotheses about **error covariance structures** is a relative strength of SEM.

The addition of each disturbance correlation to the model “costs” one degree of freedom and thus makes the model more complex (and also improves fit). If there are substantive reasons for specifying disturbance correlations, then it is probably better to analyze the model with these terms than without them. The constraint that a disturbance correlation is zero when there are common unmeasured causes tends to redistribute this association toward the exogenous end of the model, which can result in biased estimates of direct effects. In general, disturbances should be specified as correlated if there are good reasons for doing so; otherwise, one should be wary of making the model overly complex by adding parameters without a good reason.

Another complication of nonrecursive models is identification. There are some straightforward ways to determine whether some, but not all, types of nonrecursive path models are identified. These procedures are described in the next chapter, but it is worthwhile to make this point now: Adding exogenous variables is one way to remedy an identification problem of a nonrecursive model. But this typically can be done *before* the data are collected. *Thus it is critical to evaluate whether a nonrecursive path model is identified right after it is specified and before the study is conducted.*

## PATH MODELS FOR LONGITUDINAL DATA

There are many kinds of path models for repeated measures data, so just a few major types are described next; see Little (2013) and Newsom (2015) for more information. Presented in Figure 6.7 is a **panel model** for endogenous variables  $Y_1$  and  $Y_2$ , each measured at two different occasions. This model also includes exogenous variables,  $X_1$  and



**FIGURE 6.7.** A cross-lagged panel model for  $Y_1$  and  $Y_2$ . The second subscript indicates the time of measurement.

$X_2$ , both measured at time 1 only. Coefficients  $a$  and  $b$  in the figure correspond to, respectively, the **autoregressive paths**

$$Y_{11} \rightarrow Y_{12} \quad \text{and} \quad Y_{21} \rightarrow Y_{22}$$

The coefficient for each path just listed estimates the direct effect of the variables on itself, or the stability of  $Y_1$  and  $Y_2$  over time. Coefficients  $c$  and  $d$  in Figure 6.7 correspond to, respectively, the **cross-lagged paths**

$$Y_{11} \rightarrow Y_{22} \quad \text{and} \quad Y_{21} \rightarrow Y_{12}$$

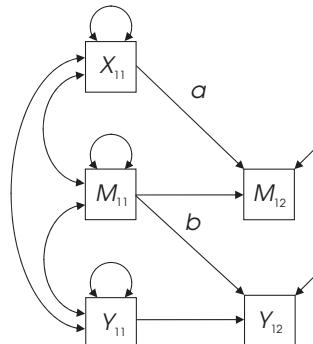
Coefficients for these paths estimate the direct effects of  $Y_1$  and  $Y_2$  on each other over time. Autoregressive and cross-lagged effects are each estimated controlling for the other. For example, coefficient  $c$  in the figure estimates the effect of  $Y_1$  at time 1 on  $Y_2$  at time 2, controlling for the autoregressive effects of  $Y_2$  from time 1. Exercise 6 asks you to count the numbers of observations and free parameters for the panel model in Figure 6.7.

Cross-lagged and autoregressive effects in panel models concern influences over time. Within-time associations in panel models are typically specified as unanalyzed, either between covariates, such as  $X_1 \rightsquigarrow X_2$  in Figure 6.7, or between the disturbances of endogenous variables measured at the same time, such as  $D_{11} \rightsquigarrow D_{21}$  at time 1. Because the pattern of disturbance correlations in the figure is bow-free, the whole model is recursive. The complexity of panel models can increase rapidly as more variables are added to the model. They can also be nonrecursive, if disturbance correlations are in a bow pattern (there are direct effects between the corresponding endogenous variables). See Frees (2004) for examples.

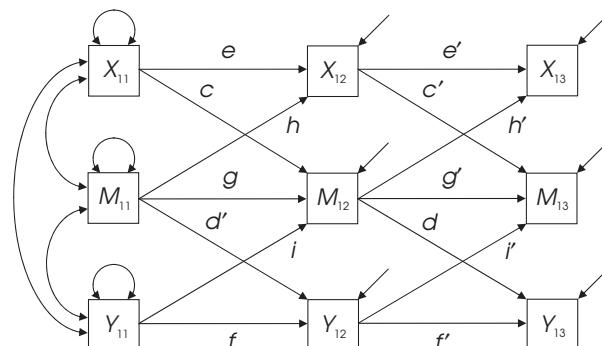
The model in Figure 6.8(a), shown using compact symbolism for disturbances, corresponds to a **half longitudinal design** for estimating mediation (Cole & Maxwell, 2003). In this model,  $X_1$  designates the causal variable,  $M_1$  the mediator, and  $Y_1$  the outcome variable. In such designs, the mediator and outcome are each measured at times 1 and 2, but the cause is measured only at time 1 (see the figure). At time 2, the disturbances of the mediator and outcome are assumed to covary, but this disturbance correlation is not shown in the figure to save space. Coefficient  $a$  in the figure model estimates the direct effect of the cause at time 1 on the mediator at time 2, controlling for the previous level of the mediator at time 1. Similarly, coefficient  $b$  in the figure estimates the direct effect of the mediator at time 1 on outcome at time 2, controlling for the previous level of the outcome. If all variables are continuous and assuming no interaction, the indirect effect of  $X$  on  $Y$  through  $M$  is estimated as the product  $ab$ . This estimator controls for autoregressive effects of both the mediator and outcome.

In a **full longitudinal design** (Cole & Maxwell, 2003), the cause, mediator, and outcome are each measured over at least three different times. This design is represented in

(a) Half longitudinal mediation



(b) Full longitudinal mediation



**FIGURE 6.8.** A half longitudinal model of mediation (a). A full longitudinal model of mediation (b). Compact symbolism is used for the disturbances. All pairwise disturbance correlations within each measurement occasion are assumed but are not shown. The second subscript indicates the time of measurement.

Figure 6.8(b) using compact symbolism. This model assumes that disturbances within the second and third measurement occasions are all pairwise correlated, but, to save space, they are not shown in the figure. Assuming no interactions and continuous variables, we estimate mediation in Figure 6.8(b) as the product of coefficients  $c$  and  $d$ , or  $cd$ . Coefficient  $c$  estimates the direct effect of the cause at time 1 on the mediator at time 2, controlling for the previous level of the mediator, and coefficient  $d$  estimates the direct effect of the mediator at time 2 on the outcome at time 3, controlling for the prior level of the outcome. Little (2013) notes that coefficients  $c$  and  $d$  correspond to the sole contiguous pathway in this model through which changes in the initial cause can indirectly affect the outcome, or

$$X_{11} \rightarrow M_{12} \rightarrow Y_{13}$$

The model for the full longitudinal design in Figure 6.8(b) contains two replications of the half longitudinal design. Specifically, (1) the product  $cd'$  for the paths

$$X_{11} \rightarrow M_{12} \quad \text{and} \quad M_{11} \rightarrow Y_{12}$$

is a proxy of the “real” estimator of mediation,  $cd$ . Likewise, (2) the product  $c'd$  for the paths

$$X_{12} \rightarrow M_{13} \quad \text{and} \quad M_{12} \rightarrow Y_{13}$$

estimates a different proxy. The values of these two proxy estimators just mentioned and that of  $cd$  should all be similar under the assumption of stationarity, which also predicts all the equivalences (within the limits of sampling error) listed next for the model in Figure 6.8(b):

$$\begin{array}{llll} c = c' & d = d' & e = e' & f = f' \\ g = g' & h = h' & \text{and} & i = i' \end{array}$$

Selig and Preacher (2009) describe additional longitudinal designs for estimating mediation.

## SUMMARY

Considered in this chapter was the specification of path models, or structural models of observed variables. Such models reflect hypotheses about spurious (noncausal) associations and direct or indirect causal effects among measured variables. Path models feature single-indicator measurement where each hypothetical construct is measured with a single manifest variable. Consequently, it is critical that single indicators have good psychometric properties. There is an emerging consensus that mediation analysis requires data from designs with time precedence, such as experimental or longitudinal

designs. Such designs guarantee that causes are measured before mediators, which are in turn measured before outcomes. Mediation always involves indirect effects, but indirect effects, estimated in cross-sectional designs where all variables are concurrently measured, do not automatically imply mediation. Rules that apply to all kinds of structural equation models without mean structures for counting the number of observations and the number of free model parameters were also presented. These rules are used to check whether a path model is identified, which is the topic of the next chapter.

### LEARN MORE

Antonakis et al. (2010) address causal inference from the perspectives of both design and analysis. Hoyle (2012) describes specification of models in SEM with observed or latent variables, and Kline (2012) outlines assumptions of different types of structural equation models.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21, 1086–1120.

Hoyle, R. H. (2012). Model specification in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 126–144). New York: Guilford Press.

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). New York: Guilford Press.

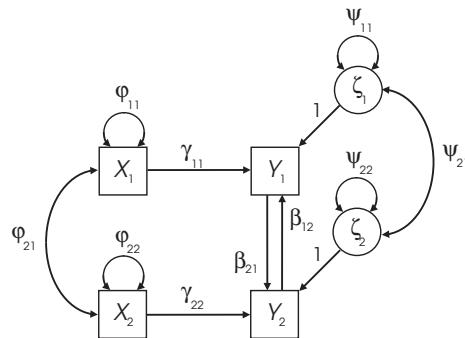
### EXERCISES

1. How would you represent membership in one of three different groups (i.e., a nominal variable) as a predictor in a path model?
2. Describe the consequences of increasing measurement error for the endogenous variable represented in Figure 6.3.
3. For the equivalent path models in Figures 6.4(a) and 6.4(b), compare path coefficients  $a$  and  $b$ , assuming that both  $X$  and  $Y$  are continuous, their relation is linear, and the estimation method is bivariate regression (i.e., OLS).
4. Use Rule 6.1 to count the number of free parameters for the path models in Figures 6.6(b) and 6.6(c).
5. Describe the hypotheses represented in Figure 6.6(c) about variables  $Y_1$  and  $Y_2$ .
6. Count the number of observations and free parameters for the panel model in Figure 6.7.

## Appendix 6.A

### LISREL Notation for Path Models

Described next is LISREL notation for path models when means are not analyzed. Measured exogenous variables are designated as  $X$  and measured endogenous variables as  $Y$ . Lowercase Greek letters in this notation include  $\beta$  (beta),  $\gamma$  (gamma),  $\zeta$  (zeta),  $\phi$  (phi), and  $\psi$  (psi); uppercase letters include  $B$  (beta),  $\Gamma$  (gamma),  $\Phi$  (phi), and  $\Psi$  (psi). Symbols for variables, parameters, and disturbances appear in their proper places in the nonrecursive path model shown next:



Structural equations for the endogenous variables are listed next:

$$\begin{aligned} Y_1 &= \gamma_{11} X_1 + \beta_{12} Y_2 + \zeta_1 \\ Y_2 &= \gamma_{22} X_2 + \beta_{21} Y_1 + \zeta_2 \end{aligned} \quad (6.2)$$

The equations for  $Y_1$  and  $Y_2$  can also be expressed in matrix algebra terms as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} = \boldsymbol{\Gamma} \mathbf{X} + \mathbf{B} \mathbf{Y} + \boldsymbol{\zeta} \quad (6.3)$$

where  $\boldsymbol{\Gamma}$  is the parameter matrix for direct effects of measured exogenous variables on the endogenous variables,  $\mathbf{X}$  is the matrix of measured exogenous variables,  $\mathbf{B}$  is the matrix for direct

effects of endogenous variables on each other,  $\mathbf{Y}$  is the matrix of endogenous variables, and  $\boldsymbol{\zeta}$  is the matrix of disturbances. Other parameter matrices include

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix} \quad (6.4)$$

where  $\boldsymbol{\Phi}$  is the covariance matrix of the measured exogenous variables and  $\boldsymbol{\Psi}$  is the covariance matrix of the disturbances. The four LISREL parameter matrices for manifest variable path models are thus

$$\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Phi}, \text{ and } \boldsymbol{\Psi}$$

# 7

## Identification of Observed-Variable (Path) Models

---

The topic of this chapter corresponds to the second step in SEM: the evaluation of identification, or whether it is theoretically possible for the computer to derive a unique set of model parameter estimates. This chapter shows you how to evaluate the identification status of path models estimated in single samples when means are not analyzed. A set of identification rules or heuristics that are relatively straightforward to apply is presented, and a research example dealt with in a later chapter is also introduced. Some of the topics discussed next are rather abstract, but many examples are offered, and chapter exercises provide more opportunities for practice. A Chinese proverb states that learning is a treasure that will follow you everywhere. After mastering the concepts in this chapter, you will be better prepared to apply SEM in your own studies.

---

### GENERAL REQUIREMENTS

There are two general—necessary, but insufficient—requirements for identifying any type of model in SEM. They are listed next and discussed afterward:

1. The model degrees of freedom must be at least zero ( $df_M \geq 0$ ).
2. Every latent variable—including disturbances or error terms—must be assigned a scale.

### Minimum Degrees of Freedom

The requirement for  $df_M \geq 0$  is the **counting rule**. For models that satisfy it, there are at least as many observations (Rule 6.2) as there are free model parameters (Rule 6.1).

Models that violate the counting rule are **underidentified** or **undetermined**. As an example of how a deficit of observations leads to nonidentification, consider the following equation:

$$a + b = 6 \quad (7.1)$$

Look at this expression as a model, the 6 as an observation, and  $a$  and  $b$  as free parameters (unknowns). Because Equation 7.1 has more free parameters than observations, it is impossible to find a set of unique estimates. In fact, there are an infinite number of solutions, including

$$(a = 4, b = 2), (a = 8, b = -2), \text{ and } (a = 2.5, b = 3.5)$$

and so on. A similar thing happens when a computer tries to derive unique estimates for an underidentified structural equation model: It is impossible to do so, and the attempt fails.

This next example shows that having equal numbers of observations and parameters does not guarantee identification. Consider the following set of formulas:

$$\begin{aligned} a + b &= 6 \\ 3a + 3b &= 18 \end{aligned} \quad (7.2)$$

Although this model has two observations (6, 18) and two free parameters ( $a, b$ ), it does not have a unique solution. Actually, an infinite number of solutions satisfy Equation 7.2, such as  $(a = 4, b = 2)$ , and so on. This happens because the second formula in Equation 7.2 is linearly dependent on the first formula: It is simply three times the first formula, so it cannot narrow the range of solutions that satisfy the first formula.

Now consider the following set of formulas with two observations and two free parameters where the second equation is not linearly dependent on the first:

$$\begin{aligned} a + b &= 6 \\ 2a + b &= 10 \end{aligned} \quad (7.3)$$

This two-observation, two-parameter model has a unique solution; it is  $(a = 4, b = 2)$ . Thus, the model defined by Equation 7.3 is **just-identified** or **just-determined**. Also, given estimates of its parameters, the equation can perfectly reproduce the observations (6, 10). Recall that most structural equation models with zero degrees of freedom that are also identified can perfectly reproduce the data (sample covariances), but such models test no particular hypothesis.

A statistical model can also have fewer parameters than observations. Consider the following set of formulas with three observations and two parameters:

$$\begin{aligned} a + b &= 6 \\ 2a + b &= 10 \\ 3a + b &= 12 \end{aligned} \tag{7.4}$$

There is no single solution that satisfies all three formulas. For example, the solution ( $a = 4, b = 2$ ) works only for the first two formulas in Equation 7.4, and the solution ( $a = 2, b = 6$ ) works only for the last two formulas. But there is a way to find a unique solution: Impose a statistical criterion that leads to an **overidentified** or **overdetermined** model with more observations than free parameters. An example for Equation 7.4 is the least squares criterion from regression analysis but with no intercept (constant) in the prediction equation. Expressed in words:

Find values of  $a$  and  $b$  that yield total scores such that the sum of squared differences between the observations (6, 10, 12) and these total scores is as small as possible.

Applying the criterion just stated to the estimation of  $a$  and  $b$  in a regression analysis yields a solution that not only gives the smallest possible squared difference (.67) but that is also unique, or ( $a = 3.00, b = 3.33$ ). This solution does not perfectly reproduce the observations (6, 10, 12): the predicted scores obtained from Equation 7.4 for the solution just stated are 6.33, 9.33, and 12.33. The fact that an overidentified model may not perfectly reproduce the data has an important role in model testing that is explored in later chapters.

The terms *just-identified* and *overidentified* do not automatically apply unless a model meets both of the necessary requirements mentioned at the beginning of this section and additional, sufficient requirements for that particular type of model described later. That is:

1. A **just-identified structural equation model** is identified and has the same number of observations as free parameters ( $df_M = 0$ ).
2. An **overidentified structural equation model** is identified and has more observations than free parameters ( $df_M > 0$ ).

A structural equation model can be underidentified in two ways. The first case occurs when there are more free parameters than observations ( $df_M < 0$ ), and thus the model fails the counting rule. The second case happens when some model parameters are underidentified because there is not enough information to estimate them but others are identified. In the second case, the whole model is considered nonidentified, even though  $df_M \geq 0$ . A general definition by Kenny (2011b) that covers both cases is

3. An **underidentified structural equation model** is one for which it is not possible to uniquely estimate all of its free parameters.

For overidentified path models, the value of  $df_M$  equals the number of “missing” paths,  $\rightarrow$  or  $\curvearrowright$ . But if all fixed-to-zero paths were added as free parameters for recursive models, the respecified path model would be just-identified ( $df_M = 0$ ). Deleted paths reflect strong causal assumptions and thus are more interesting than the specified (freely estimated) paths, which reflect weak hypotheses. This is because deleted paths can potentially falsify the model (Kenny & Milan, 2012). Fixing some paths to zero can also identify a nonrecursive path model, as we shall see.

Models with  $df_M > 0$  have **overidentifying restrictions**, which means that at least one free parameter has multiple estimates. These restrictions can be detected by resolving equations for the free parameters in terms of the covariances among the observed variables (Kenny & Milan, 2012, p. 148). Each overidentifying restriction corresponds to a particular test of local fit, but finding overidentifying restrictions in SEM can be tedious in larger models (i.e., it is impractical). Overidentifying restrictions are more transparent in graph theory (see Chapter 8), where it is not necessary to resolve equations in order to determine whether multiple estimators of a particular causal effect are available (i.e., there are overidentifying restrictions).

### Scaling Disturbances

In RAM symbolism, disturbances ( $D$ ) in structural models are represented as latent exogenous variables, and each  $D$  term requires a scale in order for the computer to estimate the error variance. Disturbance variances are usually scaled by imposing a **unit loading identification (ULI) constraint**. This means that the coefficient for the unstandardized residual path coefficient is fixed to 1.0 (i.e., a scaling constant). In diagrams, a ULI constraint is represented by the numeral “1” that appears next to the path from a disturbance to an endogenous variable. For example, the specification

$$D \rightarrow Y = 1.0$$

in Figure 6.2(b) assigns to  $D$  a scale that is related to the unexplained variance in  $Y$ . The implication is that the sum of the explained and unexplained variances in  $Y$  must equal the total observed variance (see Figure 6.3). The specification of any other scaling constant  $> 0$ , such as 17.3, would also scale the disturbance variance, but the equality just mentioned may not hold. Computer programs for SEM that automatically scale error terms impose ULI constraints as just described.

### UNIQUE ESTIMATES

The penultimate property of identification is that it must be possible to express each and every free model parameter as a unique function of elements in the population covariance matrix while satisfying the statistical criterion to be minimized. Because we typically estimate the population matrix with the sample matrix, this facet of identification

can be described by saying that there is a unique set of parameter estimates, given the data, model, and statistical criterion.

Determining whether the free parameters can be expressed as unique functions of the data is *not* an empirical question. Instead, it is a theoretical problem that can be evaluated by resolving equations for parameters in terms of symbols for elements of the data matrix. This process is basically a mathematical proof, so no actual numerical values are needed, just symbolic representations (Kenny & Milan, 2012, pp. 147–148). *This means that model identification can—and should—be evaluated before the data are collected.* As Kenny and Milan (2012) put it, what researcher wants to find out that his or her model is not identified after putting in the effort to collect the data?

You may have seen formal proofs for showing that formulas for regression coefficients and intercepts (e.g., Equations 2.2, 2.3) are, in fact, those that satisfy the least squares criterion. The derivation of a proof for a simple regression analysis would be a fairly daunting task for those without strong quantitative backgrounds, and models analyzed in SEM are often much more complicated than simple regression models. The statistical criterion minimized in maximum likelihood estimation is also more complicated than that minimized in the method of OLS estimation.

Unfortunately, SEM computer tools are of little help in determining whether a particular model is identified. Most programs perform rudimentary checks, such as applying the counting rule, but these checks do not generally involve sufficient conditions. It may surprise you to learn that SEM computer tools are rather helpless in this way, but there is a simple explanation: Computers are good at *numerical processing*, but it is *symbolic processing* that is needed for determining identification status. Computer languages for symbolic processing, such as Prolog (programming logic), form the basis of some applications in artificial intelligence. But contemporary SEM computer tools lack any real capability for symbolic processing of the kind needed to prove identification for a wide range of models.

Fortunately, one does not need to be a mathematician to deal with identification. This is because the rest of us can apply **identification heuristics**. These rules cover most of the structural equation models considered in this book. Kenny and Milan (2012) note that there may never be a single set of heuristics that cover all possible types of models; that is, the identification problem may be **undecidable**. But you can build a “toolbox” of heuristics that should work most of the time, beginning with the methods described next for path models. Graph theory (see Chapter 8) offers additional ways to evaluate identification, and related computer tools can analyze model diagrams in order to determine which causal effects are identified. These discussions assume that the two general requirements for identification are met.

## RULE FOR RECURSIVE MODELS

Because of their particular characteristics, recursive path models are identified (Bollen, 1989, pp. 95–98). This property is even more general: Recursive structural models

are identified, whether that model is a path model or the structural part of a structural regression (SR) model with latent variables. Note that whether the measurement part of an SR model with a recursive structural component is also identified is a separate question. The facts just reviewed explain the following sufficient condition for identification:

---

Recursive structural models are identified.

(Rule 7.1)

---

## IDENTIFICATION OF NONRECURSIVE MODELS

The situation concerning identification for nonrecursive structural models is more complicated because nonrecursive models that meet the general requirements may not be identified. There are algebraic methods to determine whether parameters of nonrecursive models can be expressed as unique functions of the observations (Berry, 1984, pp. 27–35), but these techniques are practical only for very simple models. Fortunately, there are special identification heuristics for nonrecursive models. Some of these heuristics concern necessary requirements, which do not guarantee identification, but others involve sufficient requirements that, if satisfied, prove identification. These heuristics are described next for nonrecursive path models, but the same principles apply to SR models with nonrecursive structural components.

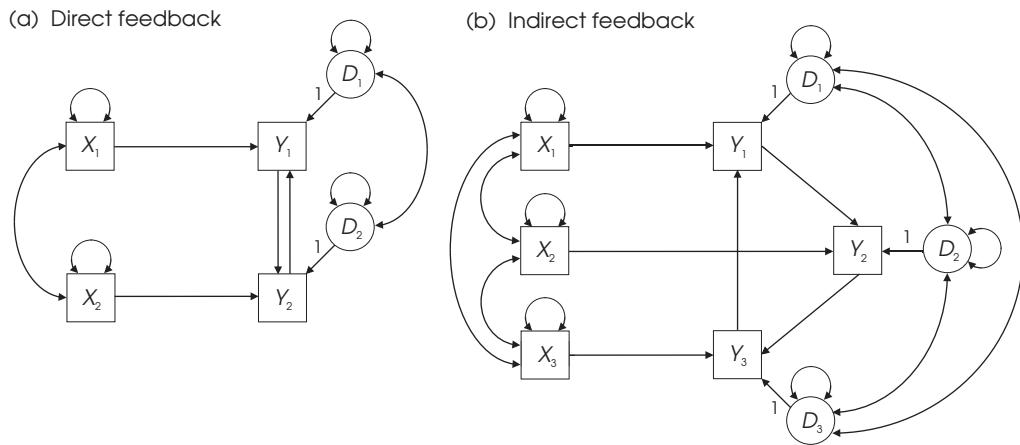
The nature and number of conditions for identification that a nonrecursive model must satisfy depend on its pattern of disturbance correlations. Considered next are identification heuristics for models with feedback loops and correlations between all pairs of disturbances either for the whole model or within blocks of recursively related endogenous variables.

### MODELS WITH FEEDBACK LOOPS AND ALL POSSIBLE DISTURBANCE CORRELATIONS

Presented in Figure 7.1(a) is the smallest identified model with all possible disturbance correlations (1 in total) and with no equality constraints for estimating direct feedback. Shown in Figure 7.1(b) is the corresponding model with all possible disturbance correlations (3 in total) and no equality constraints for estimating indirect feedback. Figure 7.1(a) is just-identified because  $df_M = 0$ , and Figure 7.1(b) is overidentified because  $df_M = 3$ . Exercise 1 asks you to verify these statements.

### Instrumental Variables

The direct feedback loop in Figure 7.1(a) consists of  $Y_1 \leftrightarrow Y_2$  and  $D_1 \curvearrowleft D_2$ , a bow-pattern disturbance correlation. These specifications imply that (1) causal variable  $Y_1$  covaries with the disturbance of its outcome, or  $D_2$ ; and (2) causal variable  $Y_2$  covaries



**FIGURE 7.1.** Two examples of nonrecursive models with feedback loops and all possible disturbance correlations.

with the disturbance of its outcome, or  $D_1$ . Thus, the requirement for pseudo-isolation in regression analysis is violated.

Nevertheless, Figure 7.1(a) is identified because each endogenous variable in a nonrecursive relation has a unique **instrument** or **instrumental variable** that identifies this relation. For example, the instrument for  $Y_1 \rightarrow Y_2$  is variable  $X_1$ , and variable  $X_2$  is the instrument for  $Y_2 \rightarrow Y_1$ . An instrument is unrelated to the disturbance of the outcome variable in a nonrecursive relation, whereas at the same time it has a direct or indirect effect on the causal variable in that relation. If the instrument has an indirect effect on the causal variable, then any intervening variable in that indirect pathway has no direct effect on the outcome variable. A variable that is predicted by the model to simply covary with the causal variable could also be an instrument if the candidate instrument has the other properties just mentioned. Finally, neither the causal nor outcome variable in a nonrecursive relation has a direct or indirect effect on the instrument; nor does any other variable in the model affect both the instrument and the outcome variable (Mulaik, 2009b). Graph theory offers a simpler definition of instruments as we will see in the next chapter.

In Figure 7.1(b), there are two instruments for each of  $Y_1-Y_3$ , the endogenous variables that make up the indirect feedback loop. Specifically, variables

1.  $X_1$  and  $X_3$  are the instruments for  $Y_1 \rightarrow Y_2$ ;
2.  $X_1$  and  $X_2$  are the instruments for  $Y_2 \rightarrow Y_3$ ; and
3.  $X_2$  and  $X_3$  are the instruments for  $Y_3 \rightarrow Y_1$ .

Instruments are defined by the model's specification (i.e., theory), not statistical analysis. Specifically, the coefficient for the direct effect of the instrument on the endogenous outcome variable is fixed to zero in the model. For example, do *not* regress  $Y_2$  on  $X_1$  and

$X_2$  in Figure 7.1(a) in order to find which predictor has a coefficient that is not statistically significant, and then use that exogenous variable as an instrument (Kenny, 2011a).

In bigger models, instruments can be exogenous or endogenous. Consider Figure 7.2 with two direct feedback loops and a disturbance correlation pattern described as **block recursive**. One can partition the endogenous variables of this model into two blocks, one with  $Y_1$  and  $Y_2$  and the other made up of  $Y_3$  and  $Y_4$ . Direct effects within each block are nonrecursive, but effects between the blocks are unidirectional (recursive). Each block contains all possible disturbance correlations (1 in total), but the disturbances across the blocks are independent. In the figure, variables  $X_1$  and  $X_2$  are the instruments for, respectively,  $Y_1$  and  $Y_2$  in the first block, and  $Y_1$  and  $Y_2$  are the instruments for, respectively,  $Y_3$  and  $Y_4$  in the second block. It is not too hard to locate instruments in smaller nonrecursive models with feedback loops, but doing so in larger models can be challenging. Fortunately, there are identification heuristics that are easier to apply.

### Order Condition

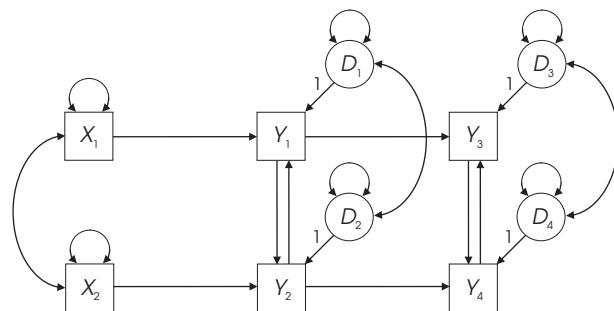
The **order condition** is a counting rule applied to each endogenous variable involved in a feedback loop in a model that has all possible disturbance correlations or that is block recursive. If the order condition is failed, the equation for that endogenous variable is underidentified. The order condition is necessary but insufficient, so failing it says that the model is not identified, but passing it does not guarantee that the model is actually identified.

The order condition is evaluated by counting the number of variables (except disturbances) that have direct effects on each endogenous variable versus the number that do not. Let's call the latter excluded variables. The condition is stated next:

---

The order condition requires that the number of excluded variables for each endogenous variable equals or exceeds the total number of endogenous variables minus 1. (Rule 7.2)

---



**FIGURE 7.2.** A block recursive model with two recursively related blocks of direct feedback and all possible disturbance correlations within each block. The whole model is nonrecursive.

For nonrecursive models with all possible disturbance correlations, the total number of endogenous variables equals that for the whole model. For example, there are three endogenous variables in Figure 7.1(b). This means that at least  $3 - 1 = 2$  variables must be excluded from the equation of each endogenous variable, which is here true. For example, variables  $X_2$ ,  $X_3$ , and  $Y_2$  are excluded from the equation for  $Y_1$ , which exceeds the minimum number (2). Because the equations for  $Y_2$  and  $Y_3$  also meet the order condition, the model passes the order condition.

For block recursive models, the total number of endogenous variables is counted separately for each block when the order condition is evaluated. For example, there are two blocks in Figure 7.2. Each block has two endogenous variables, so to satisfy the order condition, at least  $2 - 1 = 1$  variables must be excluded from the equation of each endogenous variable in both blocks, which is here true: One variable is excluded from each equation for  $Y_1$  and  $Y_2$  in the first block (e.g.,  $X_2$  for  $Y_1$ ), and three variables are excluded from each equation in the second block (e.g.,  $X_1$ ,  $X_2$ , and  $Y_2$  for  $Y_3$ ). This block recursive model passes the order condition.

## Rank Condition

The **rank condition** for identification is necessary and sufficient; thus:

---

Nonrecursive models that satisfy the rank condition are identified.	(Rule 7.3)
---	------------

---

This condition is usually described in matrix algebra terms (Bollen, 1989, pp. 101–103). One such definition is that at least one nonzero determinant must be associated with the matrix of coefficients for variables excluded from the equation of each endogenous variable involved in a feedback loop but included in the equations of other endogenous variables in the model.

The definition of the rank condition just stated is fine for those already familiar with matrix algebra. Berry (1984) devised a method for checking the rank condition that does not require detailed knowledge of matrix operations, a simpler version of which is described in Appendix 7.A. After applying this method, we can prove that the nonrecursive models in Figures 7.1(b) and 7.2 are actually identified (see the appendix). Exercise 2 asks you to evaluate the order condition and rank condition for Figure 7.1(a).

## GRAPHICAL RULES FOR OTHER TYPES OF NONRECURSIVE MODELS

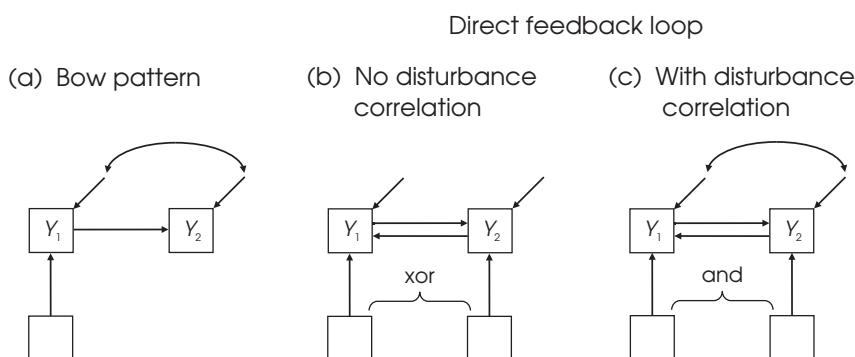
The order and rank conditions assume feedback loops and all possible disturbance correlations across the whole model or within recursively related blocks. If disturbance correlations are in any other pattern, the order condition is no longer necessary and the rank condition is no longer necessary and sufficient. Also, a structural model can be nonrecursive, yet have no feedback loops. Such models have disturbance correlations in

a bow pattern; that is, there is a direct effect between the corresponding pair of endogenous variables (e.g., Figure 6.6(c)).

Rigdon (1995) described a set of necessary and sufficient graphical rules for evaluating the identification status of nonrecursive models where the endogenous variables can be partitioned into sets of recursively related blocks. Each block contains either one or two variables. Blocks with two variables are reserved for pairs of variables in a nonrecursive relation (direct feedback or bow-pattern disturbance correlations). Blocks with a single variable are reserved for variables with strictly recursive relations to other variables. Exogenous variables are ignored when partitioning the endogenous variables.

Any single-variable block in Rigdon's graphical method is identified because it is recursive. Any two-variable nonrecursive block falls into one of eight patterns described by Rigdon (1995, p. 370), some of which correspond to identified blocks, but other patterns describe blocks that are not identified. Presented in Figure 7.3 are abstractions from Rigdon's types that represent the *minimum* required specifications for identifying blocks of two nonrecursively related variables,  $Y_1$  and  $Y_2$ . All other specifications beyond these minimum requirements are irrelevant (have no bearing on whether the nonrecursive block is identified or not identified). The requirements concern unique instruments, which are represented in the figure as unlabeled variables that could be exogenous or endogenous. All other variables in the model can be ignored, including those with direct effects on both  $Y_1$  and  $Y_2$ . This is because such variables cannot be instruments.

Figure 7.3(a) depicts a bow-pattern disturbance correlation. This nonrecursive relation is identified if the causal variable  $Y_1$  has a unique instrument. The outcome variable,  $Y_2$ , requires no instrument. This specification also provides a way to identify the direct effect of a putative exogenous variable on an outcome where the two variables share a disturbance correlation, which is a type of endogeneity. In the respecified model with an instrument, the formerly exogenous variable is now endogenous because it is regressed on the instrument. Shown in Figure 7.3(b) is a block with a direct feedback



**FIGURE 7.3.** Minimum required specifications for identifying blocks of two nonrecursively related endogenous variables,  $Y_1$  and  $Y_2$ , based on Rigdon's (1995) graphical classification system. Symbols for unlabeled variables represent unique instrumental variables; xor, exclusive or (either  $Y_1$  or  $Y_2$ , but not both).

loop between  $Y_1$  and  $Y_2$  that is identified if (1) there is only one unique instrument and (2) the disturbance correlation is fixed to zero. But if the disturbance correlation is a free parameter, then a unique instrument is needed for *both* variables in the feedback loop. This requirement is illustrated in Figure 7.3(c). Remember that the patterns in Figure 7.3 depict minimum identification requirements. Thus, any direct effects on  $Y_1$  or  $Y_2$  from variables that are not unique instruments have no bearing on the block's identification status.

Let's apply the patterns in Figure 7.3 to determine the identification status of the two nonrecursive models in Figure 7.4. Duncan (1975, pp. 84–86) algebraically proved that Figure 7.4(a) is not identified. To graphically demonstrate this fact, we partition the endogenous variables into two blocks. Block 1 is nonrecursive and includes variables  $Y_1$  and  $Y_2$  plus their disturbance correlation. Block 2 has a single variable,  $Y_3$ , which makes this block recursive and, thus, identified. But block 1 is not identified because neither  $Y_1$  nor  $Y_2$  has a unique instrument. Variable  $X_1$  is a common cause of  $Y_1$  and  $Y_2$ , but shared causes contribute nothing to the identification of a nonrecursive block. This means that Figure 7.4(a) is not identified because one of its two blocks is not identified.

Figure 7.4(b) is identified because the sole block in this model matches the pattern in Figure 7.3(b). Specifically, this model with a feedback loop but no disturbance correlation requires a unique instrument for either  $Y_1$  or  $Y_2$ , but not both. Exercise 3 asks you to verify that Figure 7.4(b) fails both the order and the rank conditions. In this case, the order condition is no longer necessary, and the rank condition is no longer necessary nor sufficient for this model with a feedback loop but no disturbance correlation. Exercise 4 asks how a researcher could respecify Figure 7.4(b) in order to test the assumption that the disturbance correlation equals zero.

## **RESPECIFICATION OF NONRECURSIVE MODELS THAT ARE NOT IDENTIFIED**

Suppose that the specification of Figure 7.4(a) faithfully reflects the hypotheses of a particular theory but, too bad, the model is not identified. (You should verify this statement.) What does the researcher do now? There are basically three options. One way is to simplify the model by dropping paths, which increases the value of  $df_M$  by one for each omitted path. For example, if the paths

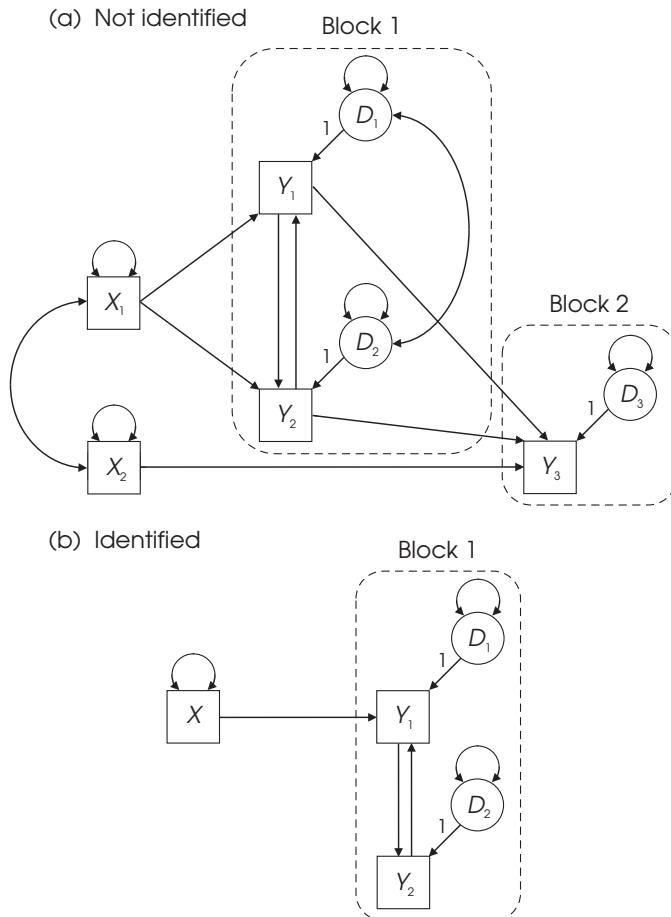
$$X_1 \rightarrow Y_2 \quad \text{and} \quad D_1 \curvearrowright D_2$$

were dropped from Figure 7.4(a), the respecified model with a feedback loop and a unique instrument,  $X_1$  for  $Y_1$  but no disturbance correlation would be identified. A different respecification is to drop the paths

$$Y_2 \rightarrow Y_1 \quad \text{and} \quad D_1 \curvearrowright D_2$$

from the original model, which results in a respecified model that is recursive, and thus identified. Both respecifications just described (among others not considered here) would fundamentally change the predictions represented in the original model of Figure 7.4(a).

A second way to effectively reduce the number of free parameters is to impose an equality constraint on the direct effects in a feedback loop. For example, the specification that both direct effects of  $Y_1 \leftrightarrow Y_2$  are equal implies that only one path coefficient is needed rather than two. A drawback of equality constraints is that they preclude the detection of unequal mutual influence. In contrast, a proportionality constraint allows for unequal mutual influence but on an a priori basis. Suppose that the value of the coefficient for  $Y_1 \rightarrow Y_2$  must be at least three times that of the coefficient for  $Y_2 \rightarrow Y_1$ . Again, only one path coefficient is needed instead of two because one coefficient is just a multiple of the other. But imposition of proportionality constraints generally requires knowledge about relative effect sizes.



**FIGURE 7.4.** Application of Rigdon's (1995) visual rules to nonrecursive models that are not identified (a) or identified (b).

The third option for respecifying a nonrecursive model that is not identified is to add unique instruments, one for each endogenous cause involved in a nonrecursive relation. Because exogenous variables are assumed to be uncorrelated with all disturbances, such variables are good candidates for instruments. Suppose that exogenous variables  $X_3$  and  $X_4$  were added to the model in Figure 7.4(a) along with the new paths

$$X_3 \rightarrow Y_1 \quad \text{and} \quad X_4 \rightarrow Y_2$$

plus unanalyzed associations between every pair of measured exogenous variables in the revised model ( $X_1$ – $X_4$ ). This respecified model would be identified—Exercise 5 asks you to verify this claim—and it includes the feedback loop and disturbance correlation from the original model. *Any respecification for the sake of identification requires theoretical justification*; see Paxton, Hipp, and Marquart-Pyatt (2011) for more information.

## A HEALTHY PERSPECTIVE ON IDENTIFICATION

Respecification of a model so that it is identified can at first seem like a shell game: Add this path, drop another, switch an error correlation, and—voilà!—the model is identified or—curses!—it is not. Although the researcher needs an identified model in SEM, it is critical to respecify models in a judicious manner. Any change to the original model for the sake of identification should therefore be guided by your hypotheses, not by empirical reasons. For example, one cannot estimate an equation, find that a path is close to zero, and then eliminate the path in order to identify the model (Kenny, Kashy, & Bolger, 1998). Don't lose sight of the ideas that motivated the analysis in the first place through haphazard specification.

## EMPIRICAL UNDERIDENTIFICATION

Although it is *theoretically* possible (that word again) for the computer to derive a unique set of estimates for the parameters of identified models, their analysis can still be foiled by other types of problems. Data-related problems are one such difficulty. For example, extreme collinearity can result in what Kenny (1979) referred to as **empirical underidentification**. If the correlation between two variables is very high (e.g.,  $r_{XY} = .95$ ), then, practically speaking, they are the same variable. This reduces the effective number of observations below the value of  $v(v+1)/2$  (the counting rule). An effective reduction in the number of observations can also shrink the effective value of  $df_M$ , perhaps to less than zero. The good news about this kind of empirical underidentification is that it can be detected through data screening.

Other types of empirical underidentification can be more difficult to spot, such as when estimates of certain key direct effects in a nonrecursive model equal a very small or a very high value. Suppose that the coefficient for the path  $X_2 \rightarrow Y_2$  in Figure 7.2 is

about zero. The virtual absence of this path means that  $Y_2$  has no unique instrument, which violates the rank condition. Other possible causes of empirical underidentification include (1) violation of normality or linearity assumptions when using normal theory estimation methods and (2) specification errors (Rindskopf, 1984).

## MANAGING IDENTIFICATION PROBLEMS

The best advice for avoiding problems was given earlier but is worth repeating: Evaluate whether a model is identified right after it is specified but before the data are collected (prevention is better than cure). If you know that your model is in fact identified and yet the analysis fails, the source of the problem may be a mistake in computer syntax or empirical underidentification. If a program error message indicates a failure of iterative estimation, another possible diagnosis is poor start values. How to specify better start values is described later in the book.

Perhaps the most challenging problems occur when analyzing a complex model for which no clear identification heuristic exists. This means that whether the model is actually identified is unknown. If the analysis fails in this case, it may be unclear whether the model is at fault (it is not really identified), the data are to blame (e.g., empirical underidentification), or the researcher made a mistake (e.g., syntax error). Ruling out a mistake does not solve the basic ambiguity. Here are some tips about how to cope:

1. A necessary but insufficient condition for the identification of a structural equation model is that the computer can generate a converged solution with no evidence of technical problems in the analysis. This empirical check can be applied to the actual data. Instead, you can use an SEM computer program as a diagnostic tool with made-up data in the form of a summary matrix that are anticipated to approximate actual values. This suggestion assumes that the data are not yet collected. Care must be taken not to generate hypothetical correlations or covariances that are out of bounds or that may result in empirical underidentification. If you are unsure about a particular made-up data matrix, then others with somewhat different but still plausible values can be constructed. If a computer program is unable to generate a proper solution, the model may not be identified; otherwise, it may be identified, but this is not a guarantee.

2. A common beginner's mistake is to specify a complex model of ambiguous identification status and then attempt to analyze it. If the analysis fails (likely), it may not be clear what caused the problem. Begin instead with a simpler model that is a subset of the target model and is also one for which the application of heuristics can prove identification. If the analysis fails, the problem is not identification; otherwise, add parameters to the simpler model one at a time. If the analysis fails after adding a particular effect, try a different order. If these analyses also fail at the same point, then adding the corresponding parameter may cause underidentification. If no combination of adding effects

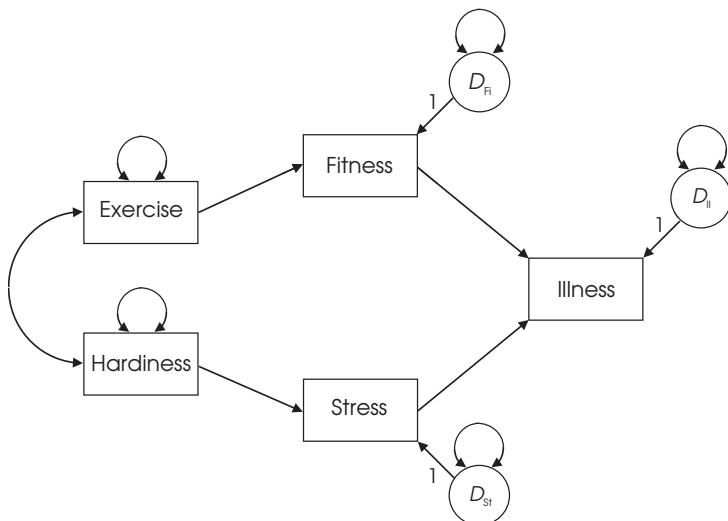
to a basic identified model gets you to the target model, think about how to respecify the target model in order to identify it and yet still respect your hypotheses.

## PATH ANALYSIS RESEARCH EXAMPLE

The data for this example were introduced earlier (see Table 4.2). Roth et al. (1989) administered measures of exercise, hardiness, fitness, stress, and illness. Because all variables were concurrently measured, we will refer to indirect effects, not mediation. The recursive path model in Figure 7.5 represents the hypotheses that the effects of exercise and hardiness on illness are purely indirect and that each effect is transmitted through a single intermediary, fitness for exercise and stress for hardiness. You should verify that  $df_M = 5$ . Detailed analysis of this model is described in Chapters 11–12.

## SUMMARY

It is easy to determine whether a recursive path model is identified. About all that is needed is to check whether the model degrees of freedom are at least zero and every disturbance is scaled. But the identification status of nonrecursive models is not always so clear. There are heuristics (order condition, rank condition) for models with all possible disturbance correlations either for the whole model or within recursively related blocks. There are also graphical criteria such that endogenous variables are partitioned into blocks of one or two variables, where the variables are nonrecursively related and instruments identify their effects. It is best to avoid analyzing a complex model of ambiguous



**FIGURE 7.5.** A recursive path model of illness.

identification status as your initial model. Instead, first analyze simpler models that you know are identified before adding parameters. The next chapter introduces graph theory for causal inference. It also offers some unique insights about the identification of path models.

### LEARN MORE

Kenny and Milan (2012) discuss identification of both structural models and measurement models, Paxton et al. (2011) describe the analysis of nonrecursive path models, and Rigdon (1995) outlines graphical identification criteria.

Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 145–163). New York: Guilford Press.

Paxton, P., Hipp, J. R., & Marquart-Pyatt, S. T. (2011). *Nonrecursive models: Endogeneity, reciprocal relationships, and feedback loops*. Thousand Oaks, CA: Sage.

Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30, 359–383.

### EXERCISES

1. Determine the value of  $df_M$  for each of the two nonrecursive models in Figure 7.1.
2. Evaluate the rank condition for the model in Figure 7.1(a).
3. Verify that the identified model in Figure 7.4(b) fails both the order and rank conditions.
4. Respecify Figure 7.4(b) to test the assumption that the disturbance correlation is zero.
5. Verify that Figure 7.4(a) is identified after adding  $X_3 \rightarrow Y_1$  and  $X_4 \rightarrow Y_2$  plus unanalyzed associations between every pair of measured exogenous variables in the revised model.

## Appendix 7.A

### Evaluation of the Rank Condition

Begin by constructing a **system matrix**, where the endogenous variables are represented in the rows and all variables are represented in the columns. In each row, a 0 or 1 appears in the column that corresponds to that row. A 1 indicates that the variable represented by the column has a direct effect on the endogenous variable represented by that row. A 1 also appears in the column that corresponds to the endogenous variable represented by that row. The remaining entries are 0's, and they indicate excluded variables. The matrix for Figure 7.1(b) is (I):

$$\begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \end{array} \left[ \begin{array}{cccccc} X_1 & X_2 & X_3 & Y_1 & Y_2 & Y_3 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{array} \right] \quad (\text{I})$$

"Reading" this matrix for  $Y_1$  indicates three 1's in its row, one in the column for  $Y_1$  itself, and the others in the columns of variables that directly affect it,  $X_1$  and  $Y_3$ . Because  $X_2$ ,  $X_3$ , and  $Y_2$  are excluded from  $Y_1$ 's equation, the entries in the columns for these variables are all 0's. Entries in the rows for  $Y_2$  and  $Y_3$  are specified in a similar way.

The rank condition is applied to the equation of each endogenous variable by working with the system matrix. The steps for models with all possible disturbance correlations are:

1. Begin with the first row of the system matrix. Cross out all entries in that row. Also cross out any column in the system matrix with a 1 in that same row. Save the entries that remain to form a reduced matrix. Variable labels are not needed in the reduced matrix.

2. Simplify the reduced matrix further by deleting any row with entries that are all zeros. Also delete any row that is an exact duplicate of another or can be reproduced by adding other rows (i.e., it is a linear combination of other rows). The number of remaining rows is the rank. For example, consider the following reduced matrix (II):

$$\left[ \begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{array} \right] \quad (\text{II})$$

The third row can be formed by adding the corresponding elements of the first and second rows, so it should be deleted. Therefore, the rank of this matrix (**II**) is 2, not 3. The rank condition is met for the equation of this endogenous variable if the rank of the reduced matrix is greater than or equal to the total number of endogenous variables minus 1.

**3.** Repeat steps 1 and 2 for every endogenous variable. If the rank condition is satisfied for every endogenous variable, then the model is identified.

Steps 1 and 2 applied to the system matrix for the model of Figure 7.1(b) are outlined here (**III**). Note that we are beginning with  $Y_1$ :

$$\blacktriangleright Y_1 \begin{bmatrix} X_1 & X_2 & X_3 & Y_1 & Y_2 & Y_3 \\ \pm & \theta & \theta & \pm & \theta & \pm \\ Y_2 & \theta & 1 & 0 & \pm & 1 & \theta \\ Y_3 & \theta & 0 & 1 & \theta & 1 & \pm \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \rightarrow \text{Rank} = 2 \quad (\text{III})$$

For step 1, all entries in the first row are crossed out. Also crossed out are three columns with a 1 in this row (i.e., those with column headings  $X_1$ ,  $Y_1$ , and  $Y_3$ ). The resulting reduced matrix has two rows. Neither row has entries that are all zero or can be reproduced by adding other rows together, so the reduced matrix cannot be simplified further. The rank of the equation for  $Y_1$  is 2, which equals the required minimum value, or one less than the total number of endogenous variables in the whole model. The rank condition is satisfied for  $Y_1$ .

The steps for the remaining endogenous variables in Figure 7.1(b) are summarized next:

Evaluation for  $Y_2$  (**IV**):

$$\blacktriangleright Y_2 \begin{bmatrix} X_1 & X_2 & X_3 & Y_1 & Y_2 & Y_3 \\ Y_1 & 1 & \theta & 0 & \pm & \theta & 1 \\ Y_2 & \theta & \pm & \theta & \pm & \pm & \theta \\ Y_3 & 0 & \theta & 1 & \theta & \pm & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \rightarrow \text{Rank} = 2 \quad (\text{IV})$$

Evaluation for  $Y_3$  (**V**):

$$\blacktriangleright Y_3 \begin{bmatrix} X_1 & X_2 & X_3 & Y_1 & Y_2 & Y_3 \\ Y_1 & 1 & 0 & \theta & 1 & \theta & \pm \\ Y_2 & 0 & 1 & \theta & 1 & \pm & \theta \\ Y_3 & \theta & \theta & \pm & \theta & \pm & \pm \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \rightarrow \text{Rank} = 2 \quad (\text{V})$$

The rank of the equations for each of  $Y_2$  and  $Y_3$  is 2, which exactly equals the minimum required value. Because the rank condition is satisfied for all three endogenous variables of this model, we conclude that it is identified.

The rank condition is evaluated separately for each block in Figure 7.2. First, construct a system matrix for each block. For example, the system matrix for the block that contains  $Y_1$  and

$Y_2$  lists only these variables plus prior variables  $X_1$  and  $X_2$ . Variables of the second block are not included in the matrix for the first block. The system matrix for the second block lists only  $Y_3$  and  $Y_4$  in its rows, but all variables are represented in its columns. Next, apply the rank condition to each system matrix of each block. The steps are outlined next. Evaluation for block 1 (**VI**):

$$\blacktriangleright Y_1 \begin{bmatrix} X_1 & X_2 & Y_1 & Y_2 \\ \theta & 0 & \pm & \pm \\ Y_2 & 0 & 1 & \pm & \pm \end{bmatrix} \rightarrow \begin{bmatrix} 1 \end{bmatrix} \rightarrow \text{Rank} = 1 \quad (\text{VI})$$

$$\blacktriangleright Y_2 \begin{bmatrix} X_1 & X_2 & Y_1 & Y_2 \\ 1 & 0 & \pm & \pm \\ Y_1 & \theta & \pm & \pm \end{bmatrix} \rightarrow \begin{bmatrix} 1 \end{bmatrix} \rightarrow \text{Rank} = 1$$

Evaluation for block 2 (**VII**):

$$\blacktriangleright Y_3 \begin{bmatrix} X_1 & X_2 & Y_1 & Y_2 & Y_3 & Y_4 \\ 0 & 0 & \pm & 0 & \pm & \pm \\ Y_4 & 0 & 0 & \theta & 1 & \pm & \pm \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \rightarrow \text{Rank} = 1 \quad (\text{VII})$$

$$\blacktriangleright Y_4 \begin{bmatrix} X_1 & X_2 & Y_1 & Y_2 & Y_3 & Y_4 \\ 0 & 0 & 1 & 0 & \pm & \pm \\ Y_3 & \theta & 0 & \theta & \pm & \pm & \pm \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \rightarrow \text{Rank} = 1$$

Because the rank of every equation is equal to or greater than the number of endogenous variables in each block minus 1 (i.e., 2 – 1), the rank condition is passed, so the block recursive model in Figure 7.2 is identified.

# 8

## Graph Theory and the Structural Causal Model

---

Basic concepts in graph theory and the structural causal model (SCM) are introduced in this chapter.<sup>1</sup> Many of these ideas build on what you already know and will help to extend your repertoire of skills. A vocabulary for causal graphs is presented next, and the concept of d-separation is explained. The d-separation concept is key to understanding the implications of causal hypotheses for model testing and whether particular causal effects are identified or not identified. This concept also provides a way to locate instruments in a causal graph. Causal mediation analysis, which is based on counterfactual definitions of direct and indirect effects and allows for interaction between the cause and mediator, is also introduced. Chapter exercises will help you to apply lessons learned from graph theory.

---

### INTRODUCTION TO GRAPH THEORY

Graph theory for causal modeling originated in Judea Pearl's work in the 1980s about **Bayesian networks**, or probabilistic graphical models that represent dependences among sets of random variables. They are represented in computer memory as graph-type structures, which are then virtually navigated by the computer in order to update conditional probabilities of events. Initially applied to discrete variables with multinomial distributions, graphical models can also represent dependence relations among other types of variables, including continuous ones with joint multivariate distributions.

During the 1980–1990s and still ongoing, the ideas behind Bayesian networks and causal graphs have been extended to the broader problem of causal inference in research (Pearl 2000, 2009b). This framework, now known as the SCM, is becoming well known

---

<sup>1</sup>I thank Judea Pearl and Bryant Chen for comments on earlier drafts of this chapter.

in epidemiology (Rothman, Greenland, & Lash, 2008), but less so in psychology, education, and related areas. This is unfortunate because the SCM has the features listed next that address fundamental challenges in causal inference (Pearl, 2009a):

1. Causal hypotheses are represented both graphically and in expressions that are a kind of mathematical language subject to theorems, lemmas, and proofs.
2. The SCM provides a precise language for communicating the assumptions behind causal questions to be answered.
3. The SCM explicitly distinguishes between questions that can be empirically tested versus those that are unanswerable, given the model. It also provides ways to determine what new measurements would be needed to address an “unanswerable” question.
4. Finally, the SCM subsumes other useful theories or methods for causal inference, including the potential outcomes model and SEM.

Causal hypotheses are represented in either a **directed acyclic graph** (DAG) with no causal loops or a **directed cyclic graph** (DCG) with causal loops. The former (DAG) is the nonparametric generalization of recursive path models with unidirectional effects regardless of the pattern of correlated disturbances (if any), and the latter (DCG) is the nonparametric analog of nonrecursive path models with feedback loops. Because graphical models in the SCM are nonparametric, the researcher makes no commitment to distributional assumptions for any individual variable. A direct causal effect is also nonparametric because it represents all forms of the functional relation between cause and effect. If variables  $X$  and  $Y$  are both continuous, for example, the specification  $X \rightarrow Y$  in a directed graph represents the linear and all curvilinear trends of the causal effect of  $X$  on  $Y$ . But in parametric path models, the specification  $X \rightarrow Y$  represents just the linear trend for continuous variables.

## Graph Vocabulary

We need a basic vocabulary for directed graphs. Variables are also referred to as **nodes** or **vertices**. Some variables in directed graphs are connected by **arcs**, also known as **edges** or **links** (i.e., paths). Arcs may designate either functional or statistical dependences between variables in the graph. A pair of vertices is **adjacent** if they are connected by an edge; otherwise, that pair is **nonadjacent**. The weak hypothesis of a direct causal effect is represented by a directional edge, or **arrow** ( $\rightarrow$ ), that points from cause to effect. The absence of an arrow between a pair of variables reflects the strong hypothesis of no direct causal effect between them. A pair of variables connected by the symbol for bidirectional edge,  $\nwarrow\swarrow$ , is assumed to share an unmeasured (latent) cause. The corresponding symbol in SEM,  $\curvearrowleft\curvearrowright$ , also allows for unmodeled causal effects between pairs of exogenous variables or even a more complex structure, such as a latent common cause plus a direct effect (Hayduk et al., 2003).

The direct causes of a variable are its **parents**, and all direct and indirect causes of a variable are its **ancestors**. In addition, all variables directly caused by a given variable are its **children**, and its **descendants** include all variables directly or indirectly caused by that same variable. All parents in a directed graph are ancestors just as all children are descendants. A variable with no parents is exogenous, and a variable with at least one parent is endogenous, just as in path models. Disturbances are not usually shown in directed graphs, if they are assumed to be independent; otherwise, a pair of endogenous variables presumed to share error variance (i.e.,  $\geq 1$  omitted confounder) is directly connected by the symbol for a bidirectional edge.

A **path** is a sequence of adjacent edges that connects two variables regardless of the directions of those edges. It passes through any variable along the path just once. In a **directed path**, all edges are arrows that point away from a cause toward an effect. Such paths convey causal information from the beginning of the path to the end. All other paths are **undirected paths** that may convey statistical association—but not causation—between the variables at either end. The goal of specifying a directed graph is the same as that for a path model: The graph represents all hypothesized connections, causal or noncausal, between any pair of variables.

## ELEMENTARY DIRECTED GRAPHS AND CONDITIONAL INDEPENDENCES

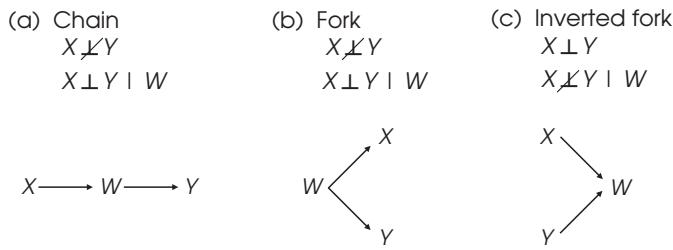
Three elementary structures make up directed graphs: chains, forks, and inverted forks. These structures correspond to, respectively, causation, confounding, and colliders (Elwert, 2013). A **contracted chain**, such as  $X \rightarrow Y$ , is the smallest causal structure. The two variables in a contracted chain are unconditionally dependent because there are no intervening variables that could disrupt or block the causal coordination between parent and child.

A larger **chain** has at least three variables, such as the directed path

$$X \rightarrow W \rightarrow Y$$

which is also shown in Figure 8.1(a). This DAG represents the indirect causal effect of  $X$  on  $Y$  through the intermediary  $W$ . The direct effect of  $X$  on  $Y$  is assumed to equal zero. The whole chain is also a **front-door path** that starts with an arrow pointing away from  $X$  toward the end of the path. Both contracted chains in the figure,  $X \rightarrow W$  and  $W \rightarrow Y$ , are also front-door paths.

Figure 8.1(a) depicts a simple **Markov chain** where only adjacent variables, such as  $X$  and  $W$ , have direct causal dependences. The nonadjacent pair  $X$  and  $Y$  is dependent due to the directed path that runs through  $W$ , but this dependence can be broken, if the intermediary  $W$  is deactivated. One way to do so is to estimate the association between  $X$  and  $Y$  conditioning on (controlling for)  $W$ . For example, regressing  $Y$  on both  $X$  and  $W$  would render  $X$  and  $Y$  statistically independent. This is because  $W$  is the sole interven-



**FIGURE 8.1.** Elementary structures in directed acyclic graphs with three variables and corresponding implied dependences and independences. (a) Chain, (b) fork, and (c) inverted fork with a collider.

ing variable between  $X$  and  $Y$ , and deactivating  $W$  blocks causal coordination between this pair of variables.<sup>2</sup> The statistical independence of  $X$  and  $Y$ , given  $W$ , is represented by the expression

$$X \perp\!\!\! \perp Y \mid W$$

Another way to describe this **conditional independence** is to say that controlling for  $W$  blocks the directed path between  $X$  and  $Y$  previously opened by  $W$ .

The conditional independence just described relies on the **Markov assumption** that every variable in a DAG is independent of all its nondescendants conditional on its parents. In Figure 8.1(a), variables  $X$  and  $Y$  are unrelated, given  $W$ , the parent of  $Y$ . As a nonparametric causal model, the same independence is expected to hold even if the variables are non-normally distributed and even if the causal effects are nonlinear (Hayduk et al., 2003). But when *not* conditioning on  $W$ , the same model also predicts that  $X$  and  $Y$  are dependent, or

$$X \not\perp\!\!\! \perp Y$$

That is, the **marginal association** between  $X$  and  $Y$  in the population is not zero, if we ignore  $W$ .

The smallest undirected path with a common ancestor of two other variables, such as

$$X \leftarrow W \rightarrow Y$$

and also shown in Figure 8.1(b), includes a **fork**, where  $W$  is specified as a direct cause of both  $X$  and  $Y$ . This fork is a **back-door path** that starts with an arrow pointing toward

<sup>2</sup>Other ways to condition on a variable include stratification, subgroup analysis, or sampling from populations with certain values on key variables (e.g., a survey of employed mothers). Conditioning can also happen inadvertently due to missing data (Elwert, 2013).

$X$  ( $\rightarrow X$ ). A back-door path may convey a spurious association between variables at either end, but never causation. The graph in the figure implies that  $X$  and  $Y$  are independent, given their common cause  $W$ . Thus, conditioning on  $W$  blocks the back-door path between  $X$  and  $Y$ , rendering them unrelated. The same graph also implies that any observed association between  $X$  and  $Y$  ignoring  $W$  is spurious. These same predictions can be summarized as

$$X \perp Y | W \quad \text{and} \quad X \not\perp Y$$

Because Figures 8.1(a) and 8.1(b) make the same predictions about conditional independencies, they are equivalent. Exercise 1 asks you to find a third DAG for the same three variables that is equivalent to the two graphs just mentioned.

The smallest graph with an undirected path that includes an **inverted fork** is

$$X \rightarrow W \leftarrow Y$$

which is also depicted in Figure 8.1(c). Here, variable  $W$  is a **collider**, or common outcome, because it lies along an undirected path with two arrows pointing into it. Note that a variable can be a collider along one path but not a collider along a different path in the same graph, but it is common to refer to variables with at least two parents as colliders. The term *collider* suggests a pileup of causal forces. Any path with a collider is blocked, closed, or inactive. This is because a collider blocks any association (including causal effects) between variables at either end of a path with a collider. A path with no collider is open, active, or unblocked, and thus potentially conveys statistical association through the path. For example, the paths between  $X$  and  $Y$  in Figures 8.1(a) and 8.1(b) are open because they have no collider, but the path between the same pair of variables in Figure 8.1(c) is blocked by the collider  $W$  (the path between  $X$  and  $Y$  is closed).

The presence of colliders in causal graphs has a special significance that may be overlooked in regression analysis. Consider Figure 8.1(c), where  $X$  and  $Y$  are specified as independent causes of  $W$ . One prediction of this graph is

$$X \perp Y$$

which says that  $X$  and  $Y$  are independent without controlling for any other variables (i.e., the conditioning set is  $\emptyset$ , the empty set). But if we condition on their common outcome—the collider—the same graph also predicts

$$X \not\perp Y | W$$

That is, controlling for a common outcome of two unrelated causes induces a spurious association between them. Even if two causes are correlated, controlling for their com-

mon outcome adds a spurious component to their observed association. It does so by unblocking the path between them previously closed by the collider.

Here is an intuitive example: Suppose that students in a private school were selected because of musical giftedness or athletic prowess, which we assume are unrelated. If we know that a student has no musical talent, then by default we can say that the student is an athlete, and vice versa; that is, refutation of one cause of an outcome confirms the action of the other cause. Likewise, given confirmation of one cause eliminates the need to invoke the other cause, which is described as the **explaining away effect** in the artificial intelligence literature and as **Berkson's paradox** in the statistical literature (Pearl, 2009b). In this example, musical and athletic abilities would be negatively related among students enrolled in the school. The composition of the sample is a collider independently caused by musicality and athleticism, and conditioning on that collider induces a negative correlation between its independent causes.

Another way to condition on a collider is through statistical control. Suppose that all variables in Figure 8.1(c) are continuous and all causal effects are linear. Given

$$\rho_{XW} = .30, \rho_{YW} = .40, \text{ and } \rho_{XY} = 0$$

the partial correlation between  $X$  and  $Y$  controlling for  $W$  is  $\rho_{XY|W} = -.137$  (Equation 2.15); see Hayduk et al. (2003, pp. 309–311) for a graphical illustration.

That conditioning on a collider imparts spurious association between its causes is not intuitive, but it is a real phenomenon. Perhaps even less intuitive but just as profound is the fact that conditioning on the *descendent* of a collider also induces spurious association. For example, if variable  $A$  were added to Figure 8.1(c) as a child of the collider (i.e.,  $W \rightarrow A$ ), the revised graph would also predict

$$X \not\perp\!\!\!\perp Y | A$$

Although variable  $A$  in the revised graph does not lie along a path between  $X$  and  $Y$ , conditioning on  $A$  will nevertheless open a path between them, where  $A$  is a descendant of the collider  $W$ . Epidemiologists are generally familiar with this concept mainly due to Pearl's formalization of causal models as directed graphs, but researchers in other areas may be less familiar with these issues.

Because causal graphs in the SCM are nonparametric, they allow for the possibility that two causes of the same outcome may interact. Interaction (moderation) means for the graph in Figure 8.1(c) that the effect of  $X$  on  $W$  varies across the levels of  $Y$ . Because moderation is symmetrical, the effect of  $Y$  on  $W$  would also vary as a function of  $X$ . In contrast, the hypothesis of moderation must be explicitly represented in a parametric causal model, such as a traditional path model. The assumption that causes and mediators interact is also part of causal mediation analysis, a topic covered in a later section of this chapter.

## IMPLICATIONS FOR REGRESSION ANALYSIS

The concepts just reviewed have implications for covariate selection in regression analysis, given a causal model for the problem. It is appropriate to control for the confounding effects of a common cause (e.g., Figure 8.1(b)); otherwise, confounding bias may occur. This idea is well known among most researchers. Inadvertently controlling for a mediator when regressing an outcome on an indirect cause can lead to **overcontrol bias** (Elwert, 2013), which eliminates some or all of the causal pathway (e.g., Figure 8.1(a)). This is why multiple regression assumes no causal effects among the predictors; that is, there is a single equation, that of the criterion.

Again assuming a causal model, controlling for a collider—or for the descendant of a collider—can lead to **collider bias** (Rothman et al., 2008) or **endogenous selection bias** (Elwert, 2013), where spurious associations are induced that may be falsely interpreted as evidence for causation (e.g., Figure 8.1(c)). This problem is especially critical when researchers control for what they believe to be background (causal) variables that are really outcomes with two or more parents.

## d-SEPARATION

Pearl's (2009b) **d-separation criterion** (d is for "directional") locates conditional independences in the data; that is, it tells us which pairs of variables are made independent by conditioning on certain other variables in the graph. These control variables block the flow of information between a focal pair of variables due to indirect causal effects or common causes (Hayduk et al., 2003). The same criterion also warns against inducing spurious association by conditioning on colliders or on their descendants. It relies on the Markov assumption that every omitted common cause is represented by the symbol for a bidirectional edge. The additional assumption of **faithfulness** states that direct versus indirect effects of one variable on another do not exactly cancel each other out (sum to zero). In other words, the graph only implies the conditional independences generated by the d-separation criterion (Elwert, 2013).

The d-separation criterion is defined next for a pair of variables (Glymour, 2006, p. 394), but it applies to sets of variables, too:

---

A pair of variables in a DAG is d-separated by a set of covariates  $Z$  if either (Rule 8.1)

1. one of the noncolliders on the path is in  $Z$ ; or
2. there is a collider on the path, but neither the collider nor any of its descendants is in  $Z$ .

A pair of variables  $X$  and  $Y$  is d-separated by  $Z$  if and only if  $Z$  blocks every path from  $X$  to  $Y$ .

---

A pair of variables is **d-connected** (unblocked, open), if not every path between them is d-separated. If the DAG is faithful, d-connectedness implies statistical dependence. Pairs of variables in parent-child relations are inherently d-connected. In graphs where every pair of variables is connected by an edge, there are no d-separated sets of variables. Because such graphs imply no conditional independences, they have no statistical implications that can be refuted with data. The analogous case in path analysis is when  $df_M = 0$ . In fact, a necessary condition for  $df_M = 0$  is that the graph implies no d-separations.

Let's try two examples. You can use the freely available Belief and Decision Network Tool (Porter et al., 1999–2009) to verify each example. Three nonadjacent pairs of variables in Figure 8.2(a) can be d-separated, including  $(X, Y)$ ,  $(X, B)$ , and  $(A, Y)$ . Variables  $X$  and  $Y$  at the ends of the directed path

$$X \rightarrow A \rightarrow B \rightarrow Y$$

are rendered independent by conditioning on any combination of the intervening variables,  $A$  or  $B$ . Doing so closes the directed path between this pair. Thus, the graph implies

$$X \perp Y | A \quad X \perp Y | B \quad \text{and} \quad X \perp Y | (A, B)$$

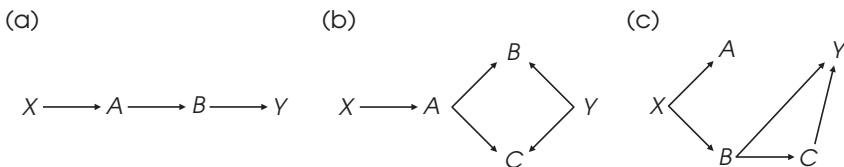
The pair  $(X, B)$  are independent, given  $A$ , the sole intermediary between them. Variable  $Y$  is the child of  $B$ , and so it contributes nothing to the association between  $X$  and  $B$ . Thus, the graph implies

$$X \perp B | A \quad \text{and} \quad X \perp B | (A, Y)$$

Finally, the nonadjacent pair  $(A, Y)$  is independent, given  $B$ . The same pair is also unrelated, given both  $B$  and  $X$ , which follows from the Markov assumption that controlling for the parent of  $Y$ , or  $B$ , shields  $Y$  from the influence of any other ancestor, or  $X$ . Thus, the graph also implies

$$A \perp Y | B \quad \text{and} \quad A \perp Y | (B, X)$$

Altogether, Figure 8.2(a) implies the seven conditional independences just listed.



**FIGURE 8.2.** Larger directed acyclic graphs.

Now let's consider the larger graph in Figure 8.2(b). There are five nonadjacent pairs of variables that can be d-separated, including

$$(X, Y), (X, B), (X, C), (B, C), \text{ and } (A, Y)$$

Listed in Table 8.1 are all conditional independences located in this graph by the d-separation criterion. Briefly, the pair  $(X, Y)$  is marginally independent with no covariates. This is because every path that connects  $X$  and  $Y$ , or

$$\begin{aligned} X &\rightarrow A \rightarrow B \leftarrow Y \\ X &\rightarrow A \rightarrow C \leftarrow Y \end{aligned}$$

is blocked by a collider,  $B$  or  $C$ . The same pair remains independent if  $A$  is the sole covariate. This is because controlling for  $A$  does not open any path between  $X$  and  $Y$  that is already closed by a collider. Conditioning on any combination of  $B$  or  $C$  without also controlling for  $A$  would open at least one path between  $X$  and  $Y$  and thus induce a spurious association. But including  $A$  in a conditioning set that includes  $B$  or  $C$  would close the path again. All five conditional independences just mentioned for the pair  $X$  and  $Y$  are listed in Table 8.1.

The logic for generating conditional independences for the pairs  $(X, B)$  and  $(X, C)$  in Figure 8.2(b) is similar, so only the first pair of variables just mentioned is considered next. Two paths connect this pair:

$$\begin{aligned} X &\rightarrow A \rightarrow B \\ X &\rightarrow A \rightarrow C \leftarrow Y \rightarrow B \end{aligned}$$

**TABLE 8.1. Conditional Independences Located by the d-Separation Criterion in Figure 8.2(b)**

Nonadjacent pair	Conditional independences	
$X, Y$	$X \perp Y$	$X \perp Y   A$
	$X \perp Y   (A, B)$	$X \perp Y   (B, C)$
	$X \perp Y   (A, B, C)$	
$X, B$	$X \perp B   A$	$X \perp B   (A, C)$
	$X \perp B   (A, Y)$	$X \perp B   (A, C, Y)$
$X, C$	$X \perp C   A$	$X \perp C   (A, B)$
	$X \perp C   (A, Y)$	$X \perp C   (A, B, Y)$
$B, C$	$B \perp C   (A, Y)$	$B \perp C   (A, X, Y)$
$A, Y$	$A \perp Y$	$A \perp Y   X$

The first (directed) path just listed is unblocked, but the second (undirected) path is blocked by the collider  $C$ . Conditioning on  $A$  alone blocks the directed path between  $X$  and  $B$  while at the same time leaving the undirected path between them blocked. Thus, variables  $X$  and  $B$  are independent, given  $A$ . For the same reasons, conditioning on both  $A$  and  $Y$  also d-separates the pair  $(X, B)$ . Including the collider  $C$  in a conditioning set would open the undirected path between  $X$  and  $B$ , but controlling for  $A$  would close the path again. Variable  $Y$  along with both  $C$  and  $X$  also d-separate the pair  $X$  and  $B$ . The four conditional independences for the pair  $(X, B)$  are listed in Table 8.1 along with the four conditional independences for the pair  $(X, C)$ .

The common causes of the pair  $(B, C)$  in Figure 8(b) are both  $A$  and  $Y$ ; the paths are

$$\begin{array}{c} B \leftarrow A \rightarrow C \\ B \leftarrow Y \rightarrow C \end{array}$$

Thus, conditioning on both  $A$  and  $Y$  renders the pair  $(B, C)$  independent. The same pair will also be independent controlling for  $A$ ,  $Y$ , and  $X$  because controlling for the parent of  $B$ , or  $A$ , isolates  $B$  from its only other ancestor, or  $X$ . Finally, both paths that connect the pair  $(A, Y)$  are blocked:

$$\begin{array}{c} A \rightarrow B \leftarrow Y \\ A \rightarrow C \leftarrow Y \end{array}$$

Thus, the pair  $(A, Y)$  is marginally independent. The same pair is also independent given  $X$ , the parent of  $A$ . All four conditional independences just described for the pairs  $(B, C)$  and  $(A, Y)$  are listed in Table 8.1. Altogether Figure 8.2(b) implies a total of 17 conditional independences. Exercise 2 asks you to find the conditional independences implied by Figure 8.2(c).

Finding conditional independences in directed cyclic graphs with causal loops requires a little adaptation. One reason is that graphical computer tools that locate conditional independences generally analyze only directed acyclic graphs. The procedure described in Appendix 8.A transforms a DCG to a special kind of DAG known as a **collapsed graph**, which implies all the essential conditional independences of the original DCG.

## BASIS SET

The **basis set** for a DAG includes the smallest number of conditional independences that imply all others (if any) located by the d-separation criterion. The size of the basis set equals the number of pairs of nonadjacent variables that can be d-separated. Any conditional independences beyond this number are implied by the basis set; that is, they are redundant, so there is no need to test them all. For example, there are three pairs of

nonadjacent variables in Figure 8.2(a) that can be d-separated, so the size of the basis set is 3. This graph implies the 7 conditional independences listed in the previous section, but not all of them are logically independent. A basis set of 3 for this graph will explain all the rest. Figure 8.2(b) generates the 17 conditional independences listed in Table 8.1, but the size of the basis set is 5, or the number of nonadjacent variables in the graph that can be d-separated. Thus, a smaller set of only 5 conditional independencies—the basis set—will generate all 17 conditional independencies listed in the table for this graph.

A basis set for a DAG may not be unique because there is more than one way to generate a basis set, but any such set predicts all conditional independencies implied by the graph. Defined next is a straightforward method by Pearl and Meshkat (1999) and Shipley (2000):

---

(Rule 8.2)

List each unique pair of nonadjacent variables in the graph that can be d-separated. Next, condition on the parents of both variables in each pair. The corresponding set of conditional independencies is a basis set.

---

Let's apply Rule 8.2 to Figure 8.2(a). Three pairs of nonadjacent variables can be d-separated, so the size of the basis set is 3. The parents of endogenous variables  $A$ ,  $B$ , and  $Y$  are, respectively,  $X$ ,  $A$ , and  $B$ . Thus, a basis set for the graph is

$$\begin{aligned} X \perp B &| A \\ A \perp Y &| (X, B) \\ X \perp Y &| B \end{aligned}$$

The basis set just listed explains all 7 conditional independencies implied by Figure 8.2(a). A researcher would need to test only the smaller basis set for this graph.

Now let's find a basis set for Figure 8.2(b). A total of five pairs of nonadjacent variables can be d-separated, so the size of the basis set is 5. Each nonadjacent pair of variables is listed in Table 8.2. Also reported in the table are the parents of both nonadjacent variables (the conditioning set) and the corresponding d-separation statement. The 17 conditional independencies implied by the graph and listed in Table 8.1 can all be derived from the 5 conditional independencies that make up a basis set listed in Table 8.2. Exercise 3 asks you to find a basis set for Figure 8.2(c).

## CAUSAL DIRECTED GRAPHS

A **causal directed graph** includes, for any pair of variables, all common causes, whether measured or unmeasured. The unmeasured causes correspond to hypothetical left-out variables and thus are latent. It is not required that all the causes of every endogenous

**TABLE 8.2. A Basis Set for Figure 8.2(b)**

Nonadjacent pair	Parents of both variables	Conditional independence
$X, Y$	None	$X \perp Y$
$X, B$	$A, Y$	$X \perp B \mid (A, Y)$
$X, C$	$A, Y$	$X \perp C \mid (A, Y)$
$B, C$	$A, Y$	$B \perp C \mid (A, Y)$
$A, Y$	$X$	$A \perp Y \mid X$

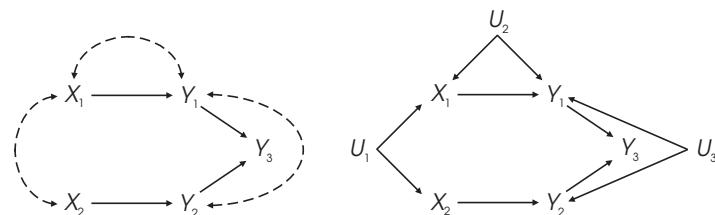
variable be included in a causal graph, but any common cause must be specified or else the graph is not truly causal. The specification error of omitting an edge (e.g.,  $\rightarrow$ ) does not make a graph noncausal. This is because the absence of an edge reflects an assumption (e.g., a direct causal effect is zero), and omitting it makes the causal graph misspecified.

A presumed latent common cause is represented in a directed graph by connecting a pair of measured variables with the symbol for a bidirectional edge. Consider the causal DAG in Figure 8.3(a). The bidirectional edge between variables  $Y_1$  and  $Y_2$  in this graph corresponds to a disturbance correlation, and the bidirectional edge that connects the two exogenous variables,  $X_1$  and  $X_2$ , assumes that they covary. The third bidirectional edge in the figure connects variables  $X_1$  and  $Y_1$ , where  $X_1$  is specified to directly cause  $Y_1$ . The specification just mentioned says that any part of their statistical association beyond that due to the causal effect of  $X_1$  on  $Y_1$  is spurious due to an unmeasured common cause.

Omitted confounders implied in Figure 8.3(a) are explicitly shown as the unmeasured variables  $U_1-U_3$  in Figure 8.3(b). The d-separation criterion is applied to the latter graph in the normal way except that  $U_1-U_3$  are never in the covariate set because they are latent and thus not measured. There are six pairs of nonadjacent measured variables, including

$$(X_1, Y_3), (X_1, Y_2), (X_2, Y_3), (X_1, X_2), (Y_1, Y_2), \text{ and } (X_2, Y_1)$$

(a) Implied omitted causes      (b) Explicit omitted causes

**FIGURE 8.3.** Causal graph with bidirectional edges that imply unmeasured common causes (a). Same graph but with explicit omitted causes  $U_1-U_3$  (b).

but the last three pairs just listed cannot be d-separated because some common causes are latent. For example, the pair  $(X_1, X_2)$  shares  $U_1$  as common cause, or

$$X_1 \leftarrow U_1 \rightarrow X_2$$

but  $U_1$  is unmeasured, so the pair  $(X_1, X_2)$  cannot be d-separated. Similarly, the path

$$Y_1 \leftarrow U_3 \rightarrow Y_2$$

for the pair  $(Y_1, Y_2)$  cannot be blocked by conditioning because  $U_3$  is latent. Exercise 4 asks you to prove that the pair  $(X_2, Y_1)$  in Figure 8.3(b) cannot be d-separated. Listed next is a set of conditional independences for the remaining three pairs of nonadjacent measured variables that can be d-separated:

$$\begin{aligned} X_1 \perp Y_3 &| (Y_1, Y_2) \quad \text{and} \quad X_1 \perp Y_3 &| (Y_1, Y_2, X_2) \\ &X_1 \perp Y_2 &| X_2 \\ X_2 \perp Y_3 &| (Y_1, Y_2) \quad \text{and} \quad X_2 \perp Y_3 &| (Y_1, Y_2, X_1) \end{aligned}$$

A smaller basis set for this graph is

$$\begin{aligned} X_1 \perp Y_3 &| (Y_1, Y_2) \\ X_1 \perp Y_2 &| X_2 \\ X_2 \perp Y_3 &| (Y_1, Y_2) \end{aligned}$$

## TESTABLE IMPLICATIONS

Each d-separation statement in a causal directed graph corresponds to a prediction that is potentially testable in sample data. If all variables are continuous in a linear model, each conditional independence matches up with a partial correlation that should equal zero. In models with independent error terms, the whole set of “vanishing” partial correlations represents all testable implications of the model. For example, the basis set for Figure 8.3(b) given in the previous section implies for continuous variables the vanishing partial correlations listed next:

$$\rho_{X_1 Y_3 \cdot Y_1 Y_2} = \rho_{X_1 Y_2 \cdot X_2} = \rho_{X_2 Y_3 \cdot Y_1 Y_2} = 0$$

If any of these predictions are inconsistent with the data, the associated conditional independence is not supported. This outcome may help to diagnose misspecification in a particular part of the graph. Suppose we observe in a sample that  $r_{X_1 Y_2 \cdot X_2} = .40$ , which contradicts the prediction of zero. This result suggests that there may be addi-

tional omitted direct or indirect causes of  $X_1$  and  $Y_2$  besides  $U_1$ , among other possibilities for specification error in Figure 8.3(b). A benefit of local fit testing for this example is that  $r_{X_1 Y_2 \cdot X_2}$  cannot be distorted by measurement error in  $Y_1$  or  $Y_3$ . Shipley (2000) describes local fit testing in path analysis based on conditional independences. We will apply some of these methods to the analysis of a path model with actual data in Chapter 11.

## GRAPHICAL IDENTIFICATION CRITERIA

Identification means basically the same thing in SEM and the SCM: whether a causal characteristic of a model can be uniquely determined by the data, but the emphasis in the two methods is somewhat different. In SEM, identification is attached to parameters of parametric models, and there is a “heavyweight” process for evaluating identification that involves manipulating equations or applying heuristics (e.g., the rank condition). In the SCM, there are graphical identification criteria for finding a **sufficient set** of covariates that identifies a particular causal effect. Controlling for a sufficient set removes spurious components (back-door paths), leaving just the causal relation. If there is no sufficient set, the corresponding causal effect may be identified by other methods, such as the specification of instruments for variables in nonrecursive relations.

Graphical identification criteria in the SCM are based on the concept of d-separation. This helps the researcher to avoid selecting inappropriate covariates that fail to remove common cause confounding or that introduce new biases such as that due to conditioning on a collider. The main idea is that a sufficient set of covariates blocks all noncausal (back-door) paths between  $X$  and  $Y$  while not blocking any causal (directed) paths. Next we assume a causal DAG. The graphical identification criteria described next do not apply to a collapsed graph based on a DCG.

### Back-Door Criterion

A set of covariates  $Z$  (which may consist of  $\emptyset$ ) is sufficient for identifying the total causal effect (the sum of all direct and indirect effects) of  $X$  on  $Y$ , if that set meets the **back-door criterion** (Pearl, 2009b, 79–80):

---

A set of covariates  $Z$  satisfies the back-door criterion relative to the total causal effect of  $X$  on  $Y$  if (Rule 8.3)

1. no variable in  $Z$  is a descendant of  $X$ ; and
2.  $Z$  blocks all back-door paths between  $X$  and  $Y$ .

---

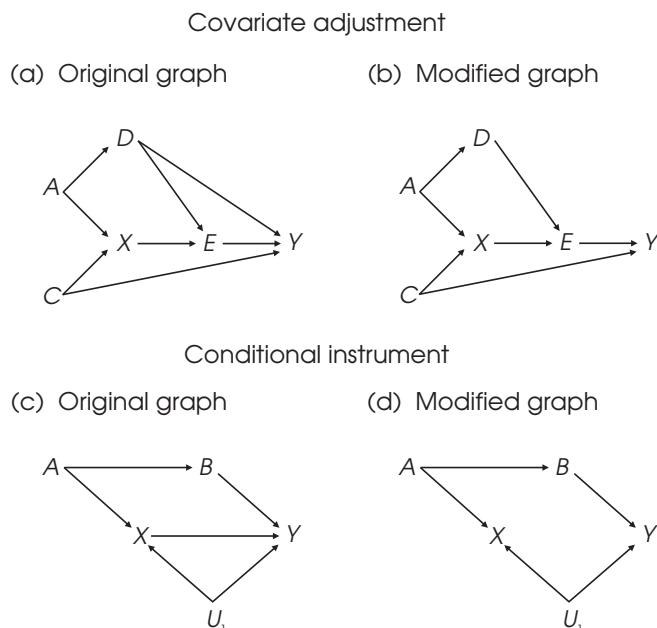
If  $Z$  meets the back-door criterion, then  $Z$  is sufficient to identify the total effect of  $X$  on  $Y$ .

---

Shown in Figure 8.4(a) are a total of three back-door paths between  $X$  and  $Y$ . They are

$$\begin{aligned} X &\leftarrow C \rightarrow Y \\ X &\leftarrow A \rightarrow D \rightarrow Y \\ X &\leftarrow A \rightarrow D \rightarrow E \rightarrow Y \end{aligned}$$

Two sufficient sets that block all the back-door paths just listed and thus identify the total causal effect of  $X$  on  $Y$  are  $(A, C)$  and  $(C, D)$ . Each set just listed is also a **minimally sufficient set** for which no proper subset is itself a sufficient set. (A proper subset does not include the original set.) This means that both covariates in each of the two minimally sufficient sets just listed are needed to block all back-door paths between  $X$  and  $Y$ . The larger set  $(A, C, D)$  is also sufficient, but it is not minimally sufficient. This is because two of its proper subsets include the minimally sufficient sets  $(A, C)$  and  $(C, D)$ . This example shows that there can be multiple sets of sufficient covariates, each of which identifies the same causal effect. This is a kind of overidentifying restriction in that estimates based on using different sufficient sets should all be equal, if the graph is correct. Exercise 5 asks you to prove that  $(A)$  and  $(X, C)$  are each minimally sufficient sets that identify the total effect of  $D$  on  $Y$  in Figure 8.4(a).



**FIGURE 8.4.** An original directed acyclic graph (a). The graph modified by deleting the direct effect from  $D$  to  $Y$  in the original graph (b). An original directed acyclic graph with omitted confounder  $U_1$  (c). The graph modified by deleting the direct effect from  $X$  to  $Y$  in the original graph (d).

Because exogenous variables have no ancestors, there are no back-door paths when the causal variable is exogenous. In this case, the sufficient set is  $\emptyset$  (i.e., no covariates). For example, variable  $A$  in Figure 8.4(a) is exogenous and has several indirect effects on  $Y$  through variables  $D$ ,  $X$ , and  $E$ . There are no back-door paths between  $A$  and  $Y$ , so the sufficient set is  $\emptyset$ . Thus, regressing  $Y$  on  $A$  with no covariates would estimate the total effect. Because the back-door criterion is sufficient to identify a total effect, there is no need to identify parameters along all the separate paths from  $A$  to  $Y$  in the figure.

In a smaller graph, it is not too difficult to spot sufficient sets that block back-door paths (if any), but it can be challenging in bigger graphs. This is where a freely available computer tool for analyzing directed acyclic graphs comes in handy. For example, DIGitty (Textor et al., 2011) and the DAG Program (Knüppel & Stang, 2010) each automatically list minimally sufficient sets (if any) that identify the total effect between a pair of variables selected by the user. Both programs also allow the specification of omitted confounders when analyzing a causal DAG. These capabilities allow the researcher to explicitly assess whether unmeasured confounders preclude the identification of any total causal effect through covariate adjustment.

As an exercise, specify the causal DAG in Figure 8.3(b) with confounders  $U_1-U_3$  in either computer tool just mentioned. You will see that some total effects in this graph cannot be identified through covariate adjustment, including the total effects of

$$X_1 \text{ on } Y_1, X_1 \text{ on } Y_3, \text{ and } X_2 \text{ on } Y_3$$

all due to omitted confounders. The observation that some causal effects are not identified should motivate the researcher to find approximate measures of omitted causes—or add instrumental variables to the model as another option—in order to identify those causal effects.

Other total effects in Figure 8.3(a) are identified through covariate adjustment. For example, the total effect of  $X_2$  on  $Y_2$  (which is also a direct effect) is identified because there are no unblocked back-door paths in Figure 8.3(b) between this pair of variables. Thus, (1) the sufficient set is  $\emptyset$  (i.e., no covariates are needed) and (2) regressing  $Y_2$  on  $X_2$  estimates their causal relation. You should also verify for the figure that the total effects of  $Y_1$  on  $Y_3$  and of  $Y_2$  on  $Y_3$  are identified by, respectively, the sufficient sets ( $Y_2$ ) and ( $Y_1$ ). This example shows that it may be possible in the SCM to “test the testable” (estimate identified causal effects) for graphs in which some, but not all, causal effects are identified (Pearl, 2009b, pp. 144–145).

## Single-Door Criterion

In a recursive linear model with continuous variables, the **single-door criterion** tells us whether the coefficient for a particular direct effect ( $\rightarrow$ ) is identified by covariate adjustment, and what variables should serve as the conditioning set (Pearl, 2009b, pp. 150–152):

A set of variables  $Z$  satisfies the single-door criterion relative to the pair variables  $X$  and  $Y$  if (Rule 8.4)

1. no variable in  $Z$  is a descendant of  $Y$ ; and
2.  $Z$  d-separates  $X$  and  $Y$  in the modified graph formed by deleting the edge  $X \rightarrow Y$  from the original graph.

If  $Z$  meets the single-door criterion, the coefficient for the direct effect of  $X$  on  $Y$  is identified.

---

Look back at Figure 8.4(a). We proved earlier that the total effect of  $D$  on  $Y$  is identified through covariate adjustment. This total effect consists of a direct effect and an indirect effect through variable  $E$ . Is the direct effect of  $D$  on  $Y$  also identified? Assuming a linear model and continuous variables, we can apply the single-door criterion. First, we delete the path  $D \rightarrow Y$  from the original model. Presented in Figure 8.4(b) is the graph so modified. There are no descendants of  $Y$ . Two minimally sufficient sets d-separate  $D$  and  $Y$  in the modified graph,  $(C, E)$  and  $(A, X, E)$ ; that is, both

$$D \perp Y | (C, E) \quad \text{and} \quad D \perp Y | (A, X, E)$$

are true in the modified graph. This means that the coefficient for the path  $D \rightarrow Y$  is identified in the original graph (Figure 8.4(a)). Exercise 6 asks you to find minimally sufficient sets of covariates that identify the coefficient for the path  $C \rightarrow Y$  in Figure 8.4(a).

## INSTRUMENTAL VARIABLES

Another way to identify direct effects in linear models involves instrumental variables. For example, an instrument for a causal variable  $X$  involved in a nonrecursive relation with outcome variable  $Y$  should be correlated with  $X$  but should be uncorrelated with the error term of  $Y$  (e.g., Figure 7.3). In larger models with multiple outcomes or omitted confounders, it can be difficult to determine proper instruments. Graph theory offers a clear definition of instruments based on the concept of d-separation (Pearl, 2012, pp. 82–83):

---

A variable  $Z$  is a proper instrument for the path  $X \rightarrow Y$  if (Rule 8.5)

1.  $Z$  is d-separated from  $Y$  in the modified graph formed by deleting the edge  $X \rightarrow Y$  from the original graph; and
  2.  $Z$  is not d-separated from  $X$  in the modified graph.
-

Consider the causal DAG in Figure 8.4(c) for a linear model and continuous variables. The direct effect of  $X$  on  $Y$  is not identified. This is because in the modified graph where the path  $X \rightarrow Y$  is deleted, two unblocked paths remain, including

$$\begin{aligned} X &\leftarrow A \rightarrow B \rightarrow Y \\ X &\leftarrow U_1 \rightarrow Y \end{aligned}$$

The second path cannot be closed by conditioning on the latent variable  $U_1$ ; thus, the direct effect of  $X$  on  $Y$  cannot be identified by covariate adjustment. Is variable  $A$  in Figure 8.4(c) a proper instrument for  $X$ ? No, it is not. This is because in the modified graph without the path  $X \rightarrow Y$ , which is shown in Figure 8.4(d), variable  $A$  is d-connected to  $Y$  by the path

$$A \rightarrow B \rightarrow Y$$

which violates the first requirement of Rule 8.5. With no instrument, it might seem that the direct effect of  $X$  on  $Y$  in Figure 8.4(c) is not identified unless more measured variables are added.

But we can actually find a proper instrument for  $X$  in Figure 8.4(c) by creating a **conditional instrument** that satisfies Rule 8.5. Here, variable  $A$  is rendered a proper instrument after controlling for  $B$ . This is because the conditional instrument  $A | B$  is d-separated from  $Y$ , or

$$(A | B) \perp Y$$

in the original graph (Figure 8.4(c)), but  $A | B$  is not d-separated from  $X$ , or

$$(A | B) \not\perp X$$

in the modified graph where the path  $X \rightarrow Y$  is deleted (Figure 8.4(d)). Because Rule 8.5 is satisfied, variable  $A | B$  is a proper instrument that identifies the direct effect of  $X$  on  $Y$ . For more examples of conditional instruments, see Brito and Pearl (2002).

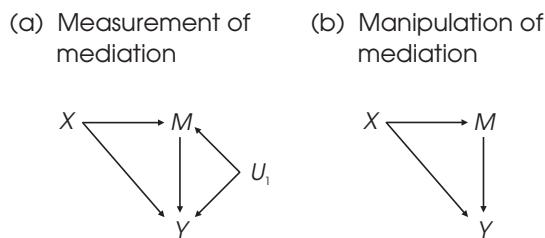
## CAUSAL MEDIATION

Estimation of mediation in the SCM is referred to as **causal mediation analysis**. It is based on Pearl's (2014) mediation formula, which (1) allows for interaction between the causal variable and the mediator and (2) defines mediation in a consistent way for linear versus nonlinear models and also for continuous versus binary mediators or outcomes. These definitions are based on a counterfactual approach to mediation. Next, we assume a design where  $X$  is an experimental variable with two levels (treatment, control),  $M$  is the presumed mediator, and  $Y$  is the outcome.

Presented in Figure 8.5(a) is a basic **measurement-of-mediation model** (Bullock, Green, & Ha, 2010), where the mediator is an individual difference variable that is measured, not manipulated (e.g., patient compliance). Because  $X$  is a manipulated variable, over replications it will be isolated from confounders that also affect  $M$  or  $Y$ . This explains why there are no confounders in Figure 8.5(a) for either the pair ( $X, M$ ) or the pair ( $X, Y$ ). But because neither  $M$  nor  $Y$  are manipulated, it is plausible that they share at least one confounder (i.e., their disturbances are correlated), represented as  $U_1$  in the figure. Consequently, the coefficient for the path  $M \rightarrow Y$  is not identified.

One way to identify the coefficient for the path  $M \rightarrow Y$  in a measurement-of-mediation model is to specify an instrument for the mediator, such as a variable that directly affects  $M$  but not  $Y$  and is also unrelated to the disturbance of  $Y$  (Antonakis et al., 2010). Exogenous variables make ideal instruments because by definition they are unrelated to all disturbances in the model. An example is a manipulated instrument where cases are randomly assigned to conditions that should directly affect  $M$  but not  $Y$ . Suppose that patients are randomly assigned to conditions that offer varying levels of incentive (including none) for adherence to treatment. This randomized manipulation to change  $M$  may indirectly affect  $Y$ , but over replications there should be no direct effect. Instruments can also be measured variables, but finding nonexperimental instruments can be challenging. In designs with a single mediator, it must be assumed that  $M$  completely mediates the relation between  $X$  and  $Y$ . This is a strong assumption that is not always convincing. See MacKinnon and Pirlott (2015) for additional discussion of instrumental variable methods in mediation analysis.

A stronger model is a **manipulation-of-mediation model** (Bullock et al., 2010), in which the mediator is also a manipulated variable. Examples of mediators that are potentially manipulable include self-efficacy, goal difficulty, performance norms, and arousal, but manipulating other kinds of internal states or situational factors may be difficult or unethical in some cases (Stone-Romero & Rosopa, 2011). It must be assumed in experimental mediational designs that manipulation affects just the mediator in question and no other mediators. This requirement can be tricky given multiple mediators. The results may apply only to the subset of participants who responded to the manipulation of the mediator and not to the whole sample. Challenges of manipulating mediators



**FIGURE 8.5.** Basic nonparametric mediational models where  $X$  is randomized and where the mediator  $M$  is a measured variable (a) versus a manipulated variable (b). The omitted confounder of  $M$  and  $Y$  is  $U_1$ .

are substantial, but successfully doing so isolates over replication studies any confounders of  $X$  or  $M$  and  $Y$ . It also identifies the coefficient for the path  $M \rightarrow Y$  without an explicit instrument.

As nonparametric causal models, both Figure 8.5(a) and Figure 8.5(b) allow for the possibility that  $X$  and  $M$  interact in how they affect  $Y$ . In contrast, the parametric counterpart of Figure 8.5(b)—which is presented in Figure 6.5(b) as a standard path model—assumes that the cause and the mediator do *not* interact. In linear models with continuous mediators and outcomes and assuming no interaction, there is a single estimator of the direct effect of  $X$  on  $Y$  (e.g., coefficient  $e$  in Figure 6.5(b)). But if we allow for interaction between  $X$  and  $M$ , there is more than one direct effect of  $X$  because it changes over the levels of  $M$ . If the latter is continuous, then theoretically  $X$  has an infinite number of direct effects. This plurality of direct effects changes the meaning of indirect and total effects, too.

In Pearl's SCM, the researcher routinely assumes that  $X$  and  $M$  interact, and thus estimates the magnitude of this interaction in causal mediational analysis. This assumption underlies the distinction between a **controlled direct effect** (CDE) and a **natural direct effect** (NDE).<sup>3</sup> Each type of direct effect just mentioned is also associated with a different statement of counterfactual propositions. They are defined next for a binary causal variable  $X$  (0 = control, 1 = treatment) (Petersen, Sinisi, & van der Laan, 2006).

The CDE estimates how much  $Y$  would change as  $X$  changes from one level to the other if the mediator  $M$  were controlled uniformly at the same fixed value for all cases. If  $X$  and  $M$  interact, the value of the direct effect of  $X$  changes depending on the level of  $M$ . If  $M$  is a continuous variable, then  $X$  has an infinite number of direct effects on  $Y$ . In practice, the CDE may be estimated as the level of this direct effect at a weighted average value of  $M$ . The NDE allows for variation in the mediator. Specifically, it estimates how much  $Y$  would change on average if  $X$  were allowed to change from control to treatment, but the mediator  $M$  were kept at the level it would have taken in the control condition ( $X = 0$ ). This says that  $X$  is allowed to change, but  $M$  is held constant at the values that would be naturally observed among untreated cases. Unlike the case for the CDE, the level of  $M$  is not fixed to the same value for all cases. If there is no interaction, then estimates of the controlled direct effect and the natural direct effect are equal for continuous variables in a linear model.

The parallel to the NDE is the **natural indirect effect** (NIE). It estimates the amount of change in  $Y$  in the treatment group ( $X = 1$ ) if the mediator were changed from the level that would be observed in the control group to the level it would take in the treatment group; that is,  $Y$  is influenced by  $X$  due solely to its influence on  $M$  (Muthén, 2011). The total effect is thus

$$\text{TE} = \text{NDE} + \text{NIE} \quad (8.1)$$

---

<sup>3</sup>With no interaction of  $X$  and  $M$  in a linear model with continuous mediator and outcome variables, CDE = NDE for causal variable  $X$ .

and it estimates the causal effect of  $X$  on  $Y$  both directly and indirectly through  $M$ . The effect decomposition in Equation 8.1 holds in both linear and nonlinear models regardless of interaction. In contrast, the CDE of  $X$  does not have a simple additive relation to any of the quantities in Equation 8.1, so it is not part of the definition of a total effect when  $X$  and  $M$  interact.

At first glance, the counterfactual definitions of the various effects just reviewed can seem awfully abstract, but this example by Petersen et al. (2006) may help to clarify things: Suppose that  $X = 1$  is an antiretroviral therapy for HIV-infected persons and  $X = 0$  is control. The mediator  $M$  is the blood level of HIV (viral load), and outcome  $Y$  is the level of CD4 T-cells, or “helper” white blood cells that are part of the immune system. The CDE of treatment is the difference in average CD4 T-cell count if viral load were controlled at a single level for all cases. In contrast, the NDE is the effect of treatment on CD4 T-cell count as it would have been observed if viral load were as in the control condition. The NIE is the change in CD4 T-cell count in treatment if viral load shifted to what it would be without treatment to the level under treatment. The total effect of antiretroviral therapy on CD4 T-cell count both directly and indirectly through viral load is the sum of its NDE and NIE. See Appendix 8.B for definitions of these various effects in the language of counterfactuals and expected values.

Estimation of the direct and indirect effects just defined generally requires the assumption of no confounding of the relation between any pair of variables among  $X$ ,  $M$ , and  $Y$ . These assumptions are probably met when both  $X$  and  $M$  are manipulated (Figure 8.5(b)), but probably not when  $M$  is measured, not manipulated (Figure 8.5(a)). For natural effects, it is also assumed that no confounder of the relation between  $M$  and  $Y$  is caused by  $X$ . Altogether these requirements are very strict. Estimating mediation typically relies on very strong assumptions, but it is better to be aware of all that is assumed in mediational analyses (Bullock et al., 2010). Estimation of mediation in traditional path analysis has the same general requirements *plus* the additional assumption that  $X$  and  $M$  do not interact.

## SUMMARY

Graph theory and the SCM offer unique perspectives on causal modeling. One is the capability to generate testable implications about conditional independences between pairs of measured variables. Evaluation of each specific prediction with data is a local fit test. Another contribution is a principled framework for specifying covariates in regression analysis when there is a causal model for the problem. By analyzing a causal graph, the researcher can tell whether to include or exclude certain variables as covariates when estimating causal effects between a pair of variables  $X$  and  $Y$ . Specifically, variables that block noncausal associations, or back-door paths, between  $X$  and  $Y$  should be selected as covariates, but other variables that are mediators or colliders along paths that connect  $X$  and  $Y$  should generally *not* be selected. If no set of measured variables meets both requirements just listed—that is, there is no sufficient set of covariates—then the

researcher must rely on other methods, such as instrumental variables, to identify the causal effect. In causal mediation analysis, the cause and the mediator are assumed to interact, and counterfactual-based definitions of direct and indirect effects apply to mediators or outcomes that are continuous or binary in linear or nonlinear models. Considered in the next chapter are the specification and identification of CFA measurement models in SEM.

## LEARN MORE

Elwert (2013) and Glymour (2006) offer accessible introductions to graph theory, and Pearl (2012) elaborates on the causal foundation of SEM.

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). New York: Springer.

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In M. Oakes & J. Kaufman (Eds.), *Methods in social epidemiology* (pp. 387–422). San Francisco: Jossey-Bass.

Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York: Guilford Press.

## EXERCISES

1. Specify a DAG that implies the same conditional independence as Figures 8.1(a) and 8.1(b).
2. List all the conditional independences implied by the graph in Figure 8.2(c).
3. Find a basis set for Figure 8.2(c).
4. Show that the pair  $X_2$  and  $Y_1$  in Figure 8.3(b) cannot be d-separated.
5. Show that  $(A)$  and  $(X, C)$  are minimally sufficient sets for the total effect of  $D$  on  $Y$  in Figure 8.4(a).
6. Assuming a linear model, find all minimally sufficient sets that identify the coefficient for the direct effect of  $C$  on  $Y$  in Figure 8.4(a). Hint: There are three.

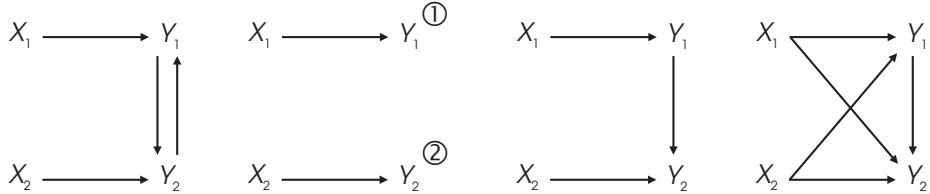
## Appendix 8.A

### Locating Conditional Independences in Directed Cyclic Graphs

Spirites (1995) described a set of rules to transform a DCG with causal loops to a special DAG with no causal loops for the sake of obtaining conditional independences. This transformation is done by constructing a collapsed graph as follows:

1. Remove all the arrows between the variables within each causal loop.
2. Arbitrarily number the variables in step 1, and then add an arrow pointing from each lower-numbered variable to the variable with the next-higher number.
3. Add an arrow pointing from each parent of a variable inside the loop to all variables inside the loop.
4. Apply the d-separation criterion in the usual way to the resulting DAG.

A collapsed graph is not unique owing to the arbitrariness in numbering, but all collapsed graphs based on the same DCG imply the same d-separation relations (Spirites, 1995). Illustrated next from left to right is the application of the first three steps beginning from a DCG with a single causal loop (far left):



In a linear model, the sole independence implied by the collapsed graph (far right) is  $X_1 \perp X_2$ . This is the fourth step in Spirites's method for this example.

## Appendix 8.B

### Counterfactual Definitions of Direct and Indirect Effects

We assume that  $X$  is a binary experimental variable ( $0 = \text{control}$ ,  $1 = \text{treatment}$ ),  $M$  is a mediator, and  $Y$  is the outcome, and that  $X$  and  $M$  interact. The controlled direct effect of  $X$  on  $Y$  is

$$\text{CDE} = E [ Y(X = 1, M = m) ] - E [ Y(X = 0, M = m) ] \quad (8.2)$$

where the whole expression is the expected (average) difference on  $Y$  between treated and control cases for a particular value of the mediator,  $m$ , for all cases. The natural direct effect is

$$\text{NDE} = E [ Y(X = 1, M = m_0) ] - E [ Y(X = 0, M = m_0) ] \quad (8.3)$$

which is the expected difference in outcome if the mediator for each case were equal to that in the control group ( $M = m_0$ ). The natural indirect effect is

$$\text{NIE} = E [ Y(X = 1, M = m_1) ] - E [ Y(X = 1, M = m_0) ] \quad (8.4)$$

or the expected outcome among treated cases if the mediator changed from what it would be in the control condition to what it would be in the treatment condition. The total effect is  $\text{TE} = \text{NDE} + \text{NIE}$ , which algebraically simplifies to

$$\text{TE} = E [ Y(X = 1, M = m_1) ] - E [ Y(X = 0, M = m_0) ] = E [ Y(X = 1) ] - E [ Y(X = 0) ] \quad (8.5)$$

which is the expected difference in outcome through the direct and indirect effects of treatment.

## 9

# Specification and Identification of Confirmatory Factor Analysis Models

---

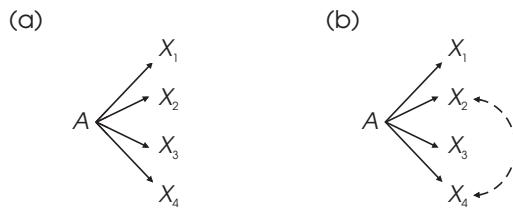
This chapter focuses on the specification of CFA measurement models and their evaluation for identification and describes types of latent variables in CFA. The technique of CFA is contrasted with that of exploratory factor analysis (EFA), which is not part of the SEM family. A key difference between the two techniques is that unrestricted measurement models are analyzed in EFA, but CFA deals with restricted measurement models. The chapter also discusses indicator selection and the representation of hypotheses about measurement dimensionality and directionality in CFA and considers how to determine whether a CFA model is identified. A research example dealt with in more detail in a later chapter is introduced.

---

## LATENT VARIABLES IN CFA

A substantive latent variable in CFA is conceptualized as a single dimension, or continuum, along which cases (or other units of analysis) can vary. It is also an explanatory variable that corresponds to the **local independence** assumptions that (1) one or more latent variables create the association between observed variables, and (2) when the latent variables are held constant, the indicators are independent, if their error variances are independent of both one another and of the latent variables as well (Bollen, 2002). In other words, the indicators are conditionally independent, given the correctly specified latent variable model.

Local independence corresponds to d-separation. For example, consider the graphs in Figure 9.1, which represent latent variable  $A$  as the common cause of observed variables  $X_1-X_4$ . Because latent variables are estimated as factors in CFA, factor  $A$  is allowed to appear in the conditioning set that d-separates  $X_1-X_4$ . Listed next are all the conditional independences implied by Figure 9.1(a) with independent errors:



**FIGURE 9.1.** Measurement models as directed acyclic graphs with independent errors (a) and with an error correlation between a pair of indicators (b).

$$\begin{array}{lll} X_1 \perp X_2 \mid A & X_1 \perp X_3 \mid A & X_1 \perp X_4 \mid A \\ X_2 \perp X_3 \mid A & X_2 \perp X_4 \mid A & X_3 \perp X_4 \mid A \end{array}$$

For continuous variables and assuming linearity, Figure 9.1(a) implies the vanishing partial correlations listed next:

$$\rho_{X_1 X_2 \cdot A} = \rho_{X_1 X_3 \cdot A} = \rho_{X_1 X_4 \cdot A} = \rho_{X_2 X_3 \cdot A} = \rho_{X_2 X_4 \cdot A} = \rho_{X_3 X_4 \cdot A} = 0$$

Exercise 1 asks you to prove that Figure 9.1(b) with correlated errors for a pair of indicators implies fewer conditional independences compared with Figure 9.1(a) with independent errors.

Another view is the **nondeterministic function of observed variables** definition: The equation for a latent variable in a linear model cannot be manipulated so as to express that variable as a function of just the indicators (Bollen, 2002). This is the idea behind **factor indeterminacy**, which states that  $v$  indicators cannot be uniquely transformed into  $v + m$  variables, where  $m$  is the number of factors. So although the results of the analysis might suggest that a particular model fits the data, there are, in theory, infinitely many other models that are just as consistent with the same data. Also, indicators are never perfectly precise (i.e.,  $r_{XX} < 1.0$ ). With infinitely many indicators in an infinitely large sample, factor indeterminacy would be nil, but this goal is not practical; thus, latent variables are inherently estimated with uncertainty.

## FACTOR ANALYSIS

The basic logic and mathematics of factor analysis were developed in the early 1900s in order to test theories about the nature of intelligence (e.g., Spearman, 1904). This makes factor analysis one of the oldest statistical techniques for discovering and describing latent variables, given initially only sample covariances among a set of indicators (Mulaik, 1987). Factor analysis is still widely used today. It is also a primary technique for many researchers, especially those who conduct measurement-related studies. The rationale of factor analysis is also exactly half that of SEM—the other half comes from

regression analysis—so it is important to know about the basics of factor analysis when learning about SEM.

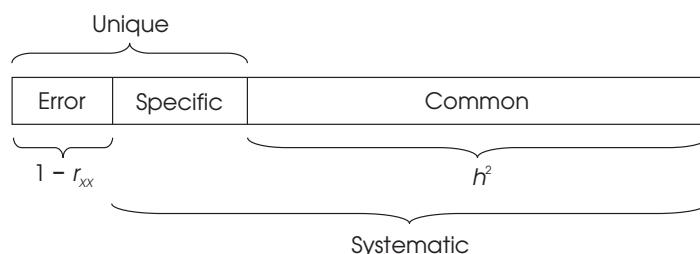
Basically, all factor-analytic methods partition standardized indicator variance in the way shown in Figure 9.2. **Common variance** is shared among the indicators and is a basis for observed covariances among them that depart appreciably from zero. It is generally assumed in factor analysis that (1) common variance is due to the factors and (2) the number of factors of substantive interest is less than the number of indicators. It is impossible to estimate more factors than indicators, but for parsimony's sake, there is no point in retaining a model with just as many explanatory entities (factors) as there are entities to be explained (indicators) (Mulaik, 2009a).

The proportion of total variance that is shared is called **communality**, which is estimated by the statistic  $h^2$ . For example, if  $h^2 = .70$ , then 70% of total indicator variance is common and thus potentially explained by the factors. The rest, or 30%, is **unique variance**, which consists of specific variance and random measurement error. **Specific variance** is systematic variance that is not explained by any factor in the model. It may be due to characteristics of individual indicators, such as the particular stimuli that make up a task. Another source is method variance, or the use of a particular measurement method (e.g., self-report) or informant (e.g., parents) to obtain the scores.

### Kinds of Factor Analysis

There are two broad categories of factor analysis: exploratory (EFA) and confirmatory (CFA). The differences between these two methods are as follows.

1. The method of EFA does not require a priori specification of the number of factors. Without specific instruction to do otherwise, an EFA computer procedure could theoretically generate all possible solutions, from a one-factor model up to a model with as many factors as indicators. In some EFA computer procedures the researcher can optionally request a solution with a specified number of factors (e.g., three), and the computer will analyze just that particular model. But in CFA, the researcher must always specify the exact number of factors.



**FIGURE 9.2.** Basic partition of standardized indicator variance in factor analysis.  $h^2$ , proportion of common variance, or communality;  $r_{xx}$ , score reliability coefficient.

**2.** There is no possibility in EFA to specify the exact correspondence between indicators and factors. This means that indicators are allowed to depend on (theoretically, measure) all factors; thus, it is **unrestricted measurement models** that are analyzed in EFA. But in CFA, each indicator is allowed to depend on only the factor(s) specified by the researcher; that is, **restricted measurement models** are analyzed in CFA.

**3.** Models with multiple factors in EFA are not actually identified because such models have more free parameters than observations. Thus, there is no unique set of statistical estimates for a particular multifactor EFA model. This property concerns the rotation phase in EFA. In contrast, CFA models must be identified before they can be analyzed, so there is only one exclusive set of parameter estimates. Accordingly, CFA has no rotation phase.

**4.** It is generally assumed in EFA that the specific variance of each indicator is not shared with that of any other indicator. But CFA permits, depending on the model, estimation of whether specific variance is shared between certain pairs of indicators (i.e., error correlations).

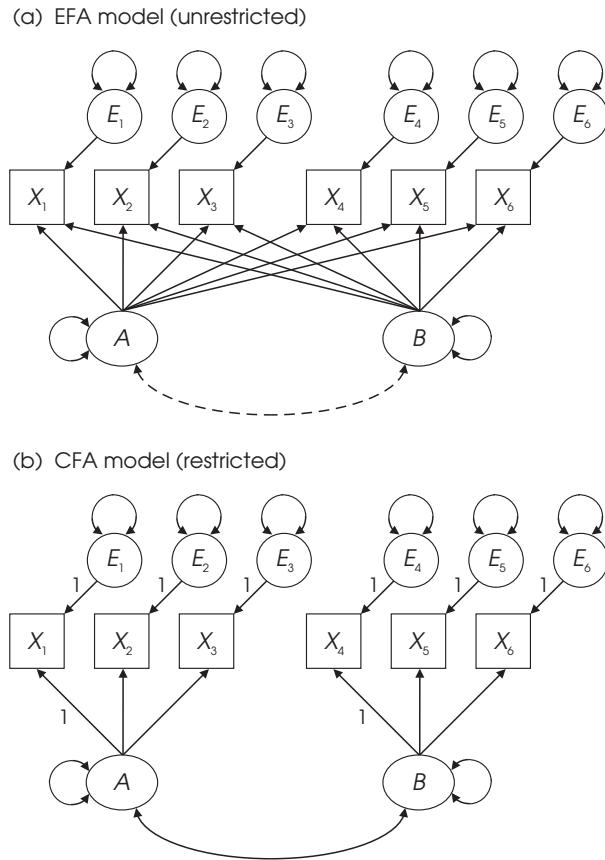
## CHARACTERISTICS OF EFA MODELS

Depicted in Figure 9.3(a) is a measurement model for six continuous indicators and two factors of the kind analyzed in EFA. Linearity is assumed plus a method of factor extraction, such as the principal-axis method, that replaces the 1.0s in the main diagonal of the sample correlation matrix with estimated communalities. Such methods analyze common variance and, consequently, explicitly distinguish between observed and latent variables.<sup>1</sup> This distinction also implies that each indicator has an error term that reflects unique variance (see Figure 9.2).

Figure 9.3(a) is unrestricted in that every indicator is regressed on both factors. The paths that point from factors to indicators represent the direct effects of factors on indicators. The proper name of the statistical estimates of these direct effects is **pattern coefficients**. Many researchers refer to pattern coefficients as *factor loadings* or just *loadings*, but these terms are ambiguous for reasons explained later and thus are not used here. The larger point is that all possible pattern coefficients are estimated for each indicator in EFA. The arc with two arrowheads in the figure represents the possibility to estimate the factor correlation. But because it is not required in EFA to estimate factor correlations, the symbol for an unanalyzed association in the figure is shown as dashed instead of solid. Most EFA computer procedures by default do not analyze correlated factors. Instead, the researcher must typically select a rotation option that permits the factors to covary when the goal is to analyze correlated factors.

---

<sup>1</sup>The principal components method analyzes total indicator variance, not common variance. Thus, it assumes  $r_{XX} = 1.0$  for all indicators, and factors are estimated as linear combinations of the indicators. For this reason, some methodologists do not consider it a “true” method of factor analysis.



**FIGURE 9.3.** An exploratory factor analysis (EFA) model and a confirmatory factor analysis (CFA) model for six indicators and two factors.

The goal of rotation in EFA is to enhance the interpretability of retained factors. It works by reweighting the initial solution (the factor axes are shifted) according to statistical criteria that vary with the particular method. The target outcome is a solution that exhibits **simple structure**, where each factor explains as much variance as possible in nonoverlapping sets of indicators. This means that absolute values of correlations between factors and indicators should shift toward either 0 or 1.0, which makes factor-indicator associations more distinct. There are infinitely many rotations for models with two or more factors, and they all explain the data to the same degree (they are equivalent). This describes **rotational indeterminacy**. In practice, either the researcher specifies a particular rotation or the computer will use its default method. This choice identifies the model, but the estimates are unique only for that particular method.

In **orthogonal rotation**, the factors are all uncorrelated just as they are extracted in the initial solution. The most widely used method is varimax rotation, which is the default in many EFA computer procedures. There are many other orthogonal rotation

methods (quartimax, equamax, etc.), but all such methods assume uncorrelated factors, which is sometimes implausible. For example, it makes little sense to believe that cognitive abilities, such as verbal versus spatial reasoning, would be independent. It would be just as implausible to assume that anxiety and depression are unrelated owing to their high comorbidity in clinical populations. Methods for **oblique rotation** allow correlated factors. Promax rotation is probably the most widely used oblique rotation method, but there are others (e.g., oblimin). It is important to know that specifying oblique rotation does not “force” the factors to covary. Instead, such methods estimate factor correlations, given the model and data, so these estimates are “allowed” to be close to zero, if such estimates are consistent with the data.

There are many other rotation methods, and it can be difficult to decide which method is best. In addition, their use entails some trial and error. For example, two different methods may generate appreciably different results for the same model and data. There may be little basis for preferring one solution over another. But given a robust population measurement model with simple structure assessed with psychometrically sound indicators, similar estimates should be obtained using different rotation methods, but there is no guarantee; see Mulaik (2009a).

When analyzing raw data, some EFA computer procedures can calculate and save estimated factor scores, which estimate relative standing on the factors. Different methods can be used to compute factor scores, including multiple regression applied to estimated correlations between factors and indicators. There is also **factor score indeterminacy**, which means that actually an infinite number of sets of factor scores would all be equally consistent with the same pattern coefficients. It can also happen that a case with a high standing on a particular set of estimated scores could obtain a low ranking on a different set of scores for the same factor. Given such indeterminacy, researchers should probably refrain from making too fine a distinction on estimated factor scores (DiStefano, Zhu, & Măndrila, 2009).

## CHARACTERISTICS OF CFA MODELS

A restricted measurement model for six indicators and two factors of the kind analyzed in CFA is presented in Figure 9.3(b). It represents the hypothesis that  $X_1-X_3$  and  $X_4-X_6$  measure, respectively, factors A and B, which are assumed to covary. This figure is a **standard CFA model** with the following characteristics:

1. Each indicator is continuous with two causes—a single factor that the indicator is supposed to measure and all unique sources of influence represented by the error term.
2. The error terms are independent of each other and of the factors.
3. All associations are linear and the factors covary. This is why the symbol for an unanalyzed association in Figure 9.3(b) is depicted as solid instead of dashed.

Some pattern coefficients are zero in standard CFA models with two or more factors. For example, there is no direct path from factor  $B$  to indicator  $X_1$  in Figure 9.3(b). This specification implies that the corresponding pattern coefficient is zero and, consequently, the computer will not estimate it. But the specification for a zero direct causal effect does *not* say that  $B$  and  $X_1$  are uncorrelated. This is because there is an undirected, back-door path between  $B$  and  $X_1$ , or

$$X_1 \leftarrow A \curvearrowright B$$

This path may convey statistical association but not causation. Although the pattern coefficient for the pair  $B$  and  $X_1$  is zero, the **structure coefficient** for the same pair may not also equal zero. The structure coefficient estimates the Pearson correlation between a factor and a continuous indicator, and it reflects any source of association, causal or noncausal. If the estimated structure coefficient for the pair  $B$  and  $X_1$  is not zero, then all of this association is spurious, given the model. The failure to correctly distinguish between pattern coefficients and structure coefficients is a source of confusion in both EFA and CFA (Graham, Guthrie, & Thompson, 2003).

The numerals (1) in Figure 9.3(b) that appear next to paths from the factors to one of their indicators are scaling constants, or unit loading identification (ULI) constraints. The specifications that

$$A \rightarrow X_1 = 1.0 \quad \text{and} \quad B \rightarrow X_4 = 1.0$$

scale the factors in a metric related to that of the explained (common) variance of the corresponding indicator, or **reference (marker) variable**. Assuming that indicators of the same factor have equally reliable scores, it is arbitrary in single-sample analyses which indicator is selected as the reference variable. This is because the choice does not usually change model fit. Most computer tools for SEM that automatically scale the factors use the reference variable method. Whatever indicator is listed first in the syntax that regresses a set of indicators on its common factor is usually selected by the computer as the reference variable, and its unstandardized pattern coefficient is automatically fixed to equal 1.0. Other options for scaling factors are described later in this chapter. The other scaling constants in Figure 9.3(b), such as

$$E_1 \rightarrow X_1 = 1.0$$

are ULI constraints that assign to error terms a metric related to that of the unexplained variance in the corresponding indicator. There are actually scaling constants in EFA (not shown in Figure 9.3(a)), but EFA computer procedures automatically impose such constraints in the analysis.

Because restricted measurement models are identified through their specification, there is no rotation phase in CFA. There is also little need to work with estimated factor scores because the factors are themselves available in the analysis as either predictors or

outcomes of other variables in the model (Brown, 2006). This possibility describes SR models, not CFA models, but SEM offers much flexibility for testing hypotheses about latent variables.

## OTHER CFA SPECIFICATION ISSUES

Considered next are additional matters of specification. How CFA models are represented in LISREL notation is described in Appendix 9.A.

### Indicator Selection

Selection of the indicators is critical because the quality of the results in factor analysis depends on the quality of the scores analyzed. In summary, Fabrigar and Wegener (2012) make the following suggestions: First, define the hypothetical constructs of interest. If the goal is to delineate dimensions of anxiety, for example, then consult relevant theoretical and empirical works about the nature and number of factors, such as state anxiety, trait anxiety, and social anxiety. Next, identify candidate indicators that as a set adequately sample the various domains. Ideally, not all indicators of the same factor will rely on the same measurement method. This is because common method variance can affect all the scores, and method variance could mistakenly be attributed to the factors. This explains why the standard practice in anxiety studies is to measure physiological variables, such as galvanic skin response, in addition to self-report.

The minimum number of indicators per factor for CFA models with two or more factors is two. But CFA (or SR) models where some factors have only two indicators are more prone to technical problems, such as failure of iterative estimation. It can be difficult to estimate error correlations for factors with only two indicators, which may lead to specification error. A better practical minimum is three to five indicators for each anticipated factor. Kenny's (1979) rule of thumb about the number of indicators is apropos: "Two *might* be fine, three is better, four is best, and anything more is gravy" (p. 143; emphasis in original).

### Dimensionality

The specifications of standard CFA models where (1) each observed variable is a **simple indicator** that depends on a single factor and (2) the errors are independent describe **unidimensional measurement**. It is **multidimensional measurement** that is specified in nonstandard CFA models, which have at least one **complex indicator** that is caused by two or more factors or have at least one error correlation. For example, adding the path  $B \rightarrow X_1$  to Figure 9.3(b) would respecify  $X_1$  as a complex indicator. There is controversy about the specification of complex indicators. On the one hand, some tests may actually measure more than one domain. Suppose that the items of an engineering aptitude test are either text-based or involve the interpretation of data graphics. The test

yields a single total score, which may reflect both verbal reasoning and visual-spatial ability. On the other hand, unidimensional models offer more precise tests of convergent validity. If many indicators depend on the same two factors, the distinctiveness of those factors may be blurred, which can also muddy the evaluation of discriminant validity.

Error correlations can be specified as a way to test hypotheses about shared sources of variation besides the factors. For repeated measures indicators, this specification represents autocorrelated errors. The same specification can also reflect the hypothesis that the two indicators share common method variance. In latent-variable models, omission of theoretically defensible error correlations may not in some cases harm fit, but their absence could change the meaning of those variables and thus lead to inaccurate results (Cole, Ciesla, & Steiger, 2007).

The specification of multidimensional measurement makes a CFA model more complex ( $df_M$  gets smaller, fit improves) relative to standard (unidimensional) model. There are also implications for identification, which are considered later in this chapter. Thus, it is important to evaluate whether a nonstandard CFA model is identified when it is specified and before the data are collected. One way to respecify a nonidentified CFA model is to add indicators, which increases the number of observations available to estimate effects.

## Directionality

Each indicator in a standard CFA model has two unrelated causes, such as

$$A \rightarrow X_1 \leftarrow E_1$$

in Figure 9.3(b). This specification is consistent with the view in classical reliability theory that observed scores are determined by a true (systematic) component and an error component. This also describes **reflective measurement**, where latent variables are assumed to cause observed variables. Observed variables in reflective measurement models are called **effect (reflective) indicators**. Measurement models analyzed in CFA are reflective, and all indicators are effect indicators.

The rationale for reflective measurement comes from the **domain sampling model** (Nunnally & Bernstein, 1994), where effect indicators of the same construct should be internally consistent. Their intercorrelations should therefore be positive and at least moderately high in magnitude. Also, correlations between indicators of the same factor should be greater than cross-factor correlations with indicators that are supposed to measure different factors. These patterns correspond to, respectively, convergent validity and discriminant validity. The domain sampling model also assumes that effect indicators of the same construct with equally precise scores are interchangeable, which means that they can be substituted for one another without appreciably affecting construct measurement.

Sometimes a measure is negatively worded compared with other indicators of the same factor. Suppose that high scores on two indicators of life satisfaction mean greater

contentment, but a third indicator is scaled to reflect the degree of unhappiness, which implies negative correlations with scores from the other two indicators. Negative correlations reduce the reliability of factor measurement (see Equation 4.7 for analyses with observed variables). In this case, the researcher could use **reverse coding** or **reverse scoring** to reflect the scores on the negatively worded measure. That is, multiply the scores by  $-1.0$  and then add a constant to the reflected scores so that the minimum score is at least 1.0. Now high scores on the original measure are reflected as low scores, and vice versa. After reflection, intercorrelations among the three indicators for this example should all be positive.

It makes no sense in reflective measurement to specify a factor with indicators that do not measure something in common. Suppose that gender, ethnicity, and level of education are specified as indicators of a “background” factor. There are two problems here. First, gender and ethnicity are unrelated in representative samples, so one could not claim that these variables somehow measure a common domain. Second, none of these indicators, such as gender, is in any way “caused” by some latent variable that corresponds to a “background” continuum.

The specification that latent variables affect indicators is not always appropriate. For example, **causal indicators** have a conceptual unity, and they are specified to cause the factor, not the other way around (Bollen & Bauldry, 2011). Suppose that income, education, and occupation are used to measure the construct of socioeconomic status (SES), which is usually viewed as a continuum. In a CFA model, these observed variables would be specified as effect indicators. But we usually think of SES as the *outcome* of these indicators (and others), not vice versa. For instance, a change in any one of these indicators, such as an income boost due to winning a lottery, may change SES; that is, it makes more sense to see the indicators as *causes* of SES, not the other way around. This describes **formative measurement**, where indicators are specified as causes of latent variables and where the latent variables have disturbances that reflect unexplained variation. The hypothesis of formative measurement cannot be represented in CFA, but it may be possible to do so when analyzing an SR model where some factors are specified as endogenous (outcomes). Analysis of formative measurement models in SEM is described later in the book.

### **Do Not Reify “Exploratory” versus “Confirmatory”**

You should not overinterpret the labels “exploratory” versus “confirmatory” (i.e., EFA vs. CFA). It is true that EFA requires no a priori hypotheses about factor–indicator correspondence or even the number of factors. There are also more confirmatory modes in EFA, such as instructing the computer to extract a specific number of factors based on theory. The technique of CFA is not strictly confirmatory. It happens in many, if not most, analyses that the initial restricted measurement model does not fit the data. In this case, the researcher typically modifies the hypotheses on which the initial model was based and specifies a revised model. The respecified model is then tested again with the same data. This process should be guided by theory (see Figure 6.1), but relatively few applications of CFA are strictly confirmatory.

### CFA after EFA

You should also know that CFA does not generally “verify” or “confirm” EFA results for the same data and number of factors. Accordingly, it is neither required nor even advisable to conduct CFA as a follow-up analysis to an EFA (if a model is retained in EFA). It can happen that the specification of CFA model based on EFA outcomes and analyzed with the same data will lead to rejection of the CFA model (van Prooijen & van der Kloot, 2001). This is because indicators in EFA often have relatively high secondary pattern coefficients for factors other than the one for which they have their primary pattern coefficients. These secondary coefficients may account for relatively high proportions of variance, so constraining them to zero in CFA may be too conservative. Consequently, the more restrictive CFA model may not fit the data.

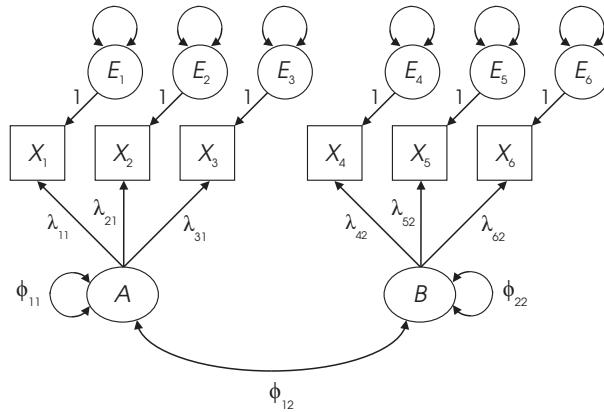
The best way to replicate EFA results is to collect more data and apply the same method in a replication sample. There are procedures for evaluating whether EFA results for the same variables replicate over independent samples (Osborne & Fitzpatrick, 2012). If somehow a researcher in a single-sample analysis gets lucky and finds that the CFA version of an EFA model is retained when fitted to the same data, he or she cannot justifiably claim evidence for replication. This is because (1) there was no replication sample and (2) the two procedures, EFA and CFA, may have capitalized on the same chance variation. This is more likely to happen if the same estimation method, such as maximum likelihood, is used in both analyses. Less mature research areas may not be ready for the more restrictive CFA, and in this case there is nothing wrong with using EFA; that is, CFA is not inherently superior to EFA. Use the right tool for the right job, and CFA is not always the right technique in factor analytic studies.

## IDENTIFICATION OF CFA MODELS

Measurement models analyzed in CFA must meet the same two general requirements for identification as any other type of structural equation model: Every latent variable (including errors) must be scaled, and the model degrees of freedom must be at least zero ( $df_M \geq 0$ ). The number of free parameters and the number of observations are both counted the same way in CFA as in path analysis (respectively, Rules 6.1 and 6.2) when all indicators are continuous. Residual terms are also scaled the same way in both techniques by imposing ULI constraints.

### Scaling Factors

There are three basic options for scaling factors. Use of one method or another does not generally affect model fit for standard CFA models, but Millsap (2001) describes some exceptions for nonstandard models with many complex indicators. Each method is described next with reference to Figure 9.4, which shows parameters for the pattern



**FIGURE 9.4.** Parameters for the pattern coefficients ( $\lambda$ ) and factor variances and covariance ( $\phi$ ) in a standard confirmatory factor analysis model.

coefficients ( $\lambda$ ) and the factor variances and covariance ( $\phi$ ) in LISREL notation (see Appendix 9.A).

The first method, the reference variable method, was described earlier. If  $X_1$  and  $X_4$  are the reference variables for, respectively, factors A and B in Figure 9.4, then we impose a ULI constraint on the unstandardized pattern coefficient of each reference variable, or

$$\lambda_{11} = \lambda_{42} = 1.0 \quad (9.1)$$

The remaining pattern coefficients (4) and the factor variances and covariance (3) are freely estimated (7 free parameters altogether). Because the factors are unstandardized, the term  $\phi_{12}$  for their statistical association is estimated as a covariance.

The second method to scale factors standardizes them by imposing a **unit variance identification (UVI) constraint** on each of their variances. For the model in Figure 9.4,

$$\phi_{11} = \phi_{22} = 1.0 \quad (9.2)$$

which standardizes both factors. In this method, all six pattern coefficients are freely estimated as is the term  $\phi_{12}$  (7 free parameters in total). Because the factors are standardized, the term  $\phi_{12}$  for the unanalyzed association between the factors is estimated as a Pearson correlation. Note that scaling the factors through either ULI constraints (Equation 9.1) or UVI constraints (Equation 9.2) reduces the number of free parameters by one for each factor.

The choice between these two methods is usually based on the relative merits of analyzing factors in unstandardized versus standardized form. When a CFA model is analyzed in a single sample and there are no repeated measures variables, either method is probably fine. Fixing the variance of a latent factor to 1.0 to standardize it has the

advantage of simplicity and does not require the selection of reference variables. A shortcoming of this method, though, is that it is usually applicable only to exogenous factors. Although basically all SEM computer tools allow the imposition of constraints on any free model parameter, the variance of endogenous factors are *not* free parameters. Only some programs, such as LISREL, SEPATH, and RAMONA, allow the *predicted* variances of endogenous factors to be constrained to 1.0. This is not an issue for CFA models, wherein all factors are exogenous, but it can be for SR models, wherein some factors are endogenous.

There are times when standardizing factors is *not* appropriate—for example, (1) the analysis of a structural equation model across independent samples that differ in their variabilities and (2) longitudinal measurement of variables that show changing variances over time. In both cases, important information may be lost when factors are standardized. (How to appropriately scale factors in multiple-samples CFA is considered in a later chapter.)

Little, Slegers, and Card (2006) describe a third method to scale factors in models where (1) the indicators of the same factor all have the same raw score metric and (2) most indicators are specified to measure a single factor (they are simple indicators). Their **effects coding method** does not require the selection of a reference variable, such as when ULI constraints are imposed (Equation 9.1), nor does it standardize factors, such as when UVI constraints are imposed (Equation 9.2). Instead, it relies on the capability of modern SEM computer tools to impose linear constraints on a set of two or more parameter estimates—in this case, the unstandardized pattern coefficients for indicators of the same factor.

The effects coding method works by telling the computer to constrain the *average* pattern coefficient across all indicators of the same factor to equal 1.0 in the unstandardized solution. So scaled, the variance of the factor will be estimated as the average explained variance across all the indicators in their original metric, weighted by the degree to which each indicator contributes to factor measurement. In this way, all indicators contribute to the scale of their common factor. For the indicators of factor A in Figure 9.4, the average pattern coefficient for indicators  $X_1$ – $X_3$  is fixed to equal unity, or

$$\frac{\lambda_{11} + \lambda_{21} + \lambda_{31}}{3} = 1.0 \quad (9.3)$$

which is algebraically equivalent to any of the three expressions listed next:

$$\begin{aligned}\lambda_{11} &= 3 - \lambda_{21} - \lambda_{31} \\ \lambda_{21} &= 3 - \lambda_{11} - \lambda_{31} \\ \lambda_{31} &= 3 - \lambda_{11} - \lambda_{21}\end{aligned} \quad (9.4)$$

The researcher selects any one of the three formulas in Equation 9.4 and then specifies that linear constraint in the syntax of an SEM computer tool. Exercise 2 asks you to derive the corresponding constraints that scale factor B in Figure 9.4 using this method. This exercise assumes that scores on  $X_4$ – $X_6$  are all based on the same metric.

## RULES FOR STANDARD CFA MODELS

Sufficient identification requirements for CFA models are described next. For standard CFA models, some straightforward rules concern the minimum numbers of indicators per factor,<sup>2</sup> as follows in Rule 9.1:

- 
- |  |            |
|--|------------|
| If a standard CFA model  | (Rule 9.1) |
| 1. with a single factor has at least three indicators, or                |            |
| 2. has two or more factors where each factor has two or more indicators, |            |
| then the model is identified.  |            |
- 

That's it. The first part of Rule 9.1 for single-factor models is the **three-indicator rule**, and the second part for models with multiple factors is the **two-indicator rule**. Recall that CFA (and SR models, too) with factors that have only two indicators per factor are prone to technical problems in the analysis, especially in small samples. It is better to have at least three to five indicators per factor to prevent such problems, but two indicators per factor is the required minimum for CFA models with multiple factors.

Let's apply these requirements to the standard CFA models in Figure 9.5. The model in Figure 9.5(a) has a single factor with two indicators. This model is underidentified because it has only two indicators, or one short of the minimum number (3) (Rule 9.1). Exercise 3 asks you to verify that  $df_M = -1$  for this model. The imposition of a constraint, such as one of equality, or

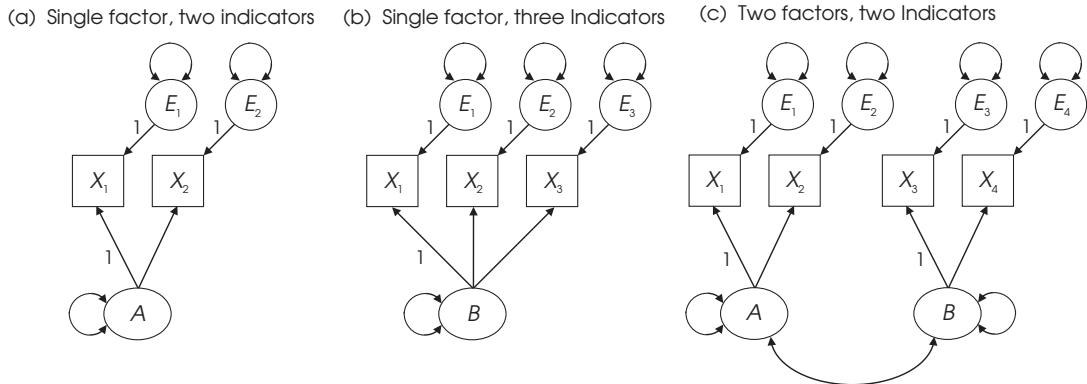
$$A \rightarrow X_1 = A \rightarrow X_2 = 1.0$$

may render this model analyzable because  $df_M$  would be zero in the respecified single-factor, two-indicator model. For such models, Kenny (1979) noted that if the correlation between the two indicators is negative, then the just-identified model that results by imposing an equality constraint on the pattern coefficients does not exactly reproduce the observed correlation. This is an example of a just-identified model that does not perfectly fit the data.

Because the single-factor model in Figure 9.5(b) has three indicators, it is identified; specifically, it is just-identified. Exercise 4 asks you to prove that  $df_M = 0$  for this model. Note that a standard CFA model with a single factor must have at least four indicators in order to be overidentified ( $df_M > 0$ ). Because each of the two factors in Figure 9.5(c) has two indicators, it is also identified (Rule 9.1). Exercise 5 asks you to verify for this model that  $df_M = 1$  (i.e., it is overidentified).

---

<sup>2</sup>Measurement models in CFA with ordinal indicators, such as Likert-scale items, have special identification requirements that are considered later in the book.



**FIGURE 9.5.** Identification status of standard confirmatory factor analysis models.

## RULES FOR NONSTANDARD CFA MODELS

There is a different—and more complicated—set of identification heuristics for non-standard CFA models with complex indicators or error correlations. The potential benefit for dealing with this greater complexity is that the researcher can represent an even wider range of hypotheses about measurement in nonstandard CFA models compared with standard CFA models.

O'Brien (1994) describes a set of rules for nonstandard measurement models where every indicator depends on just one factor but some error correlations are freely estimated. These rules are applied “backwards” starting from patterns of independent (uncorrelated) pairs of error terms to prove the identification of pattern coefficients, then of error variances, next of factor covariances in models with multiple factors, and finally of measurement error correlations. The O'Brien rules work well for relatively simple measurement models, but they can be awkward to apply to more complex models. A different set of identification rules by Kenny, Kashy, and Bolger (1998) that may be easier to apply is listed in Table 9.1 as Rule 9.2. This rule spells out requirements that must be satisfied by each factor (Rule 9.2a), pair of factors (Rule 9.2b), and individual indicator (Rule 9.2c) in order to identify models with error correlations.

Rule 9.2a in Table 9.1 is a requirement for a minimum number of indicators per factor—either two or three depending on the pattern of error correlations or constraints imposed on pattern coefficients. Rule 9.2b refers to the specification that for every pair of factors, there must be at least two indicators, one from each factor, whose error terms are not correlated. Rule 9.2c concerns the requirement for every indicator that there is at least one other indicator in the model with which it does not share an error correlation. Rule 9.2 assumes that all factor covariances are free parameters and that there are multiple indicators of every factor. Kenny et al. (1998) describe additional rules for exceptions to these cases that are not considered here.

**TABLE 9.1. Identification Rule 9.2 for Nonstandard Confirmatory Factor Analysis Models with Correlated Errors**

For a nonstandard CFA model with error correlations to be identified, all three of the conditions listed next must hold: (Rule 9.2)

*For each factor*, at least one of the following must hold: (Rule 9.2a)

1. There are at least three indicators whose errors are uncorrelated with each other.
2. There are at least two indicators whose errors are uncorrelated and either
  - a. the errors of both indicators are not correlated with the error term of a third indicator for a different factor, or
  - b. an equality constraint is imposed on the loadings of the two indicators.

*For every pair of factors*, there are at least two indicators, one from each factor, whose error terms are uncorrelated. (Rule 9.2b)

*For every indicator*, there is at least one other indicator (not necessarily of the same factor) with which its error term is not correlated. (Rule 9.2c)

*Note.* These requirements are described as Conditions B–D in Kenny, Kashy, and Bolger (1998, pp. 253–254).

Kenny et al. (1998) also describe identification heuristics for complex indicators that depend on two or more factors. The first requirement is listed in the top part of Table 9.2 as Rule 9.3, and it concerns sufficient requirements for the identification of the multiple pattern coefficients of complex indicators. Basically, this rule requires that each factor on which a complex indicator depends has a sufficient number of indicators (i.e., each factor meets Rule 9.2a in Table 9.1). Rule 9.3 also requires that each one of every pair of such factors has an indicator that does not share an error correlation with a corresponding indicator of the other factor (see Table 9.2). If a complex indicator shares error correlations with other indicators, then the additional requirements listed as Rule 9.4 in Table 9.2 must also be satisfied. This rule requires that for each factor on which a complex indicator depends, at least one simple indicator does not share an error correlation with the complex indicator. The requirements of Rules 9.3 and 9.4 are usually addressed by specifying that some indicators depend on just a single factor (i.e., the model has a sufficient number of simple indicators).

Let's apply the identification heuristics just discussed to the nonstandard CFA models in Figure 9.6. To save space, I use compact symbolism in the figure where indicators are designated as  $X$  and factors are represented as  $A$ ,  $B$ , or  $C$ . But do not forget the variance parameter associated with each exogenous variable in the figure that is normally represented by the  $\Omega$  symbol in diagrams elsewhere in this book. Scaling constants are also not shown in the figure, but they are assumed. The single-factor, four-indicator model in Figure 9.6(a) has two error correlations, or

$$E_2 \curvearrowright E_4 \quad \text{and} \quad E_3 \curvearrowright E_4$$

**TABLE 9.2. Identification Rule 9.3 for Pattern Coefficients of Complex Indicators in Nonstandard Confirmatory Factor Analysis Models and Rule 9.4 for Error Correlations of Complex Indicators**Pattern coefficients

For every complex indicator in a nonstandard CFA model:

(Rule 9.3)

In order for the multiple pattern coefficients to be identified, both of the following must hold:

1. *Each factor on which the complex indicator depends* must satisfy Rule 9.2a for a minimum number of indicators.
2. *Every pair of those factors* must satisfy Rule 9.2b that each factor has an indicator that does not have an error correlation with a corresponding indicator on the other factor of that pair.

Error correlations

In order for *error correlations* that involve complex indicators to be identified, both of the following must hold: (Rule 9.4)

1. Rule 9.3 is satisfied.
2. For each factor on which a complex indicator depends, there must be at least one indicator with a single loading that does not have an error correlation with the complex indicator.

*Note.* These requirements are described as Condition E in Kenny, Kashy, and Bolger (1998, p. 254).

This model is just-identified because it has zero degrees of freedom ( $df_M = 0$ ), its factor (A) has at least three indicators whose error terms are uncorrelated ( $X_1-X_3$ ) (Rule 9.2a), and all other requirements of Rule 9.2 (Table 9.1) are met.

The single-factor, four-indicator model in Figure 9.6(b) also has two error correlations, but in a different pattern, or

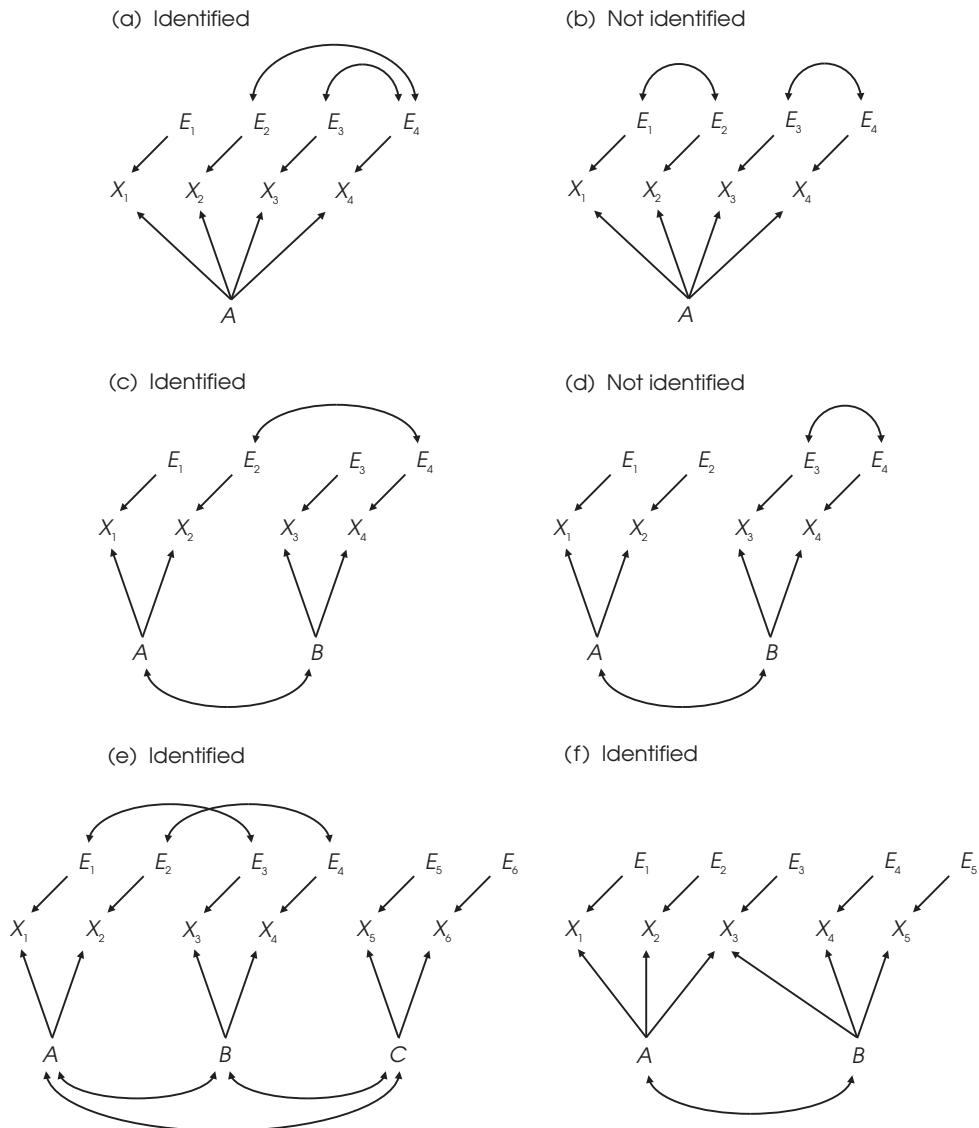
$$E_1 \curvearrowright E_2 \quad \text{and} \quad E_3 \curvearrowright E_4$$

Although this model has at least two indicators whose error terms are independent, such as  $X_2$  and  $X_3$ , it nevertheless fails Rule 9.2a because factor A does not have three indicators whose errors are independent of each other. There are also no other factors in the model, so the alternative requirement in Rule 9.2 that factor A have at least two indicators whose errors are unrelated and the errors of both those indicators are not correlated with the error term of a different factor (Table 9.1) does not apply; therefore, Figure 9.6(b) is not identified. But this model would be identified if an equality constraint were imposed on the pattern coefficients of  $X_2$  and  $X_3$ . That is, the specification that

$$A \rightarrow X_2 = A \rightarrow X_3$$

would be sufficient to identify the model in Figure 9.6(b) because then Rule 9.2 would be met.

The two-factor, four-indicator model in Figure 9.6(c) with a single error correlation,  $E_2 \curvearrowleft E_4$ , is just-identified because  $df_M = 0$  and all three requirements for Rule 9.2 are satisfied (Table 9.1). But the two-factor, four-indicator model in Figure 9.6(d) with a different error correlation,  $E_3 \curvearrowleft E_4$ , is not identified because it violates Rule 9.2a. Specifically, factor B in this model does not have two indicators whose errors are independent.



**FIGURE 9.6.** Identification status of nonstandard confirmatory factor analysis models.

It is generally easier to identify cross-factor error correlations (e.g., Figure 9.6(c)) than within-factor error correlations (e.g., Figure 9.6(d)) when there are only two indicators per factor without imposing additional constraints.

The three-factor, six-indicator model in Figure 9.6(e) with two cross-factor error correlations, or

$$E_1 \curvearrowleft E_3 \quad \text{and} \quad E_2 \curvearrowleft E_4$$

is overidentified because the degrees of freedom are positive ( $df_M = 4$ ) and Rule 9.2 is satisfied. This model also demonstrates that adding indicators—along with a third factor—identifies additional error correlations compared with the two-factor model in Figure 9.6(c). The model in Figure 9.6(f) has a complex indicator that depends on two factors, or

$$A \rightarrow X_3 \quad \text{and} \quad B \rightarrow X_3$$

Because this model meets the requirements of Rule 9.3 (see Table 9.2) and has positive degrees of freedom ( $df_M = 3$ ), it is overidentified. Exercise 6 asks you to add an error correlation to this model and then evaluate Rule 9.4 in order to determine whether the respecified model is identified or not identified.

## EMPIRICAL UNDERIDENTIFICATION IN CFA

The phenomenon of empirical underidentification can affect the analysis of CFA models—and SR models, too—that are actually identified. Suppose that the estimated pattern coefficient for the path  $A \rightarrow X_2$  in the single-factor, three-indicator model of Figure 9.5(b) is close to zero. Practically speaking, this model would resemble the one in Figure 9.5(a) in that factor A has only two indicators, which is too few for a standard single-factor CFA model. The two-factor model of Figure 9.5(c) may be empirically underidentified if the estimated covariance (or correlation) between factors A and B is close to zero. The virtual elimination of the path  $A \curvearrowleft B$  from this model transforms it into two single-factor, two-indicator models, each of which is underidentified. The non-standard model in Figure 9.6(f) where  $X_3$  depends on both factors may be empirically underidentified if the absolute estimate of the factor correlation is close to 1.0. Specifically, this extreme collinearity, but now between factors instead of observed variables, can complicate estimation of the pattern coefficients for the complex indicator  $X_3$ .

## CFA RESEARCH EXAMPLE

The first edition of the Kaufman Assessment Battery for Children (KABC-I; Kaufman & Kaufman, 1983) is an individually administered cognitive ability test for children ages

2½ to 12½ years. The test's authors claimed that the KABC-I's eight subtests represented in Figure 9.7 measure two factors. The three tasks believed to reflect sequential processing all require the correct recall of auditory stimuli (Number Recall, Word Order) or visual stimuli (Hand Movements) in a particular order. The other five tasks—Gestalt Closure, Triangles, Spatial Memory, Matrix Analogies, and Photo Series—are supposed to measure more holistic, less order-dependent reasoning, or simultaneous processing. The data for this example come from the test's standardization sample for 10-year-old children ( $N = 200$ ).

Results of some CFA analyses of the KABC-I conducted in the 1980s–1990s generally supported the two-factor model presented in Figure 9.7 (Cameron et al., 1997). But other results indicated that some subtests, especially Hand Movements, may measure both factors and that some of the error terms may covary (Keith, 1985). Exercise 7 asks you to prove that the model in Figure 9.7 is overidentified with  $df_M = 19$ . Detailed analysis of the CFA model just described is considered in Chapter 13.

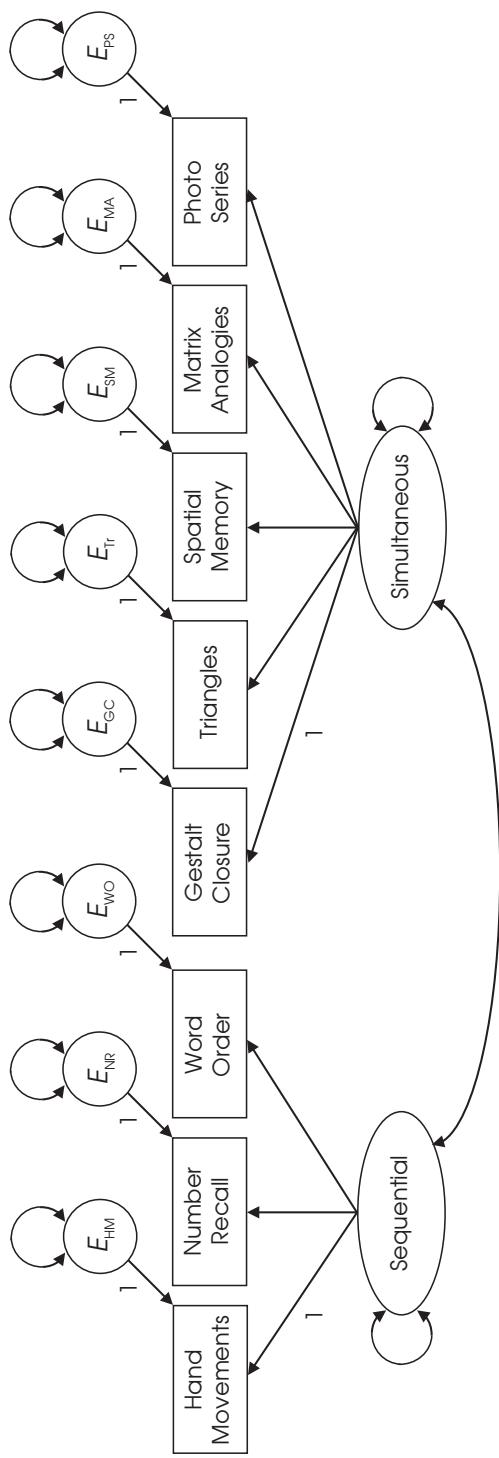
## SUMMARY

The technique of CFA analyzes restricted measurement models, where the researcher must specify in advance the number of factors, the correspondence between factors and indicators, and pattern of error correlations (if any). These models assume reflective measurement, where the factors cause the indicators, not the reverse. Standard CFA models feature continuous indicators that each depend on just a single factor with independent errors. This combination specifies unidimensional measurement, and the evaluation of standard models with two or more factors tests hypotheses of convergent validity and discriminant validity. In nonstandard CFA models, some indicators depend on two or more factors or share error variance. It is more difficult to determine whether a nonstandard model is identified, but there are heuristics for certain types of nonstandard models. Empirical underidentification is more likely to happen with CFA models where some factors have just two indicators and the sample size is not large. The next chapter deals with structural regression models where some factors are endogenous (they are outcomes).

## LEARN MORE

Bollen and Hoyle (2012) describe the specification of latent variable models in SEM. A non-technical introduction to EFA is available in Fabrigar and Wegener (2012), and Kline (2013b) elaborates on similarities and differences between EFA and CFA.

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



**FIGURE 9.7.** A confirmatory factor analysis model of the first-edition Kaufman Assessment Battery for Children.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York: Oxford University Press.

Kline, R. B. (2013b). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 171–207). New York: Routledge.

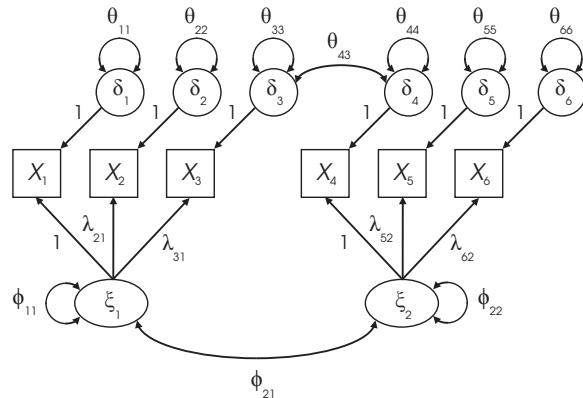
## EXERCISES

1. Show that Figure 9.1(b) implies fewer conditional independences than Figure 9.1(a).
2. Show how to scale factor  $B$  in Figure 9.4 using the effects coding method.
3. Verify that  $df_M = -1$  for Figure 9.5(a).
4. Show that  $df_M = 0$  for Figure 9.5(b).
5. Prove that  $df_M = 1$  for Figure 9.5(c).
6. If  $E_3 \rightsquigarrow E_5$  were added to Figure 9.6(f), would the resulting respecified model be identified?
7. Determine whether the model in Figure 9.7 is identified and show that  $df_M = 19$ .

## Appendix 9.A

### LISREL Notation for CFA Models

Described next for CFA models when means are not analyzed is **LISREL all-X notation**, where the symbol  $X$  represents the indicators of exogenous factors. Lowercase Greek letters in this notation include  $\delta$  (delta),  $\theta$  (theta),  $\lambda$  (lambda),  $\xi$  (xi), and  $\phi$  (phi); uppercase letters include  $\Theta$  (theta),  $\Lambda$  (lambda), and  $\Phi$  (phi). Symbols for variables, parameters, and error terms appear in their proper places in the CFA model presented next:



Measurement equations for the indicators are shown next:

$$\begin{aligned} X_1 &= \xi_1 + \delta_1 & X_4 &= \xi_2 + \delta_4 \\ X_2 &= \lambda_{21} \xi_1 + \delta_2 & X_5 &= \lambda_{52} \xi_2 + \delta_5 \\ X_3 &= \lambda_{31} \xi_1 + \delta_3 & X_6 &= \lambda_{62} \xi_2 + \delta_6 \end{aligned} \quad (9.7)$$

The measurement equations can be expressed in matrix algebra terms as follows:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & 1 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \end{bmatrix} = \boldsymbol{\Lambda}_X \boldsymbol{\xi} + \boldsymbol{\delta} \quad (9.8)$$

where  $\Lambda_X$  (lambda-X) is the parameter matrix for the pattern coefficients,  $\xi$  is the matrix of the exogenous factors, and  $\delta$  is the matrix of error terms for the indicators. The other parameter matrices are

$$\Phi = \begin{bmatrix} \phi_{11} & \\ \phi_{21} & \phi_{22} \end{bmatrix} \quad \text{and} \quad \Theta_\delta = \begin{bmatrix} \theta_{11} & & & & & \\ 0 & \theta_{22} & & & & \\ 0 & 0 & \theta_{33} & & & \\ 0 & 0 & \theta_{43} & \theta_{44} & & \\ 0 & 0 & 0 & 0 & \theta_{55} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix} \quad (9.9)$$

where  $\Phi$  is the covariance matrix of the factors and  $\Theta_\delta$  (theta-delta) is the covariance matrix of the errors. Thus, LISREL parameters matrices for CFA models in all-X notation include

$$\Lambda_X, \Phi, \text{ and } \Theta_\delta$$

## 10

# Specification and Identification of Structural Regression Models

---

The most general kind of model in SEM is an SR model, also called a full LISREL model. This term reflects the fact that LISREL was one of the first computer programs to analyze SR models, but nowadays any modern SEM computer tool can do so. The structural part of an SR model represents hypotheses about direct or indirect effects among observed or latent variables, and the measurement part represents the correspondence between latent variables and their indicators. The capability to test hypotheses about both structural and measurement relations within a single model affords much flexibility. This chapter outlines the specification of SR models with continuous indicators and requirements for their identification. Research examples considered in more detail later in the book are also introduced.

---

## CAUSAL INFERENCE WITH LATENT VARIABLES

In contrast to CFA models where all factors are exogenous and assumed to simply covary, causal effects between factors are represented in SR models. But causal inference in latent variable modeling is potentially more difficult compared with the analysis of path models, where each substantive domain is measured with a single indicator. One reason is factor indeterminacy: Just as in CFA models, factors in SR models are theoretical variables that are measured only indirectly through their indicators. Because theoretical variables and their proxies (indicators) are almost never identical, estimates of causal relations between latent variables are approximate at best. In other words, factor indeterminacy limits predictive utility, blurring estimates of correlations between a factor and external variables (Rigdon, 2014). This is because such correlations are not unique and can be estimated only within a particular range (Steiger & Schönemann, 1978). These issues argue for caution against overinterpretation of results from analyses of SR models.

## TYPES OF SR MODELS

Presented in Figure 10.1(a) is a traditional path model. Exogenous variable  $X_1$  is assumed to be measured without error, an assumption usually violated in practice. This assumption is not required for the endogenous variables in this model, but random error in  $Y_1$  or  $Y_3$  is manifested in their disturbances. Figure 10.1(b) is an SR model with both structural and measurement components. Its measurement model has the same three manifest variables represented in the path model,  $X_1$ ,  $Y_1$ , and  $Y_3$ . Unlike the path model, though, each of these three indicators in the SR model is specified as one of a pair for a latent variable.<sup>1</sup> Consequently, all observed variables in Figure 10.1(b) have error terms.

The structural model of Figure 10.1(b) represents the same basic pattern of direct and indirect causal effects as the path model of Figure 10.1(a) but among latent variables, or

$$A \rightarrow B \rightarrow C$$

The structural model just listed is recursive, but it is also generally possible to specify an SR model with a nonrecursive structural component. Each endogenous factor in Figure 10.1(b) has a disturbance ( $D_B$ ,  $D_C$ ). Unlike in path models, disturbances for endogenous factors in Figure 10.1(b) reflect only omitted causes and not also measurement error in that factor's indicators. For the same reason, estimates of the coefficients for the paths

$$A \rightarrow B \quad \text{and} \quad B \rightarrow C$$

in Figure 10.1(b) are adjusted for measurement error, but those for the paths

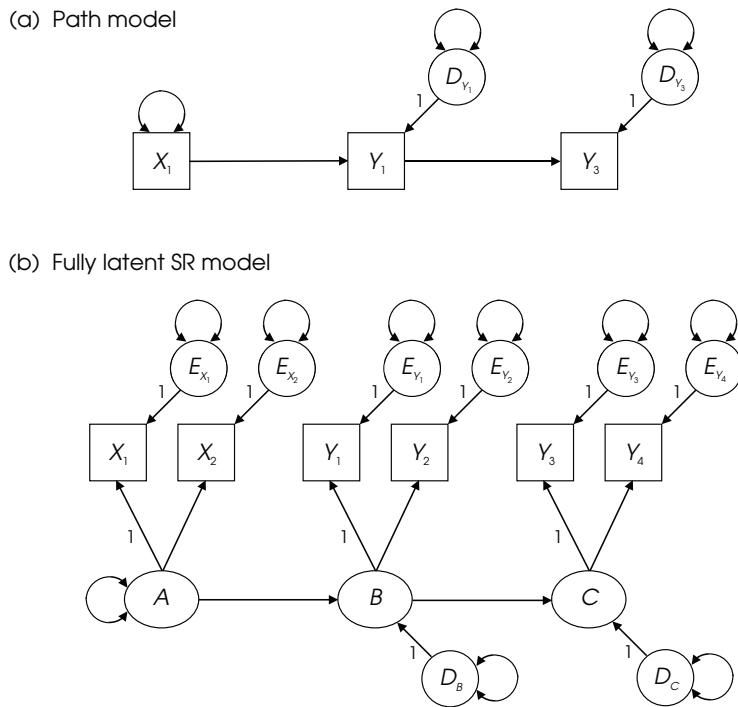
$$X_1 \rightarrow Y_1 \quad \text{and} \quad Y_1 \rightarrow Y_3$$

in Figure 10.1(a) are not. Exercise 1 asks you to calculate the degrees of freedom ( $df_M$ ) for the SR model in Figure 10.1(b). Observations and free parameters are counted for SR models in the same ways as they are for path models and CFA models (see Rules 6.1 and 6.2). Described in Appendix 10.A is LISREL notation for SR models.

Figure 10.1(b) could be described as a **fully latent SR model** because every variable in its structural part is latent with multiple indicators. It is also possible to represent single-indicator measurement in SR models. This reflects the reality that sometimes there is just a single measure of some domain of interest. It also happens that the researcher collects data on multiple indicators but later finds that some of those indicators have poor psychometrics, so their scores are not further analyzed. Models with single indicators could be called **partially latent SR models** because at least one variable

---

<sup>1</sup>I saved space in Figures 10.1–10.4 by showing only two indicators per factor, but remember that having so few indicators may cause technical problems in the analysis.



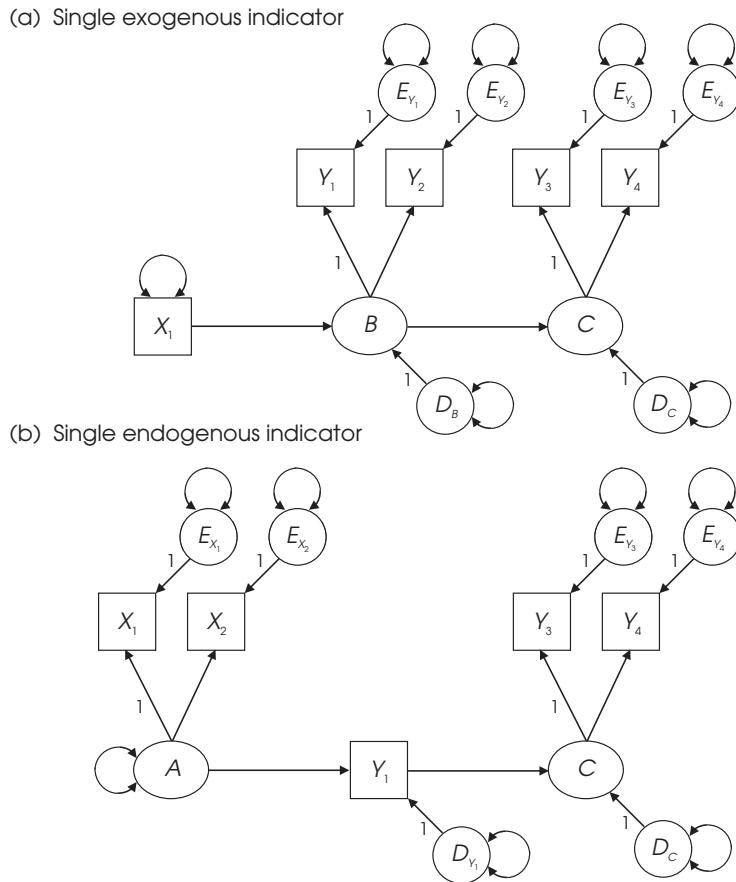
**FIGURE 10.1.** Examples of a path analysis model (a) and a corresponding fully latent structural regression model (b).

in their structural part is a single indicator. Two examples are presented in Figure 10.2. Variable  $X_1$  in Figure 10.2(a) is a single indicator. Because  $X_1$  is specified as exogenous, it is assumed to have no measurement error. Variable  $Y_1$  in Figure 10.2(b) is also a single indicator, but it is specified as endogenous; thus, scores on  $Y_1$  are *not* assumed to be perfectly reliable, but measurement error is confounded with omitted causes of  $Y_1$ .

## SINGLE INDICATORS

There is an alternative to representing a single indicator in the structural part of an SR model as one would in path analysis. It requires an a priori estimate of the proportion of variance in a single indicator that is due to measurement error (.10, .20, etc.). This estimate may be based on the researcher's experience or on results of prior empirical studies. Recall that one minus a reliability coefficient,  $1 - r_{XX}$ , estimates the proportion of total variance due to error. Because a particular reliability coefficient may estimate only one kind of error, the quantity  $1 - r_{XX}$  may *underestimate* the extent of measurement error.

Suppose that  $X_1$  is the only measure of an exogenous construct  $A$ . Given  $r_{XX} = .80$ , we can say that at least  $1 - .80 = .20$ , or 20% of  $X_1$ 's total variance is due to random error. Now we can specify an SR model like the one in Figure 10.3(a). Note in the figure that

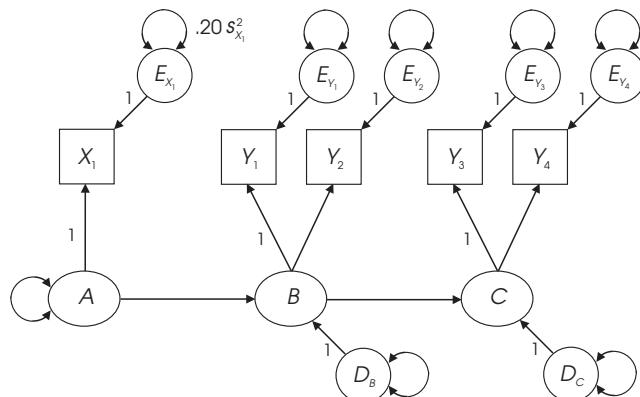
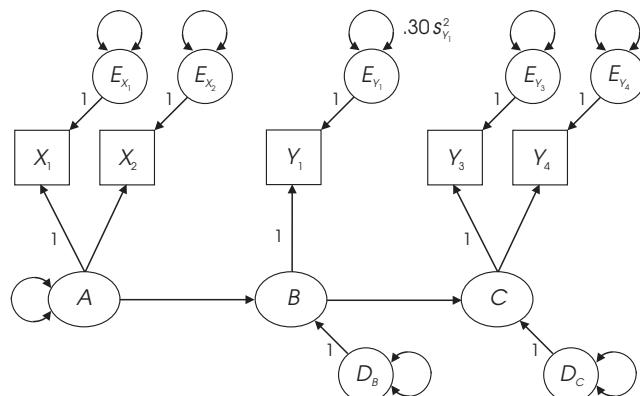


**FIGURE 10.2.** Examples of partially latent structural regression models with a single indicator of an exogenous construct (a) and an endogenous construct (b).

$X_1$  is specified as the single indicator of factor  $A$  and has an error term. The unstandardized error variance is specified as a fixed parameter that equals .20 times the observed variance, or  $.20s_{X_1}^2$ . For example, if the observed variance of  $X_1$  is 30.00, then 20% of this value, or  $.20(30.00) = 6.00$ , is specified as the error variance. Because factor  $A$  must be scaled, the unstandardized pattern coefficient for  $X_1$  in Figure 10.3(a) is fixed to equal 1.0. With the specification of an error term for  $X_1$ , the direct effect of factor  $A$  and the disturbance variance of factor  $B$  are both estimated controlling for measurement error in the single indicator.

Now look at Figure 10.3(b), in which  $Y_1$  is specified as the single indicator of endogenous factor  $B$ . Given  $r_{YY} = .70$ , the proportion of total variation in  $Y_1$  due to measurement error is estimated to be .30. This means that the variance of the error term for  $Y_1$  is fixed to equal .30 times the observed variance of  $Y_1$ . Because  $Y_1$  has an error term, the direct effects of factors  $A$  and  $B$  and the disturbance variance for factor  $C$  are all estimated controlling for measurement error in the single indicator. Four points should be noted about this method for single indicators:

1. It does not affect the complexity of the model (i.e.,  $df_M$  is not changed). Exercise 2 asks you to verify this fact. Model fit is also unchanged.
2. A common question is, why not just specify that the error variance for a single indicator as a free parameter and let the computer estimate it? Such a specification may result in an identification problem (Bollen, 1989, pp. 172–175). It is safer to fix the error variance to a constant based on a prior estimate.
3. A related question is, what if the researcher is uncertain about the estimate of error variance for a single indicator? The model can be analyzed with a range of estimates, which allows evaluation of the impact of different assumptions about measurement error on the solution.
4. A path model can be respecified in order to control for measurement error in every single indicator. This tactic is akin to fitting a path model to a data matrix

(a)  $r_{xx} = .80$  for  $X_1$ (b)  $r_{yy} = .70$  for  $Y_1$ **FIGURE 10.3.** Two structural regression models with single indicators that correct for measurement error. It is assumed that the proportion of error variance is .20 for  $X_1$  and .30 for  $Y_1$ .

based on correlations disattenuated for unreliability (Equation 4.9). Exercise 3 asks you to apply this method to Figure 10.1(a).

Hayduk and Littvay (2012) recommend the single-indicator specification for demographic variables because demographics are sometimes measured with error (e.g., a participant reports the wrong age). Specifying a small, nonzero error variance, such as .05, or 5% of the total variance is safer than assuming that demographics are perfectly measured. The same authors also remind us that multiple-indicator measurement is not always better than single-indicator measurement. For example, given a choice between a single indicator with good psychometrics and strong theoretical connection to the target construct versus a set of multiple indicators that are more or less thrown together with little regard for theory, the single indicator is preferred. Among multiple indicators, there may be a **best indicator** with the greatest relevance to theory. If so, then fixing the error variance of that best indicator to a constant using the method just described may improve factor measurement compared with freely estimating the error variances of all indicators. This is because freely-estimated error variances or covariances can become a “fudge factor” that absorbs different types of potential specification errors.

## IDENTIFICATION OF SR MODELS

If one understands something about the identification of path models and CFA models, there is relatively little new to learn about SR models. This is because the evaluation of whether SR models are identified is conducted separately for each part of the model, measurement and structural. A theme of this evaluation is that a valid (i.e., identified) measurement model would be needed before it would make sense to assess the structural part of an SR model.

As with CFA models, meeting the two necessary requirements— $df_M \geq 0$  and every latent variable is scaled—does not guarantee the identification of SR models. Additional requirements reflect the view that the analysis of a fully latent SR model is essentially a path analysis conducted with estimated variances and covariances among the factors. Thus, it must be possible for the computer to derive unique estimates of factor variances and covariances before specific direct effects among them can be estimated. Bollen (1989) describes this requirement as the **two-step identification rule**, and the steps to evaluate it are outlined in Rule 10.1.

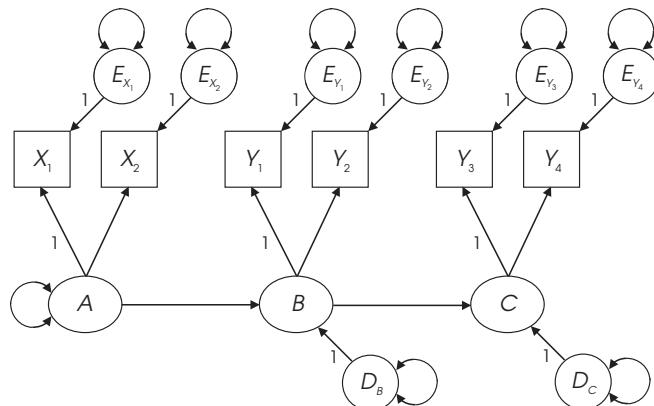
---

A fully latent SR model is identified if the (Rule 10.1)

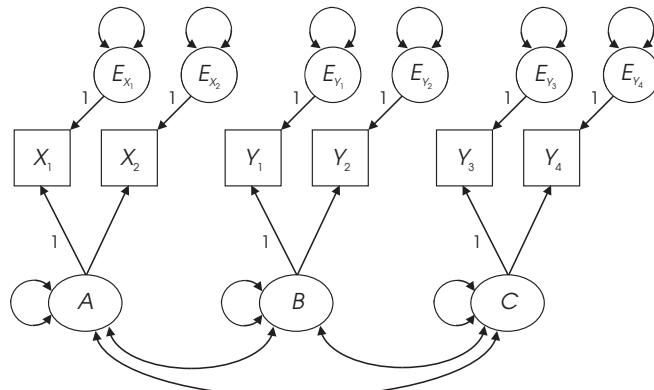
1. measurement part respecified as a CFA model is identified (evaluate the CFA model against Rules 9.1–9.4); and the
  2. structural part is identified (evaluate the structural model against Rules 7.1–7.3 or apply the graphical rules in Figure 7.3).
-

The two-step rule is a sufficient condition: Fully latent SR models that satisfy both parts of Rule 10.1 are identified. Evaluation of the two-step rule is demonstrated next for Figure 10.4(a). This model meets the necessary requirements because every latent variable (including the errors) is scaled and there are more observations than free parameters. (You should verify this statement.) But we still do not know whether Figure 10.4(a) is identified. To find out, we apply the two-step rule. The specification of this fully latent SR model as a CFA measurement model is presented in Figure 10.4(b). Because this standard CFA model has at least two indicators per factor, it is identified. The first part

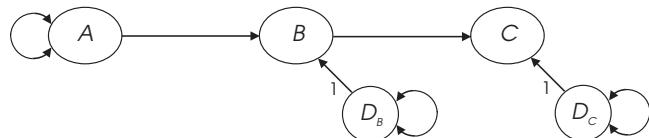
(a) Original SR model



(b) Respecified as CFA model



(c) Structural model



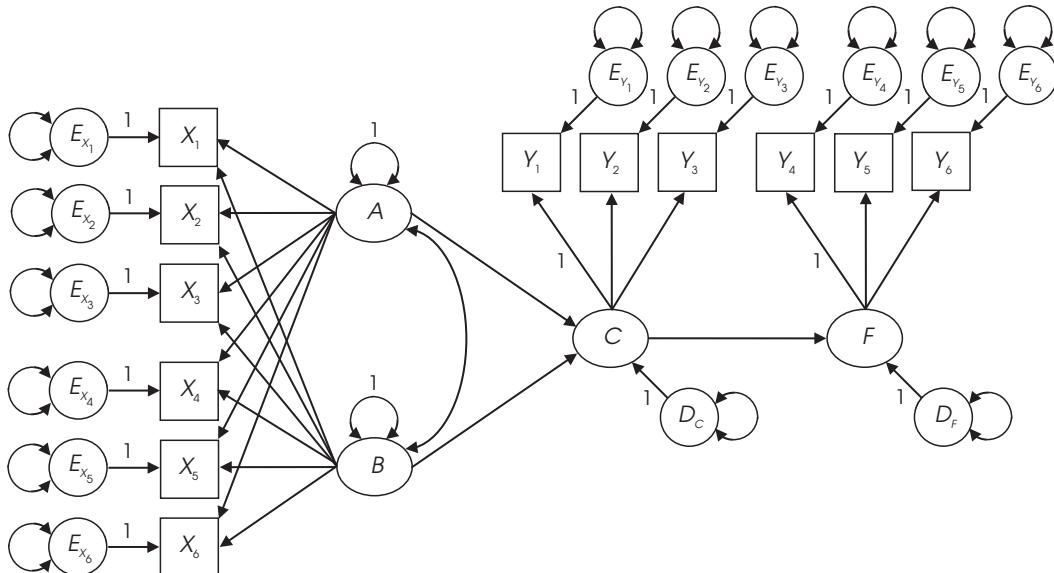
**FIGURE 10.4.** Evaluation of the two-step rule for identification of a fully latent structural regression model.

of the two-step rule is satisfied. The structural part of the original SR model is presented in Figure 10.4(c). Because the structural model viewed as a path model is recursive, it, too, is identified. Because the original SR model in Figure 10.4(a) meets both parts of the two-step rule (Rule 10.1), it is identified—specifically, overidentified.

The two-step rule does not apply to partially latent SR models with single indicators. Such models will always fail the first part of Rule 10.1, which requires at least two indicators per factor for CFA models with multiple factors. For example, if either model in Figure 10.3 is respecified as a CFA model, one factor (*A* or *B*) will have only one indicator, which is one less than the minimum required number (2). Fixing the error variance of  $X_1$  in Figure 10.3(a) or  $Y_1$  in Figure 10.3(b) to a constant along with setting a scale for the corresponding factor, however, identifies the measurement model. The structural parts of both Figures 10.3(a) and 10.3(b) are recursive, which means that they are identified. Because both the measurement and structural parts of Figures 10.3(a) and 10.3(b) are actually identified, the original SR models are identified, too.

## EXPLORATORY SEM

Special types of SR models are analyzed in **exploratory structural equation modeling** (ESEM). Some part of the measurement model in ESEM is unrestricted in that indicators depend on all factors, just as in EFA. But other parts of the measurement model are



**FIGURE 10.5.** An exploratory structural equation model with an unrestricted measurement component (for indicators of factors *A* and *B*) and a restricted measurement component (for indicators of factors *C* and *F*).

restricted in that indicators depend only on factors specified by the researcher, just as in CFA. This type of analysis may be suitable when the researcher has weaker hypotheses about measurement for some constructs than is ordinarily allowed in SEM. Consider the ESEM model in Figure 10.5. The measurement model for  $X_1$ - $X_6$  is unrestricted in that all these variables are specified to depend on both factors, A and B. In the Mplus program, the factor solution for this part of the model is rotated according to the method requested by the user. Factors A and B are scaled by fixing their variances to 1.0, which standardizes them. In contrast, the measurement model for  $Y_1$ - $Y_6$  is restricted in that each indicator depends on a single factor. There is a structural model in Figure 10.5, too, and it features direct or indirect effects among factors A, B, C, and F. Marsh, Morin, Parker, and Kaur (2014) describe applications of ESEM in clinical psychology research.

## **SR MODEL RESEARCH EXAMPLES**

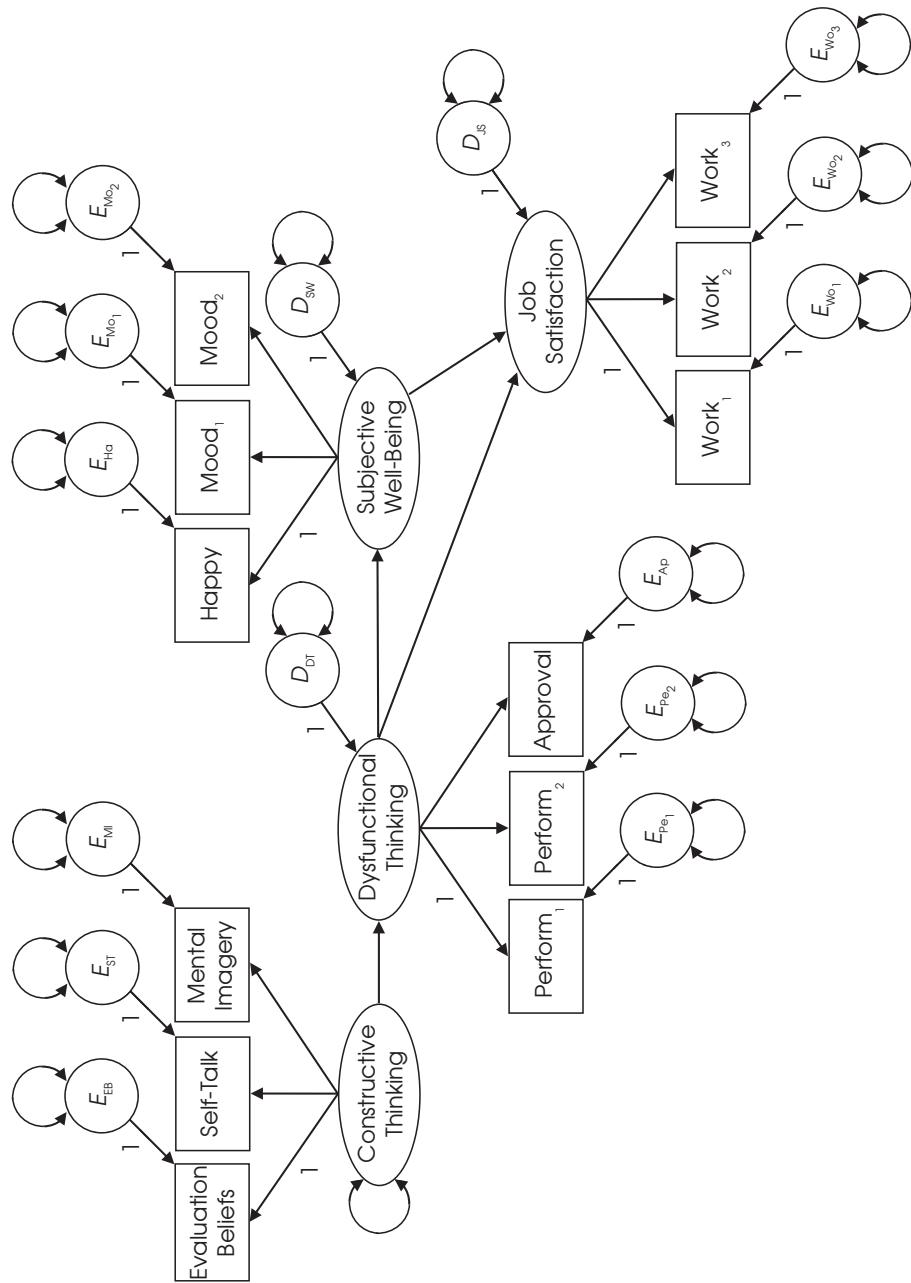
Introduced next are two research examples in which the hypotheses are represented in SR models. The analysis of both models is described in Chapter 14.

### **Fully Latent SR Model of Job Satisfaction Factors**

Within a sample of 263 full-time university employees, Houghton and Jinkerson (2007) administered measures of four different theoretical domains: constructive (opportunity-oriented) thinking, dysfunctional (obstacle-oriented) thinking, subjective well-being, and job satisfaction. Based on their review of relevant theory and empirical results, Houghton and Jinkerson (2007) specified the 4-factor, 12-indicator fully latent SR model presented in Figure 10.6. The structural part of the model represents the hypotheses that (1) dysfunctional thinking and subjective well-being each have direct effects on job satisfaction; (2) constructive thinking has a direct effect on dysfunctional thinking; (3) constructive thinking indirectly affects subjective well-being through the intermediary variable of dysfunctional thinking; and (4) constructive thinking affects job satisfaction indirectly through the other two factors (see the figure). Because there is no time precedence in this design, the indirect effects are not described as mediation.

The measurement part of the SR model in Figure 10.6 features three indicators per factor. Briefly, indicators of (1) constructive thinking include measures of belief evaluation, positive self-talk, and positive visual imagery; (2) dysfunctional thinking includes two scales regarding worry about performance evaluations and a third scale about need for approval; (3) subjective well-being include ratings about general happiness and two positive mood rating scales; and (4) job satisfaction include three scales that reflect one's work experience as positively engaging. Exercise 4 asks you to verify for this model that  $df_M = 50$ .

The article by Houghton and Jinkerson (2007) is exemplary in that the authors describe the theoretical rationale for each and every direct effect among the four factors in the structural model, give detailed descriptions of all measures including internal



**FIGURE 10.6.** A fully latent structural regression model of thought strategies and job satisfaction.

consistency score reliabilities, report the correlations and standard deviations for the covariance matrix they analyzed, and tested alternative models. But the authors did not report unstandardized parameter estimates, nor did they consider equivalent versions of their final model.

### **Single Indicators in a Nonrecursive Model of Organizational and Occupational Turnover Intention**

Within a sample of 177 nurses out of 30 similar hospitals in Taiwan, Chang, Chi, and Miao (2007) administered measures of occupational commitment (i.e., to the nursing profession) and organizational commitment (i.e., to the hospital that employs the nurse). Each commitment measure consisted of three scales: affective (degree of emotional attachment), continuance (perceived cost of leaving), and normative (feeling of obligation to stay). The affective, continuance, and normative aspects belong to a three-part theoretical and empirical model.

Results of studies reviewed by Chang et al. (2007) indicate that commitment predicts turnover intention concerning careers (occupational turnover intention) and place of employment (organizational turnover intention). That is, workers who report low levels of organizational commitment are more likely to seek jobs in different organizations but in the same field, and workers with low occupational commitment are more likely to change careers altogether. The authors also predicted that organizational turnover intention and occupational turnover intention are reciprocally related: Plans to change one's career may prompt leaving a particular organization, and vice versa. Accordingly, Chang et al. (2007) also administered measures of occupational turnover intention and organizational turnover intention to the nurses in their sample. Reported in Table 10.1 are values of sample standard deviations and internal consistency reliability (Cronbach's alpha) coefficients for all measured variables.

Besides mutual causation between organizational turnover intention and occupational turnover intention, Chang et al. (2007) also hypothesized that (1) the three components of organizational commitment (affective, continuance, normative) directly affect organizational turnover intention and (2) the three components of occupational

**TABLE 10.1. Sample Standard Deviations and Score Reliability Coefficients for Measures of Organizational Commitment, Occupational Commitment, and Turnover Intention**

Statistic	Organizational commitment			Occupational commitment			Turnover intention	
	1	2	3	4	5	6	7	8
SD	1.04	.98	.97	1.07	.78	1.09	1.40	1.50
$r_{xx}$	.82	.70	.74	.86	.71	.84	.86	.88

*Note.* These data are from H.-T. Chang et al. (2007);  $N = 177$ . Score reliabilities are internal consistency (Cronbach's alpha) coefficients. 1, 4 = affective; 2, 5 = continuance; 3, 6 = normative; 7 = organizational; 8 = occupational.

commitment directly affect occupational turnover intention. A nonrecursive path model that represents these hypotheses would consist of a direct feedback loop between organizational turnover intention and occupational turnover intention, direct effects of the three organizational commitment variables on organizational turnover intention, and direct effects of the three occupational commitment variables on occupational turnover intention. But a conventional path analysis would not permit the explicit representation of measurement error in any of the single indicators. Fortunately, the availability of score reliability coefficients for these data (Table 10.1) allows specification of the SR model in Figure 10.7 that controls for measurement error.

Each manifest variable in Figure 10.7 is represented as the single indicator of an underlying factor. The unstandardized pattern coefficient of each single indicator is fixed to 1.0 in order to scale the corresponding factor. Error variance for each indicator is fixed to equal the product of the sample variance of that indicator, or  $s^2$ , and one minus the score reliability for that indicator, or  $1 - r_{xx}$ . For instance, the reliability coefficient for scores on the affective organizational commitment variable is .82, and the sample standard deviation is 1.04 (see Table 10.1). The quantity

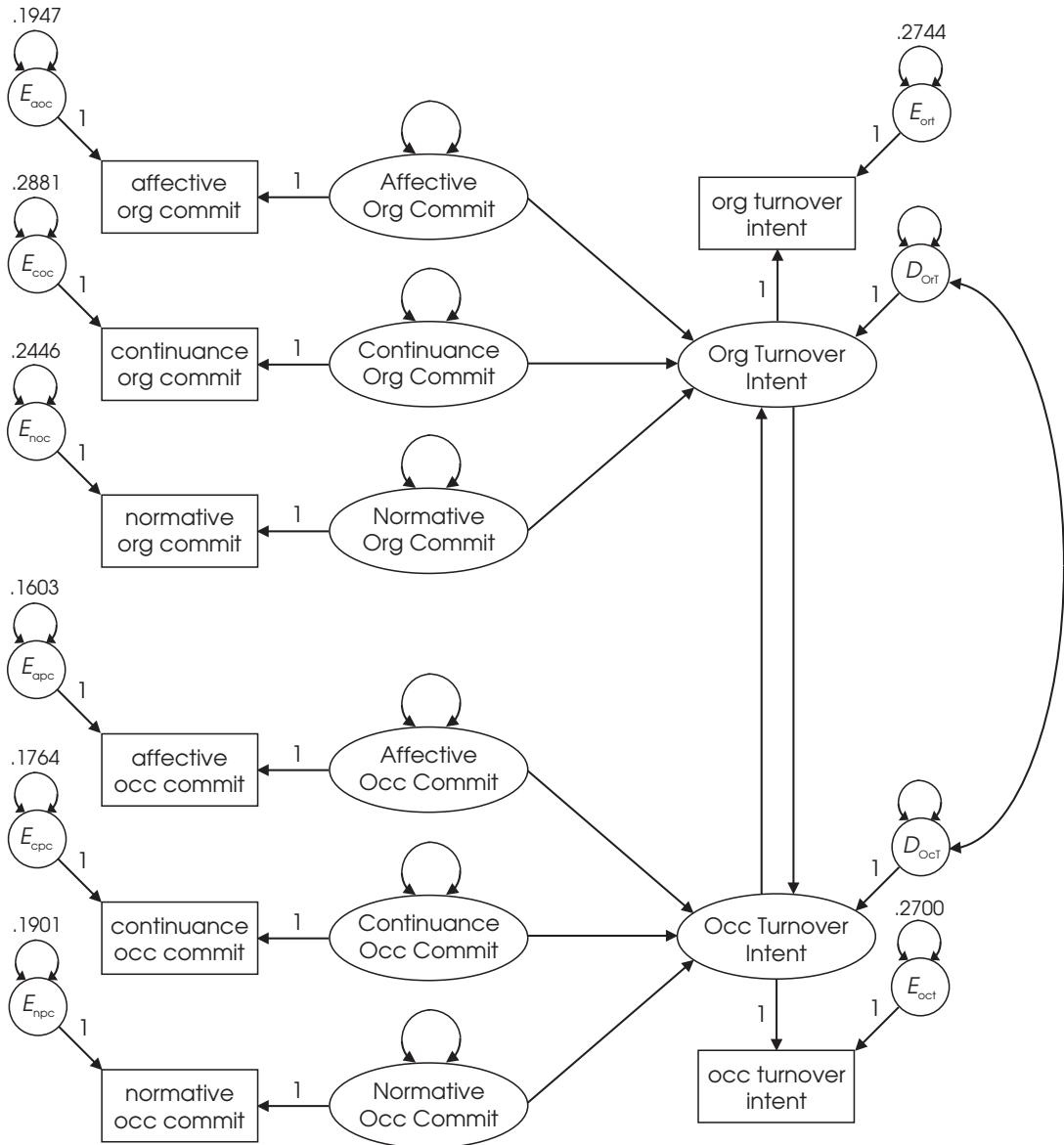
$$(1 - .82) 1.04^2 = .18 (1.0816) = .1947$$

estimates the amount of the unstandardized total variance that is due to measurement error. Accordingly, the unstandardized error variance for the affective organizational commitment indicator is fixed to equal .1947 (see Figure 10.7). Error variances for the remaining seven single indicators in the figure are calculated in similar ways. Exercise 5 asks you to calculate the value of the fixed error variance for the continuance organizational commitment indicator in Figure 10.7, given the data in Table 10.1.

Given the single indicator specifications just described, the estimation of the direct effects and disturbance variances and covariance for the structural part of the model in Figure 10.7 controls for measurement error. To save space, not all possible covariances among the six exogenous factors are shown in the figure, but they are assumed. Exercise 6 asks you to prove that  $df_M = 4$  for this model. Chang et al. (2007) analyzed a nonrecursive path model that involved the eight observed variables in Figure 10.7 but without controlling for measurement error. They also did not report the unstandardized solution.

## SUMMARY

Analysis of CFA models tests no hypotheses about causal relations between factors, but SR models have both exogenous and endogenous factors, where the endogenous factors are specified as the outcomes of other variables in the model. If every factor in the structural part of the model has multiple indicators, the SR model is fully latent; otherwise, the model is partially latent such that at least one hypothetical construct has a single indicator. It is possible to specify that the error variance of a single indicator in a par-



**FIGURE 10.7.** A nonrecursive model of organizational and occupational commitment and turnover intention with single-indicator specification that controls for measurement error. Names of observed variables are presented in lowercase characters. Pairwise covariances between the exogenous factors are omitted to save space. Unstandardized error variances for single indicators are fixed to the values indicated.

tially latent SR model is fixed to equal a constant provided by the researcher and based on an estimate of score precision. Doing so forces the computer to control for measurement error in single indicators when estimating parameters for the structural model. In order for an SR model to be identified, both its measurement and structural parts must be identified. This requirement reflects the view that the analysis of an SR model is basically a path analysis conducted with estimated covariances among its factors. We are ready to consider the analysis phase of SEM in Part III of this book.

## LEARN MORE

Cole and Preacher (2014) describe potential negative consequences of failing to control for measurement error in single indicators; Hayduk and Littvay (2012) consider possible disadvantages of multiple-indicator measurement compared with the use of best indicators; and Marsh et al. (2014) describe ESEM in clinical research.

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.

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, 12(159). Retrieved from [www.biomedcentral.com/1471-2288/12/159](http://www.biomedcentral.com/1471-2288/12/159)

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: Integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85–110.

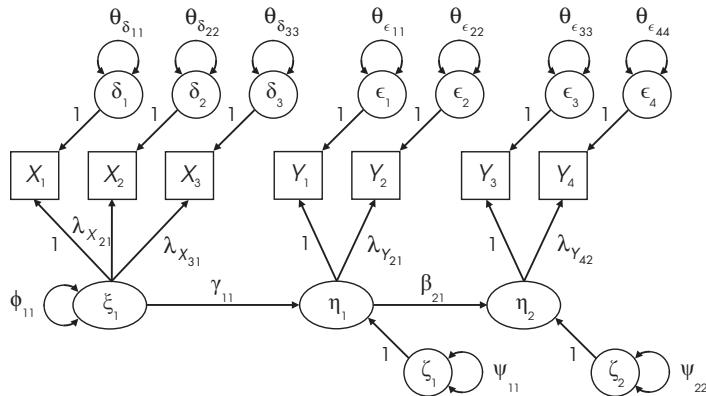
## EXERCISES

1. Calculate  $df_M$  for Figure 10.1(b).
2. Show that  $df_M$  is the same for Figures 10.2(a) and 10.3(a).
3. Respecify Figure 10.1(a) to control for measurement error in all single indicators. Assume reliability coefficients of .80, .75, and .90 for, respectively, variables  $X_1$ ,  $Y_1$ , and  $Y_3$ .
4. Calculate  $df_M$  for Figure 10.6.
5. Calculate the error variance for the continuance affective organizational commitment indicator in Figure 10.7 using the data in Table 10.1.
6. Calculate  $df_M$  for Figure 10.7.

## Appendix 10.A

### LISREL Notation for SR Models

Described next is LISREL notation for fully latent SR models when means are not analyzed. Indicators of exogenous factors are designated  $X$ , and indicators of endogenous factors are designated  $Y$ . Relevant lowercase Greek letters include  $\beta$  (beta),  $\gamma$  (gamma),  $\delta$  (delta),  $\epsilon$  (lunate epsilon),  $\zeta$  (zeta),  $\eta$  (eta),  $\theta$  (theta),  $\lambda$  (lambda),  $\xi$  (xi),  $\phi$  (phi), and  $\psi$  (psi); uppercase letters include  $B$  (beta),  $\Gamma$  (gamma),  $\Theta$  (theta),  $\Lambda$  (lambda),  $\Phi$  (phi), and  $\Psi$  (psi). Symbols for variables, parameters, and residual terms appear in their proper places in the SR model shown next:



Following are the measurement equations for the indicators:

$$\begin{aligned}
 X_1 &= \xi_1 + \delta_1 & Y_1 &= \eta_1 + \epsilon_1 \\
 X_2 &= \lambda_{X_{21}} \xi_1 + \delta_2 & Y_2 &= \lambda_{Y_{21}} \eta_1 + \epsilon_2 \\
 X_3 &= \lambda_{X_{31}} \xi_1 + \delta_3 & Y_3 &= \eta_2 + \epsilon_3 \\
 && Y_3 &= \lambda_{Y_{42}} \eta_2 + \epsilon_4
 \end{aligned} \tag{10.1}$$

These equations can be expressed in matrix algebra terms as follows:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \lambda_{X_{21}} \\ \lambda_{X_{31}} \end{bmatrix} \begin{bmatrix} \xi_1 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix} = \boldsymbol{\Lambda}_X \boldsymbol{\xi} + \boldsymbol{\delta} \tag{10.2}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{Y_{21}} & 0 \\ 0 & 1 \\ 0 & \lambda_{Y_{42}} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{bmatrix} = \boldsymbol{\Lambda}_Y \boldsymbol{\eta} + \boldsymbol{\epsilon} \quad (10.3)$$

where  $\boldsymbol{\Lambda}_X$  (lambda-X) is the parameter matrix of pattern coefficients for the X indicators and  $\boldsymbol{\Lambda}_Y$  (lambda-Y) is the corresponding matrix for the Y indicators. Other parameter matrices for the measurement model are

$$\boldsymbol{\Phi} = [\phi_{11}] \quad \boldsymbol{\Theta}_\delta = \begin{bmatrix} \theta_{\delta_{11}} & & \\ 0 & \theta_{\delta_{22}} & \\ 0 & 0 & \theta_{\delta_{33}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Theta}_\epsilon = \begin{bmatrix} \theta_{\epsilon_{11}} & & & \\ 0 & \theta_{\epsilon_{22}} & & \\ 0 & 0 & \theta_{\epsilon_{33}} & \\ 0 & 0 & 0 & \theta_{\epsilon_{44}} \end{bmatrix} \quad (10.4)$$

where  $\boldsymbol{\Phi}$ ,  $\boldsymbol{\Theta}_\delta$  (theta-delta), and  $\boldsymbol{\Theta}_\epsilon$  (theta-epsilon) are the covariances matrices of, respectively, the exogenous factors, error terms of the X indicators, and error terms of the Y indicators.

Equations for the structural part of the example SR model are

$$\begin{aligned} \eta_1 &= \gamma_{11} \xi_1 + \zeta_1 \\ \eta_2 &= \beta_{21} \eta_1 + \zeta_2 \end{aligned} \quad (10.5)$$

and the corresponding matrix algebra expression is

$$\begin{aligned} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} &= \begin{bmatrix} \gamma_{11} \\ 0 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \zeta_1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \\ &= \boldsymbol{\Gamma} \boldsymbol{\xi} + \mathbf{B} \boldsymbol{\eta} + \boldsymbol{\zeta} \end{aligned} \quad (10.6)$$

where  $\boldsymbol{\Gamma}$  is the parameter matrix for direct effects of exogenous factors on endogenous factors and  $\mathbf{B}$  is the parameter matrix for direct effects of endogenous factors on each other. The only other parameter matrix is

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_{11} & \\ 0 & \psi_{22} \end{bmatrix} \quad (10.7)$$

where  $\boldsymbol{\Psi}$  is the covariance matrix for the disturbances of endogenous factors. Thus, full LISREL notation for SR models consists of a total of eight parameter matrices listed next:

$$\mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Theta}_\delta, \boldsymbol{\Theta}_\epsilon, \boldsymbol{\Lambda}_X, \boldsymbol{\Lambda}_Y, \boldsymbol{\Phi}, \text{ and } \boldsymbol{\Psi}$$

Notation in LISREL for path models or CFA models is just a subset of the notation for SR models (Appendices 6.A, 9.A).

For SR models some authors use the simpler **LISREL all-Y notation**, which does not distinguish between exogenous and endogenous variables. For example, all factors and indicators are designated, respectively, with the symbols  $\eta$  and  $Y$ . Variances and covariances among exogenous factors are now represented in the matrix  $\boldsymbol{\Psi}$  along with the corresponding terms among the dis-

turbances of the endogenous factors. Each element in  $\Psi$  is designated as associated with either an exogenous or endogenous variable in the original model. In addition, direct effects on endogenous factors from exogenous factors or other endogenous factors are now represented in the  $\mathbf{B}$  matrix, and error variances and covariances for all indicators are represented in the  $\Theta_\epsilon$  matrix. Thus, the parameter matrices in LISREL all-Y notation are

$$\mathbf{B}, \Theta_\epsilon, \Lambda_Y \text{ and } \Psi$$

## **Part III**

# Analysis



# Estimation and Local Fit Testing

---

This chapter is organized into three main parts. Described in the first part are the two basic categories of estimators of causal effects in SEM, methods that analyze a single equation at a time versus methods that simultaneously analyze the whole model. Characteristics of maximum likelihood estimation, a simultaneous method that is the default in most SEM computer tools, are also outlined. Next, local fit testing is described. Although global fit testing predominates in SEM, it is critical to evaluate the particulars of model–data correspondence in local fit testing. These topics are illustrated through the detailed analysis of a recursive path model with continuous outcomes. The last part of this chapter considers alternative estimators for outcomes that are not continuous. The topics dealt with here will help prepare you to learn about global fit testing, the subject of the next chapter.

---

## TYPES OF ESTIMATORS

There are two main kinds of estimation methods in SEM. The first category, **single-equation methods**—also known as **partial-information methods** or **limited-information methods**—analyzes the equation for a single endogenous variable at a time. These methods do not assume multivariate normality, nor do they require identified models; in addition, they can be less affected by specification error than simultaneous methods. A drawback of single-equation methods is that there are no significance tests of global model fit or continuous measures of model–data correspondence. For this reason, local fit testing is emphasized when single-equation methods are used (there is no practical alternative).

The second category, **simultaneous methods** or **full-information methods**, estimate all free model parameters at once. These methods require identified models. Under conditions that may not hold in many actual studies, simultaneous methods are gen-

erally more efficient than single-equation methods. An **efficient estimator** has lower variation among estimates of the same parameter than a less efficient estimator, when analyzing a correctly specified model in random samples. This property of simultaneous methods is realized because they take more advantage of information in the data than single-equation methods. But this potential benefit may be more theoretical than tangible because researchers rarely analyze correctly specified models in representative samples. Global fit testing is usually emphasized when simultaneous methods are used, but local fit testing can—and should—be conducted, too.

## CAUSAL EFFECTS IN PATH ANALYSIS

Let's assume that every endogenous variable in a path model is continuous and there are no interactions. A direct effect quantifies the sensitivity of  $Y$  to changes in  $X$  while controlling for other variables (covariates) that sever all paths from  $X$  to  $Y$  except for the direct link  $X \rightarrow Y$ , for which there are no intermediaries (e.g., mediation) (Pearl, 2009b). It is also the slope of the tangent line for the functional relation between  $X$  and  $Y$ , controlling for other parents of  $Y$ . For linear effects, this slope is constant over all levels of  $X$  (e.g., Figure 2.1), and a single quantity—the path coefficient—estimates the direct effect. If the relation is curvilinear, the slope of the tangent line changes across the levels of  $X$ ; that is, the direct effect is not constant. From this perspective, curvilinear relations are a special kind of interaction effect where the magnitude or direction of the association changes across the levels of  $X$  (e.g., Figure 1.1). In this case, the direct effect cannot be estimated with a single number.

A path coefficient for a linear effect is interpreted exactly as a regression coefficient that may be either in unstandardized or standardized form. Specifically, an unstandardized path coefficient estimates the amount of raw score change in  $Y$ , given a change of one point in the original metric of  $X$ , controlling for other parents of  $Y$ . The coefficient for the standardized direct effect estimates the corresponding amount of change in  $Y$  as the proportion of a standard deviation, given a change in  $X$  of a full standard deviation (i.e., it is interpreted as beta weight).

With no interactions, an indirect effect of  $X$  on  $Y$  is estimated as the product of the individual coefficients for each direct effect that makes up that indirect causal pathway (e.g., the quantity  $cd$  for Figure 6.5(b)). This product is also interpreted as a regression coefficient, but one that estimates the amount of change in  $Y$ , given a unit change in  $X$ , through the indirect pathway while controlling for the direct effect of  $X$  on  $Y$ . In models with multiple indirect effects of  $X$  on  $Y$ , the **total indirect effect** is estimated as the sum of the coefficients for each individual indirect effect. Coefficients for total indirect effects are also interpreted as regression coefficients, but now those that represent the effect of  $X$  on  $Y$  through all indirect pathways between them, again controlling for the direct effect of  $X$ .

The total causal effect of  $X$  on  $Y$  is estimated controlling for other variables that sever all back-door (noncausal) paths between  $X$  and  $Y$ , leaving only direct or indirect

causal paths between them. Total effects are also interpreted as regression coefficients, but those that estimate the expected change in  $Y$ , given a unit change in  $X$  through all direct or indirect causal connections between them, after eliminating noncausal associations. The total effect is the sum of the direct effect and total indirect effect of  $X$  on  $Y$ .

## SINGLE-EQUATION METHODS

Described next are single-equation estimators for path models with continuous endogenous variables.

### Multiple Regression

Multiple regression (i.e., OLS estimation) can be used for recursive path models. The total effect of  $X$  on  $Y$  is estimated by regressing  $Y$  on  $X$  and the covariates that meet the back-door criterion (Rule 8.3). These covariates for the total effect block all noncausal paths between  $X$  and  $Y$ . The direct effect of  $X$  on  $Y$  is estimated by regressing  $Y$  on  $X$  and the covariates that satisfy the single-door criterion (Rule 8.4). These covariates for the direct effect d-separate  $X$  and  $Y$  in the modified model formed by deleting the path  $X \rightarrow Y$  from the original model. Depending on the model, there can be two or more sets of covariates that identify the same total effect or direct effect. The values of these different estimators should be similar, if the model is correct. Other analysis details are summarized next:

1. Variances and covariances of measured exogenous variables are simply the observed (sample) values. For example, the Pearson correlation between a pair of continuous exogenous variables estimates their association in the standardized solution, and the covariance for the same pair estimates their unstandardized association.
2. To estimate a disturbance variance, record  $R^2$  from each analysis where an endogenous variable is regressed on all its parents.<sup>1</sup> The product  $(1 - R^2)s_Y^2$ , where  $s_Y^2$  is the observed variance of the corresponding endogenous variable, equals the unstandardized disturbance variance. The quantity  $(1 - R^2)$  estimates the standardized disturbance variance as the proportion of unexplained variance.
3. Disturbance covariances for pairs of endogenous variables in a bow-free pattern are estimated as follows: The unstandardized estimate is the partial covariance between the endogenous variables controlling for their common causes, and the standardized estimate is the corresponding partial correlation (Kenny, 1979, pp. 52–61).

---

<sup>1</sup>An alternative to  $R^2$  in samples that are not large is  $\hat{R}^2$ , the shrinkage-corrected estimate (Equation 2.14).

## Two-Stage Least Squares

Standard OLS estimation is not appropriate for nonrecursive path models with causal loops or bow-pattern disturbance correlations because regression residuals are calculated to be independent of the predictors. In a causal model, this aspect of OLS translates to the requirement that the causes of an endogenous variable cannot covary with its disturbance. Nonrecursive causal relations violate this requirement, so an alternative is needed.

The method of **two-stage least squares (2SLS)** is a type of **instrumental variables regression** that is suitable for analyzing nonrecursive models.<sup>2</sup> It involves instruments that identify nonrecursive causal relations (e.g., Figure 7.3). As its name suggests, 2SLS is actually nothing more than OLS estimation but applied over two steps. The aim of the first stage is to replace a problematic causal variable with a newly created predictor. A “problematic” causal variable is correlated with the disturbance of the outcome variable. The problematic causal variable is regressed on the instrument(s). The predicted criterion variable in this analysis will be uncorrelated with the disturbance of the outcome variable. When similar replacements are made for all problematic causal variables, we proceed to the second stage of 2SLS, which is just standard OLS estimation conducted for each outcome variable but using the predictors created in the first step whenever the original causal variable was replaced.

As an example, look back at Figure 6.6(a). This nonrecursive model specifies two direct causes of  $Y_1$ , the variables  $X_1$  and  $Y_2$ . From the perspective of OLS estimation,  $Y_2$  is a problematic cause because it covaries with the disturbance of  $Y_1$ . The predicted association is represented in the figure by the noncausal path

$$Y_2 \leftarrow D_2 \rightsquigarrow D_1$$

The instrument here is  $X_2$  because it is excluded from the equation for  $Y_1$  and has a direct effect on  $Y_2$ , the problematic causal variable (i.e.,  $X_2$  satisfies Rule 8.5). Therefore, we regress  $Y_2$  on  $X_2$  in a standard regression analysis. The predicted criterion variable from this first analysis,  $\hat{Y}_2$ , replaces  $Y_2$  as a predictor of  $Y_1$  in a second regression analysis where  $X_1$  is the other predictor. The coefficients from the second regression analysis are taken as the estimates of the direct effects of  $X_1$  and  $Y_2$  on  $Y_1$ .

The 2SLS technique is widely used in disciplines such as epidemiology and economics. Many computer programs for general statistical analysis have 2SLS procedures. The LISREL program uses a special form of 2SLS estimation that calculates start values for latent-variable models. A variation known as **three-stage least squares (3SLS)** adds a third step to the basic 2SLS method that controls for correlated errors in the model. This makes 3SLS more like a simultaneous method in that it takes account of features in the whole model when estimating its parameters. See Bollen (2012) for more examples of analyses with instrumental variables.

---

<sup>2</sup>The 2SLS method can also analyze recursive path models, but the results in this case are identical to those from standard multiple regression.

## SIMULTANEOUS METHODS

Because full-information methods estimate all free parameters at once, the overriding assumption is that *the model is correctly specified*. This assumption is critical due to **propagation of specification error**. Simultaneous methods tend to spread such errors throughout the entire model. That is, a specification error in one parameter can affect results for other parameters elsewhere in the model. Suppose that a common cause of a pair of endogenous variables is not measured, but their disturbances are specified as independent. This specification error may propagate to estimation of the direct effects or disturbance variances for this pair of outcomes. It is difficult to predict the direction or magnitude of this “contamination,” but the more serious the specification error, the more serious may be the resulting bias in other parts of the model.

When misspecification occurs, single-equation methods may outperform simultaneous methods. This occurs because single-equation methods may better isolate the effects of errors to misspecified parts of the model instead of allowing them to spread to other parts. In a computer simulation study, Bollen, Kirby, Curran, Paxton, and Chen (2007) found that bias in maximum likelihood (ML) estimation—a simultaneous method—and various 2SLS estimators was generally negligible when a three-factor measurement model was correctly specified. But when model specification was incorrect, there was greater bias of the ML estimator compared with that of the 2SLS estimator, even in large samples. Based on these results, Bollen et al. (2007) suggested that researchers consider a 2SLS estimator as a complement to or even a substitute for ML estimation when there is doubt about specification.

## MAXIMUM LIKELIHOOD ESTIMATION

The ML method can be applied to the whole range of structural equation models. It “knows” how to use instruments, so it can estimate nonrecursive causal relations in path models. It can also analyze models with substantive latent variables. The term **maximum likelihood** describes the principle that underlies the derivation of parameter estimates: The estimates are the ones that maximize the likelihood that the data (the observed covariances) were drawn from this population. Default ML estimation is a normal theory method that assumes multivariate normality for the joint population distribution of the endogenous variables, given the exogenous variables. Only continuous variables can have normal distributions; therefore, if the endogenous variables are not continuous or if their distributions are severely non-normal, then an alternative method may be needed.

The statistical criterion minimized, or the **fit function**, is related to the discrepancy between sample covariances and those predicted by the researcher’s model. The final set of parameter estimates minimizes squared differences between the respective elements of the two matrices just mentioned. Parameters are estimated iteratively in a nonlinear optimization algorithm that minimizes the fit function. The mathematics of ML estima-

tion are complex, and it is beyond the scope of this section to describe them in detail—see Enders (2010, chap. 3) for a gentle introduction or Mulaik (2009b, chap. 7) for a more quantitative presentation. There are points of contact between ML estimation and OLS estimation. For example, estimates of path coefficients for recursive path models are basically identical. Estimates of disturbance variances may differ slightly in small samples, but the two methods generally yield similar results in large samples.

### Variance Estimates

The population variance  $\sigma^2$  is estimated in the ML method as  $S^2 = SS/N$ , where the numerator is the total sum of squared deviations from the mean. In OLS estimation,  $\sigma^2$  is estimated as  $s^2 = SS/df$ , where  $df = N - 1$ . In small samples,  $S^2$  estimates  $\sigma^2$  with negative bias. In large samples, values of  $S^2$  and  $s^2$  are similar, and they are asymptotic in very large samples. Variances calculated as  $s^2$  in a computer program for general statistical analysis, such as SPSS, may not exactly equal those calculated in an SEM computer as  $S^2$  for the same variables. Check the documentation of your SEM computer tool to avoid confusion about this issue. Some SEM computer tools, such as the `sem` command in Stata, allow the user to specify whether variances should be estimated as  $s^2$  or  $S^2$  (i.e., with, respectively,  $N - 1$  or  $N$  in the denominator).

### Iterative Estimation and Start Values

Implementations of ML estimation are typically iterative, which means that the computer derives an initial solution and then attempts to improve these estimates through subsequent cycles of calculations. The computer therefore repeatedly “auditions” somewhat different values until it finds the set of parameters estimates that is most likely to have generated the observed data. “Improvement” means that the overall fit of the model to the data gradually gets better. For most just-identified models, the fit of the model to the data will eventually be perfect. For overidentified models, the fit of the model may be imperfect, but iterative estimation will continue until improvements in fit fall below a predefined value. When this happens, the estimation process has converged.

The *Ωnyx* program for SEM (von Oertzen et al., 2015) uses a **multi-agent estimation algorithm** (Pinter, 1996) as it attempts to fit the model to the data. For example, after the first converged solution is found and displayed onscreen, the program continues to refine the estimates in the background. If better estimates are found later, the researcher is notified. The algorithm also alerts the researcher to **multiple optima**, or the existence of multiple solutions that satisfy the same statistical criterion nearly to the same degree. In contrast, most other SEM computer tools display just the best solution regardless of whether other solutions are nearly as good. If two solutions with quite different parameter estimates generate about the same degree of fit between model and data, then little confidence may be warranted in either solution.

Iterative estimation may converge more quickly if the procedure is given reasonably accurate start values or initial estimates of some parameters. If these initial estimates

are grossly inaccurate—for instance, the start value for a path coefficient is positive when the actual direct effect is negative—then iterative estimation may fail to converge, which means that a stable solution has not been reached. Computer programs typically issue a warning if iterative estimation fails. When failure of iterative estimation occurs, whatever final set of estimates was derived by the computer warrants little confidence. Some SEM computer tools automatically generate their own start values. *But computer-derived start values do not always lead to converged solutions.* Sometimes it is necessary for the researcher to provide better start values; see Appendix 11.A. Another tactic is to increase the program’s default limit on the number of iterations to a higher value, such as from 30 to 100. Allowing the computer more “tries” may lead to a converged solution.

### Inadmissible Solutions and Heywood Cases

Although usually not a problem when analyzing recursive path models, a converged solution may be **inadmissible** in ML estimation and other iterative methods. This problem is most evident by a parameter estimate with an illogical value, such as **Heywood cases** (after the statistician H. B. Heywood). These cases include negative variance estimates (e.g., an unstandardized disturbance variance is  $-12.58$ ) or estimated absolute correlations  $> 1.0$  (e.g., the correlation between a pair of factors is  $1.08$ ). Another example of a problem is when the standard error of a parameter estimate is so large that no interpretation seems plausible. Causes of Heywood cases (Chen, Bollen, Paxton, Curran, & Kirby, 2001) include:

1. Specification errors.
2. Nonidentification of the model.
3. The presence of outliers that can distort the solution.
4. A combination of small sample sizes and only two indicators per factor in latent-variable models.
5. Bad start values.
6. Extremely low or high population correlations that result in empirical under-identification.

An analogy may help to give a context for Heywood cases: ML estimation (and related simultaneous methods) is like a religious fanatic in that it so believes the model’s specification that it will do anything, no matter how crazy, to force the model on the data. Some SEM computer tools do not permit certain Heywood cases. For example, EQS automatically imposes a lower bound—an inequality constraint—of zero on variance estimates, which precludes negative values. But solutions in which one or more estimates have been constrained by the computer to prevent an illogical value should not be trusted. It is better to try to determine the source of the problem instead of constraining an error variance to be positive and then rerunning the analysis.

Always carefully inspect the solution, unstandardized and standardized, for any sign that it is inadmissible. Computer programs for SEM generally issue warning messages about Heywood cases, but they are not foolproof. It is possible for a solution to be inadmissible but no warning was given. It is the researcher, not the computer, who provides the ultimate quality control check for solution admissibility.

### **Scale Freeness and Scale Invariance**

The ML method is generally both scale free and scale invariant. **Scale free** means that if a variable's scale is linearly transformed, a parameter estimated for the transformed variable can be algebraically converted back to the original metric. **Scale invariant** means that the value of the fit function in a particular sample remains the same regardless of the metrics of the original variables (Kaplan, 2009). These properties may be lost if a correlation matrix is analyzed instead of a covariance matrix. This is because ML estimation and most other simultaneous methods assume unstandardized variables; that is, either a covariance matrix or a raw data file of scores that are not standardized is submitted.

### **Other Requirements**

Additional requirements of the default ML method include large samples, independent scores, normally distributed errors, no missing values when a raw data file is analyzed, and independence of the exogenous variables and disturbances. An extra assumption when a path model is analyzed is that the exogenous variables are measured without error. This requirement can be relaxed if the researcher applies the single-indicator respecification that controls for measurement error (e.g., Figure 10.3(a)).

### **Variations**

Many SEM computer tools offer variations on default ML estimation, but it may be necessary to explicitly request them. A special version for incomplete raw data files was described in Chapter 4. Robust maximum likelihood (MLR) estimation is for continuous endogenous variables with severely non-normal distributions. It is an alternative to normalizing the original variables with transformations and then analyzing the transformed data with default ML estimation. The MLR estimator is a **corrected normal theory method**. That is, the original data are analyzed with a normal theory method, such as default ML, but robust standard errors and corrected model test statistics are to be used (Savalei, 2014). **Robust standard errors** are estimates of standard errors that are supposedly robust against non-normality. A **corrected model test statistic** is a significance test of fit of the whole model to the data matrix that is adjusted for non-normality. Analysis of a raw data file is required for the MLR method.

Summarized next are the possible consequences of analyzing continuous but severely non-normal outcomes with the default ML (i.e., not MLR) method (Olsson, Foss, Troye, & Howell, 2000):

1. Values of parameter estimates may be relatively accurate in large samples, but their standard errors tend to be too low, perhaps by as much as 25 to 50%, given the model and data. This results in rejection of the null hypothesis that the corresponding population parameter is zero more often than is correct (Type I error rate is inflated).
2. Values of model test statistics tend to be too high, which results in rejection of the null hypothesis that the model has perfect fit in the population more often than is correct. This means that true models are rejected too often. The actual rate of this error may be as high as 50% when the expected rate assuming normality is 5%, again depending on the model and data.

Another option is to use default ML but with nonparametric bootstrapping, which assumes only that the population and sample distributions have the same shape. In this approach, parameters and standard errors are estimated in empirical sampling distributions (e.g., Figure 3.3). The **Bollen–Stine bootstrap** (Bollen & Stine, 1993) generates adjusted  $p$  values for model test statistics. Each generated sample is drawn from transformed data that assume perfect model–data correspondence in the population, and the computer records the proportion of times that the model test statistic from the generated samples exceeds the model test statistic for the observed data. This proportion is the corrected  $p$  value. Nevitt and Hancock (2001) found in computer simulations that bootstrapped estimates were generally less biased than those from default ML estimation under conditions of non-normality and large sample sizes. But in smaller samples (e.g.,  $N < 200$ ), bootstrapped estimates had relatively large standard errors, and many generated samples had nonpositive definite data matrices. These problems are consistent with the caution that bootstrapped results in small samples can be very inaccurate.

Some SEM computer tools, such as Mplus, allow use of the MLR method with any combination of continuous, ordered-categorical (ordinal), unordered-categorical (nominal), or censored endogenous variables. For instance, analyses for binary outcomes, such as relapsed—not relapsed, can be analyzed in logistic-type regressions where the path coefficients are odds ratios. Latent continuous variables with normal distributions are generally presumed to underlie binary or ordinal data (e.g., Figure 4.4). Other estimators for noncontinuous outcomes are described later in this chapter.

## DETAILED EXAMPLE

Considered next is parameter estimation and local fit testing for the recursive path model of illness introduced in Chapter 7. In the next chapter, you will learn about how to evaluate the same model in global fit testing. Parameter estimation is discussed before global testing because too many researchers are so preoccupied with global fit that they pay insufficient attention to the meaning of the parameter estimates (Kaplan, 2009).

Briefly, Roth et al. (1989) administered measures of exercise, hardiness, fitness, stress, and illness in a sample of 373 university students. These data are summarized in

Table 4.2. The recursive path model in Figure 7.5 represents the hypotheses that effects of exercise and hardiness on illness are purely indirect through a single intermediary, fitness for exercise and stress for hardiness. You can download from this book's website all computer syntax, data, and output files for this example in Amos, EQS, LISREL, lavaan for R, Mplus, SPSS, and Stata.

### Conditional Independences

A total of five nonadjacent pairs of measured variables in Figure 7.5 can be d-separated, so the size of the basis set is 5. Listed in the first and second columns of Table 11.1 are the conditional independences of a basis set that satisfies Rule 8.2. For example, given the model, the variables exercise and stress should be independent controlling for hardiness, the parent of stress. Listed in the third column of the table is the value of the partial correlation that corresponds to each conditional independence. If the model is correct, these sample correlations should all “vanish,” or be approximately zero. Each coefficient just mentioned is also a **correlation residual**, or the difference between the observed value and a predicted value (zero). The rule of thumb is that absolute discrepancies between predicted and observed correlations of .10 or more may signal appreciable model–data disagreement. Although it is difficult to say how many absolute correlation residuals of .10 or more is “too many,” the more there are, the worse the explanatory power of the model at the level of pairs of variables.

There is one absolute correlation residual that is just .10 or more. This result, **−.103** (shown in boldface in Table 11.1), is for the pair fitness and stress. The model predicts that fitness and stress are independent, given exercise and hardiness, but their observed residual association differs appreciably from zero. In Figure 7.5, there is a single backdoor path between fitness and stress:

$$\text{Fitness} \leftarrow \text{Exercise} \curvearrowright \text{Hardiness} \rightarrow \text{Stress}$$

A possible specification error is that fitness and stress are related through paths omitted from the original model. For example, perhaps fitness affects stress ( $\text{Fitness} \rightarrow \text{Stress}$ )

**TABLE 11.1. A Basis Set of Conditional Independences for a Recursive Path Model of Illness and Corresponding Partial Correlations**

Independence	Conditioning set	Partial correlation
Exercise $\perp$ Stress	Hardiness	−.058
Exercise $\perp$ Illness	Fitness, Stress	.039
Hardiness $\perp$ Fitness	Exercise	.089
Hardiness $\perp$ Illness	Fitness, Stress	−.081
Fitness $\perp$ Stress	Exercise, Hardiness	<b>−.103</b>

or stress affects fitness (Stress → Fitness). In the next chapter, we will deal with respecification in more detail, but we have already detected a problem with model–data correspondence in local fit testing. There are additional problems concerning local fit, as we shall see.

## Single-Equation Estimation with Multiple Regression

The unstandardized coefficient for the covariance between exercise and hardness in Figure 7.5 is just the sample covariance, which is calculated from the summary statistics in Table 4.2 as

$$-.03 (66.50) (38.00) = -75.81$$

The standardized estimate is just the observed correlation, or  $-.03$ .

Estimates of direct effects are reported in Table 11.2. Because the variables exercise and fitness are already d-separated in the modified model formed by deleting the direct effect between them (see Figure 7.5), the set  $\emptyset$  (i.e., no covariates) is minimally sufficient to identify this direct effect (Rule 8.4). Thus, the bivariate regression of fitness on exercise estimates the causal effect of exercise on fitness. The unstandardized coefficient is  $.108$  (see the table), which says that fitness is expected to increase by  $.108$  points, given a one-point increase in exercise. Its standard error is  $.013$ , so  $z = .108/.013 = 8.31$ , which exceeds the critical value for two-tailed statistical significance at the  $.01$  level, or  $2.58$ . The standardized coefficient is  $.390$ , which says that fitness increases by  $.39$  standard deviations, given an increase in exercise of a full standard deviation. Exercise 1 asks you to interpret results in Table 11.2 for the direct effect of hardness on stress.

Listed next are the three minimally sufficient sets that meet Rule 8.4 and thus each identify the direct effect of fitness on illness in Figure 7.5:

(Exercise), (Hardiness), and (Stress)

This means that we can obtain three different estimators by regressing illness on fitness plus one of the covariates exercise, hardness, or stress. Coefficients for fitness in the analyses just described are reported in Table 11.2. All three sets of results are similar. For example, unstandardized estimates for fitness range from  $-1.036$  to  $-.849$ , and the corresponding standardized coefficients range from  $-.305$  to  $-.250$  (see the table). Looking now only at results with both parents of illness (shown in boldface), we find that a one-point increase in fitness predicts a decrease in illness of  $.849$  points, and an increase in fitness of one standard deviation leads to a decrease in illness of  $.250$  standard deviations, both controlling for stress. Exercise 2 asks you to interpret the results in Table 11.2 for the direct effect of stress on illness.

Reported in the second column of Table 11.3 is the observed variance ( $s^2$ ) for the endogenous variables fitness, stress, and illness (Table 4.2). Listed in the third column are values of  $R^2$  where the predictors are the parents of each outcome. Standardized

**TABLE 11.2. Ordinary Least Squares Estimates of Direct Effects for a Recursive Path Model of Illness**

Direct effect	Minimally sufficient set				
	$\emptyset$	Exercise	Hardiness	Stress	Fitness
Exercise → Fitness	.108 (.013) .390	—	—	—	—
Hardiness → Stress	-.203 (.045) -.230	—	—	—	—
Fitness → Illness	—	-.1.036 (.183) -.305	-.951 (.168) -.280	-.849 (.162) -.250	—
Stress → Illness	—	.628 (.091) .337	.597 (.093) .320	—	.574 (.089) .307

Note. Estimates are reported as unstandardized (standard error) standardized;  $\emptyset$ , empty set (no covariates). Values in boldface control for the parents of each outcome.

estimates of disturbance variances are calculated as  $1 - R^2$  and reported in the fourth column, and the unstandardized disturbance variances calculated as  $(1 - R^2) s^2$  are listed in the last column. For example,  $R^2 = .152$  for fitness where the parent is exercise, so the standardized disturbance variance is the proportion of unexplained variance, or  $1 - .152 = .848$ . Given  $s^2 = 338.56$  for fitness, the unstandardized disturbances variance is .848 (338.56), or 287.099. Exercise 3 asks you to interpret the results in Table 11.3 for the illness variable.

We now have OLS estimates for all parameters. The unstandardized values are shown in their proper places in Figure 11.1(a), and the standardized results are presented in Figure 11.1(b). Estimates of direct effects on illness control for both of its parents, fitness and stress (Table 11.2). Because not all measured variables have the same score metric, the unstandardized path coefficients for direct effects of fitness and stress on illness cannot be directly compared. This is not a problem for the standardized coefficients. These values for fitness and stress are, respectively,  $-.250$  and  $.307$ . Thus, the absolute magnitude of the standardized direct effect of stress on illness exceeds that of fitness by about 23% ( $.307/.250 = 1.23$ ).

The direct effects of both exercise and hardiness on illness are fixed to zero. Each causal variable just mentioned has a single indirect effect on illness, exercise through fitness and hardiness through stress (see Figure 11.1). Both of these indirect effects are also total effects, so they can be estimated for this example in two different ways: (1) as products of the coefficients from the direct paths that comprise each indirect pathway and (2) through covariate adjustment. Both types of estimates assume no interactions and are described next.

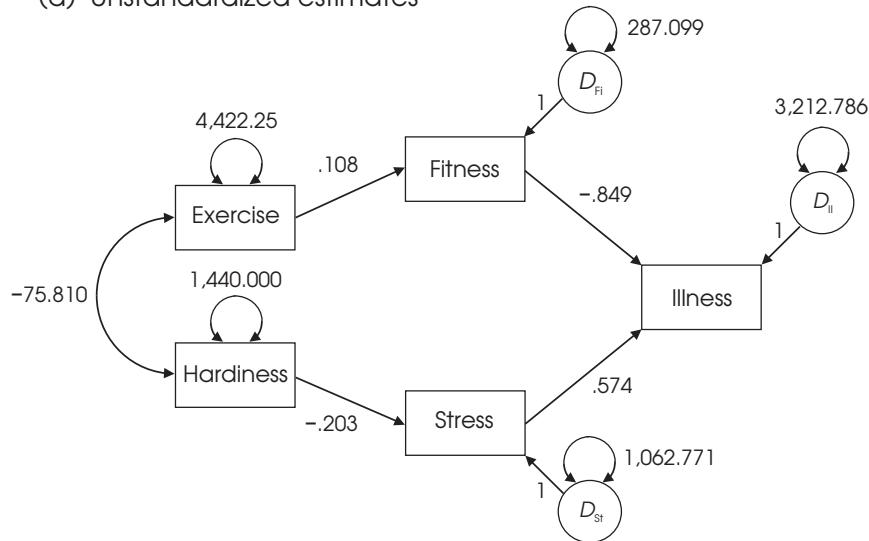
The product estimator for the unstandardized indirect effect of exercise on illness through fitness is  $.108$  ( $-.849$ ), or  $-.092$ , which equals the product of the unstandardized path coefficients for the direct effects that make up this indirect pathway (see Figure 11.1(a)). Thus, given a 1-point increase in exercise, we predict a decrease in illness of  $.092$  point through the intervening variable of fitness. For the standardized indirect effect, the product estimator is  $.390$  ( $-.250$ ), or  $-.098$ . The last-named quantity is the product of the standardized coefficients for the constituent direct effects (see Figure

**TABLE 11.3. Ordinary Least Squares Estimates of Disturbance Variances for a Recursive Path Model of Illness**

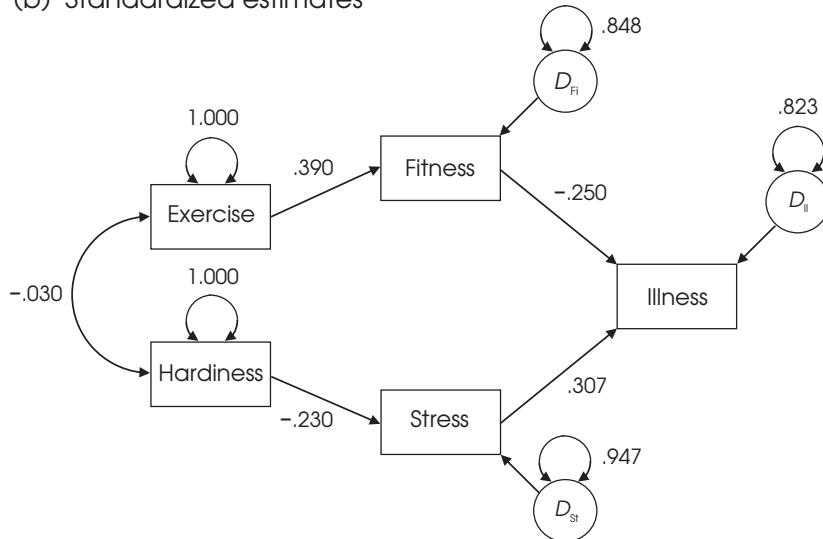
Outcome	$s^2$	$R^2$	Standardized estimate	Unstandardized estimate
Fitness	338.56	.152	.848	287.099
Stress	1,122.25	.053	.947	1,062.771
Illness	3,903.75	.177	.823	3,212.786

*Note.* The parent(s) of fitness, stress, and illness are, respectively, exercise, hardiness, and both fitness and stress.

(a) Unstandardized estimates



(b) Standardized estimates



**FIGURE 11.1.** A recursive path model of illness with ordinary least squares estimates. Standardized estimates for the disturbance variances are proportions of unexplained variance. Results for illness are based on both fitness and stress as direct causes.

11.1(b)).<sup>3</sup> This result says that given an increase in exercise of one standard deviation, we expect a decrease in illness of .098 standard deviations through the intervening variable of fitness. Exercise 4 asks you to calculate and interpret product estimators for the indirect effect of hardiness on illness through stress.

Because product estimates of indirect effects have complex distributions, it can be difficult to estimate their standard errors in significance testing. The best known example of an approximate method for unstandardized indirect effects that involve just three variables is the **Sobel test** (Sobel, 1982). Suppose that  $a$  is the unstandardized coefficient for the path  $X \rightarrow W$  and that  $SE_a$  is its standard error. Let  $b$  and  $SE_b$ , respectively, represent the same things for the path  $W \rightarrow Y$ . The product  $ab$  estimates the unstandardized indirect effect of  $X$  on  $Y$  through  $W$ . Its standard error is approximated as

$$SE_{ab} = \sqrt{b^2 SE_a^2 + a^2 SE_b^2} \quad (11.1)$$

In large samples, the ratio  $ab/SE_{ab}$  is interpreted as a  $z$  test of the unstandardized indirect effect. A webpage by K. Preacher automatically calculates the Sobel test and other variations, such as the Goodman test, that are based on somewhat different approximations of the standard error.<sup>4</sup>

Do not expect  $p$  values from the Sobel test or related methods to be accurate. The  $z$  test assumes normality, but distributions of product estimators are not generally normal. This is also more than a way to approximate standard errors, and the same unstandardized indirect effect may be “significant” in one test but not in another. The Sobel and related tests may give incorrect results in small samples. An alternative method for significance testing of product estimators of indirect effects is nonparametric bootstrapping (Preacher & Hayes, 2008), which does not assume normality. This method can also be applied to indirect effects that involve four or more variables, but bootstrapped significance tests can be inaccurate in samples that are not large. Emphasize instead whether the magnitudes of indirect effects are meaningful in your research area, not just whether or not they are statistically significant.

Now we estimate the same indirect effect (also a total effect) through covariate adjustment. Two different minimally sufficient covariate sets identify the total effect of exercise on illness and thus satisfy the back-door criterion (Rule 8.3); they are

(Hardiness) and (Stress)

Reported in Table 11.4 are values of path coefficients for exercise from two different regression analyses where hardiness is the covariate in one analysis and stress is the covariate in the other. The product estimators we just calculated for the same indi-

<sup>3</sup>Both results are based on the direct effect of fitness that controls for the other parent of illness, stress. There are other estimates of the same direct effect (see Table 11.2). This example shows that there can be multiple product estimators of the same indirect effect.

<sup>4</sup><http://quantpsy.org/sobel/sobel.htm>

**TABLE 11.4. Ordinary Least Squares Estimates of Indirect Effects (also Total Effects) of Exercise and Hardiness on Illness for a Recursive Path Model of Illness**

Indirect effect	Minimally sufficient set (covariate adjustment)		
	Product estimate	Hardiness	Stress
Exercise → Fitness → Illness	-.092 (.021) -.099	-.080 (.048) -.085	-.059 (.046) -.063
Hardiness → Stress → Illness	-.117 (.032) -.071	—	—

Note. Estimates are reported as unstandardized (standard error) standardized. Standard errors for the product estimates are Sobel standard errors.

rect effect are also reported in the table. Results across the three different estimators of the same indirect effect are similar. For example, the unstandardized coefficients range from  $-.092$  to  $-.059$ , and the standardized coefficients range from  $-.099$  to  $-.063$  (see the table). Outcomes of significance testing are inconsistent over different estimators. For example, the unstandardized product estimator is significant at the .05 level because  $z = -.092/.021 = -4.38$ ,  $p < .01$ , but neither unstandardized coefficient from covariate adjustment for the same effect is significant at the same level. (You should verify this statement working with the information in Table 11.4.) Exercise 5 asks you to interpret the results in the table for the indirect effect of hardiness on illness through stress.

In models where a cause has both a direct and an indirect effect on an outcome, a suppression effect may be indicated when the direct and indirect effects have opposite signs. This pattern is **inconsistent mediation** (MacKinnon, Krull, & Lockwood, 2000), but note that use of the term *mediation* assumes a proper design for estimating mediation. If the absolute values of inconsistent direct versus indirect effects are equal (e.g., direct =  $.30$ , indirect =  $-.30$ ), then the total effect is zero, assuming no other causal pathways between cause and outcome. But the total effect of zero in this case is due to direct versus indirect effects that exactly cancel each other out. Inconsistent mediation is contrasted with **consistent mediation**, wherein the direct and indirect effects have the same sign. See Maasen and Bakker (2001) for more information about suppression in path models. Preacher and Kelley (2011) describe measures of relative effect size when a cause has both direct and indirect effects on an outcome.

## Estimation with Maximum Likelihood

Using an SEM computer tool to estimate a path model brings with it a few conveniences. One is that values of the global fit statistics described in the next chapter are automatically computed and printed in the output. Other types of output may be optional, such as effect decompositions or graphical plots of the residuals. Some of these results can be calculated by hand, but doing so for larger models is tedious.

A drawback is that SEM computer tools do not typically inform the researcher about multiple estimators of the same effect. For example, basically all SEM computer programs estimate direct effects controlling for the parents of each endogenous variable. If the same direct effect is identified by other sets of covariates that meet the single-door criterion (Rule 8.4), the computer program will not tell you about this fact and will neither calculate nor print those other estimates. The same is true for total effects: They are typically estimated by SEM computer tools as sums of direct effects as just described and product estimators of indirect effects, not through covariate adjustment based on the back-door criterion (Rule 8.3). But knowing about graphical identification rules can help the researcher to avoid this limitation.

You may be thinking, is a researcher *required* to use an SEM computer tool to estimate a path model? The answer is no. The use of single-equation estimators, such as the 2SLS method for nonrecursive models, in path analysis combined with the application of graphical identification criteria and local fit testing, is perfectly acceptable. Doing so

is more familiar in economics and epidemiology than in the social sciences. Although SEM computer tools offer conveniences, there are also drawbacks (e.g., absence of multiple estimators of the same effect).

I used the default ML method in LISREL (Scientific Software International, 2013) to fit the path model in Figure 7.5 to a covariance matrix constructed from the data in Table 4.2. The LISREL program estimates variances for path models as  $s^2$  (i.e., the denominator is  $N - 1$ ). This makes the ML variance estimates directly comparable with those from OLS estimation for the same parameter. The analysis in LISREL converged normally to an admissible solution. Reported in Table 11.5 are the ML estimates of model parameters except for the variances and covariance of the two measured exogenous variables, exercise and hardiness. Estimates of these parameters are just the sample values (Table 4.2). Values of the unstandardized coefficients for direct effects from ML estimation in Table 11.5 are basically identical to those from OLS estimation in Table 11.2 that control for the parents of each endogenous variable. As expected, there are some slight differences in estimates of standardized direct effects or disturbance variances in ML versus OLS estimation (see Tables 11.2 and 11.3).

In this analysis of a path model, LISREL generated the standardized solution in Table 11.5 where the variances of all variables is unity (1.0). The Mplus program prints two different standardized solutions for a path model:

1. STDYX: All variables are standardized. This solution is directly comparable with the LISREL standardized solution for this example (Table 11.5).
2. STDY: All variables are standardized except for the measured exogenous variables.<sup>5</sup>

Both solutions are equally correct because there is more than one way to conceptualize standardization (Byrne, 2012b). Option STDY may be preferred for binary exogenous variables, such as gender, because change in a standard deviation metric is not very meaningful for such variables; otherwise, option STDYX for continuous predictors is fine. Check the documentation of your SEM computer to see how it derives a standardized solution for a path model. If more than one choice is available, then tell your readers which estimates are reported.

I asked LISREL to compute direct, total indirect, and total effects and to summarize these results in effect decompositions. Presented in Table 11.6 is the decomposition for the effects of exogenous variables on endogenous variables with standard errors for unstandardized results only. (Note that the direct effects in Table 11.6 match the corresponding ones in Table 11.5.) For example, hardiness is specified to have a single indirect effect on illness through stress (Figure 7.5). This sole indirect effect is also the total indirect effect because there are no other indirect causal pathways between hardiness and illness. The same indirect effect is also the total effect because there is no direct

---

<sup>5</sup>A third option in Mplus is STD, which standardizes factors only. This solution is identical to the unstandardized solution for a path model because such models have no factors.

**TABLE 11.5. Maximum Likelihood Estimates for a Recursive Path Model of Illness**

Parameter	Unstandardized	SE	Standardized
<u>Direct effects</u>			
Exercise → Fitness	.108	.013	.390
Hardiness → Stress	-.203	.044	-.230
Fitness → Illness	-.849	.160	-.253
Stress → Illness	.574	.088	.311
<u>Disturbance variances</u>			
Fitness	287.065	21.049	.848
Stress	1,062.883	77.935	.947
Illness	3,212.568	235.558	.840

Note. Standardized estimates for disturbance variances are proportions of unexplained variance. All results were computed by LISREL.

**TABLE 11.6. Decomposition for Effects of Exogenous Variables on Endogenous Variables for a Recursive Path Model of Illness**

Endogenous variables	Causal variable					
	Exercise			Hardiness		
	Unst.	SE	St.	Unst.	SE	St.
Fitness						
Direct	.108	.013	.390	0	—	0
Total indirect	0	—	0	0	—	0
Total	.108	.013	.390	0	—	0
Stress						
Direct	0	—	0	-.203	.044	-.230
Total indirect	0	—	0	0	—	0
Total	0	—	0	-.203	.044	-.230
Illness						
Direct	0	—	0	0	—	0
Total indirect	-.092	.021	-.099	-.116	.031	-.071
Total	-.092	.021	-.099	-.116	.031	-.071

Note. Unst., unstandardized; St., standardized. All results were computed by LISREL.

effect between these variables. Note that the standard errors printed by LISREL for each indirect effect in Table 11.6 match those calculated using Equation 11.1 for the Sobel test in Table 11.4. Presented in Table 11.7 is the decomposition for effects of endogenous variables on other endogenous variables. Both fitness and stress have direct effects on illness but no indirect effects through any other variables, so these direct effects are also total effects.

Not all SEM computer tools print standard errors for total indirect effects or total effects. But some programs, such as Amos and Mplus, can use the bootstrapping method to estimate standard errors for unstandardized or standardized total indirect effects or total effects. When there is a statistically significant total effect, the direct effect, total indirect effect, or both, may also be significant, but this is not guaranteed.

The standardized total effect of one variable on another estimates the part of their observed correlation due to presumed causal relations. The sum of the standardized total effects and all other noncausal associations, such as spurious associations, implied by the model equal **predicted correlations** that can be compared against the observed correlations. **Predicted covariances**, or **fitted covariances**, have the same general meaning, but they concern the unstandardized solution.

All SEM computer programs that calculate predicted correlations or covariances use matrix algebra methods. There is an older method for recursive structural models amenable to hand calculation known as **Wright's tracing rules** (Wright, 1934). The tracing rules should be understood more for their underlying principles than for their now limited utility. In these rules, a predicted correlation is the sum of all standardized causal effects and noncausal associations from all valid tracings by which the two variables are connected in the model, such that the value from each valid tracing is the product of the coefficients from the constituent paths. A **valid tracing** is defined next (Kenny, 1979):

- 
- A valid tracing means that a variable is not entered (Rule 11.1)
1. through an arrowhead and exited by the same arrowhead; nor
  2. twice in the same tracing.
- 

**TABLE 11.7. Decomposition for Effects of Endogenous Variables on Endogenous Variables for a Recursive Path Model of Illness**

Endogenous variable	Causal variable					
	Fitness			Stress		
	Unst.	SE	St.	Unst.	SE	St.
<b>Illness</b>						
Direct	-.849	.159	-.253	.574	.087	.311
Total indirect	0	—	0	0	—	0
Total	-.849	.159	-.253	.574	.087	.311

*Note.* Unst., unstandardized; St., standardized. All results were computed by LISREL.

An alternative definition comes from Chen and Pearl (2015): A valid tracing does not involve colliding arrowheads, such as



Recall that paths blocked by a collider do not convey a statistical association between the variables at either end of the path, if the collider is not included among the covariates.

Two principles follow from the tracing rule: (1) The predicted correlation for two variables connected by all possible paths in a just-identified portion of the path model will typically equal the observed (sample) value. If the whole model is just-identified, then each and every predicted correlation will exactly equal its observed counterpart. (2) But if the variables are not connected by all possible paths in an overidentified part of the model, then predicted and observed correlations may differ.

As an example of the application of the tracing rule to calculate predicted correlations with the standardized solution, look again at Figure 7.5 and find the variables hardiness and illness. There are two valid tracings between them. One is the indirect causal pathway

$$\text{Hardiness} \rightarrow \text{Stress} \rightarrow \text{Illness}$$

The product of the standardized coefficients from ML estimation in Table 11.5 for this path is

$$-.230 (.311) = -.0715$$

The other valid tracing is the noncausal path

$$\text{Hardiness} \curvearrowleft \text{Exercise} \rightarrow \text{Fitness} \rightarrow \text{Illness}$$

and the product of the standardized coefficients<sup>6</sup> for the noncausal path just listed is

$$-.030 (.390) (-.253) = .0030$$

The predicted correlation between hardiness and illness is the sum of the two products just calculated, or

$$-.0715 + .0030 = -.0685$$

The sample correlation between these two variables is  $-.16$  (see Table 4.2), so the correlation residual is

---

<sup>6</sup>Remember that the standardized estimate of the unanalyzed association between hardiness and exercise is their observed correlation,  $-.030$  (Table 4.2).

$$-.16 - (-.0685) = -.0915$$

or  $-.092$  at three-decimal accuracy. Thus, the model underpredicts the association between hardiness and illness by this amount. That the model does not perfectly reproduce the observed correlation is not surprising because there is no direct effect between these variables.

Use of the tracing rules is error-prone because it can be difficult to spot all of the valid tracings in larger models, and these rules do not apply to models with causal loops. A more complicated version of the tracing rules that include the variances of exogenous variables (including disturbances) is needed to generate predicted covariances. These are reasons to appreciate the fact that many SEM computer tools automatically calculate predicted correlations or predicted covariances for either recursive or nonrecursive path models (and other kinds of structural equation models, too).

Correlation residuals are standardized versions of **covariance residuals** or **fitted residuals**, which are differences between observed and predicted covariances. It can be difficult to interpret covariance residuals because they are not standardized. This difficulty in interpretation arises because the metric of a covariance residual depends on the scales of the two original variables that contribute to it. That is, covariance residuals for different pairs of variables are not directly comparable unless the original metric of all original variables is the same. For example, a covariance residual of, say,  $-17.50$ , for one pair of variables does not necessarily indicate greater model–data discrepancy than a covariance residual of, say,  $-5.25$ , for a different pair, if scores from all those variables are not all based on the same metric. In contrast, correlation residuals are standardized and thus are directly comparable across different pairs of observed variables regardless of their original scales.

Many SEM computer programs print **standardized residuals**, or ratios of covariance residuals over their standard errors.<sup>7</sup> In large samples, this ratio is interpreted as a  $z$  test. If this test is statistically significant, then the hypothesis that the corresponding population covariance residual is zero is rejected. This test is sensitive to sample size, which means that covariance residuals close to zero could be significant in a large sample. Likewise, a relatively large covariance residual could fail to be significant in a small sample. The interpretation of correlation residuals is not as closely bound to sample size, but there is generally no significance test for correlation residuals. Under the null hypothesis that the model perfectly fits the population covariance matrix, standardized residuals should be normally distributed, but not correlation residuals under the same hypothesis.

Some programs, such as lavaan, Mplus, and Stata, can also print **normalized residuals**, or ratios of covariance residuals over the standard error of the sample covariance, not the standard error of the difference between the sample and predicted values. (The latter is the denominator in standardized residuals.) For the same covariance residual, an absolute normalized residual is usually less than the corresponding

---

<sup>7</sup>In EQS, results labeled as “standardized residuals” are correlation residuals.

absolute standardized residual. Accordingly, normalized residuals are more conservative as significance tests than standardized residuals; that is,  $p$  values for normalized residuals are generally higher than  $p$  values for standardized residuals. For a complex latent-variable model, the computer may be unable to compute the denominator of a particular standardized residual. In this case, the corresponding normalized residual provides an alternative, but more conservative, significance test, if a significance test is really needed.

Correlation residuals for the example analysis are reported in the top portion of Table 11.8. The absolute residual for fitness and stress,  $-.133$  (shown in boldface), exceeds  $.10$ ; thus, the model does not explain very well the observed association between these two variables. Exercise 6 asks you to reproduce this correlation residual. Two other absolute correlation residuals are close to  $.10$ , including  $.082$  for fitness and hardiness and  $-.092$  for fitness and illness. (Earlier, we calculated the correlation residual just stated using the tracing rule.) Standardized residuals are reported in the bottom part of Table 11.8. The  $z$  test for the fitness–stress covariance residual is significant,  $z = 2.573$ ,  $p < .05$ . Other significant  $z$  tests (also shown in boldface) indicate that the model does not adequately explain either the observed variance of illness or its covariance with fitness and stress, but the corresponding correlation residuals are not large.

Presented in Figure 11.2 is a Q-plot of the standardized residuals in Table 11.8 generated by LISREL. In a correctly specified model, the points in a Q-plot of the standardized residuals should fall along a diagonal line, but this is clearly not the case in the figure. Taken altogether, results from local fit testing based on conditional independences (Table 11.1) and other kinds of residuals (Table 11.8, Figure 11.2) indicate that the path model in question poorly explains certain observed associations, especially for fitness and stress. We will see in the next chapter that the values of some, but not all, global fit statistics indicate problems for the same model and data. But I would conclude now that the fit of the example model is unacceptable, given the results of local fit testing and regardless of global fit testing. Models can theoretically “pass” global fit testing but still “fail” local fit testing. They say the devil is in the details, and those details regarding model fit are evaluated in local fit testing.

## FITTING MODELS TO CORRELATION MATRICES

Default ML estimation assumes the analysis of unstandardized variables. If the variables are standardized, then ML results may be inaccurate, including estimates of standard errors and model test statistics. This can happen if the model is not scale invariant (its fit depends on whether the variables are standardized or unstandardized). Whether or not a model is scale invariant is determined by a complex pattern of features, including how factors are scaled and whether certain parameter estimates are constrained to be equal (Cudeck, 1989). One symptom of scale invariance when a correlation matrix is analyzed with default ML estimation is the observation that some of the diagonal elements in a predicted correlation matrix do not equal 1.0.

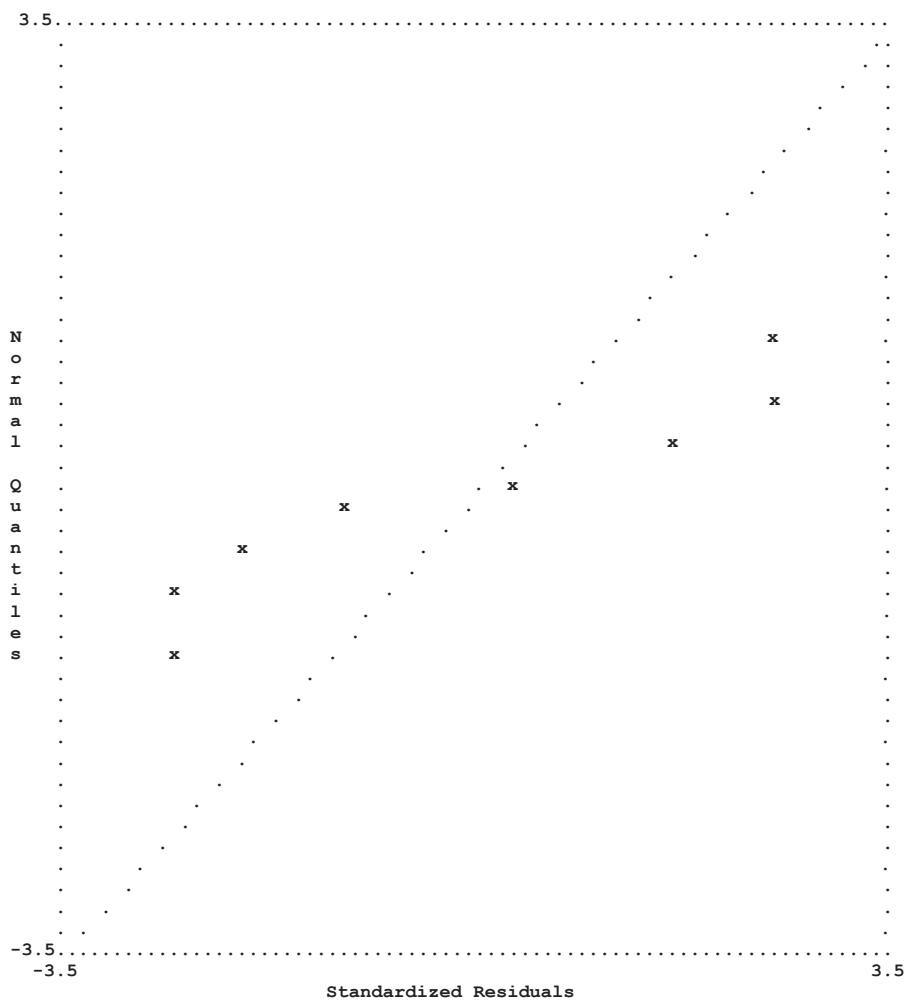
**TABLE 11.8. Correlation Residuals and Standardized Residuals for a Recursive Path Model of Illness**

Variable	1	2	3	4	5
<u>Correlation residuals</u>					
1. Exercise	0				
2. Hardiness	0	0			
3. Fitness	0	.082	0		
4. Stress	-.057	0	<b>-.133</b>	0	
5. Illness	.015	-.092	-.041	.033	.020
<u>Standardized residuals</u>					
1. Exercise	0				
2. Hardiness	0	0			
3. Fitness	0	1.714	0		
4. Stress	-1.130	0	<b>-2.573</b>	0	
5. Illness	.335	-1.951	<b>-2.539</b>	<b>2.519</b>	<b>2.333</b>

Note. The correlation residuals were computed by EQS, and the standardized residuals were computed by LISREL.

The **constrained estimation** or **constrained optimization** method can be used to correctly fit a model to a correlation matrix instead of a covariance matrix (Browne, 1982). This method involves the imposition of nonlinear constraints on certain parameter estimates to guarantee that the model is scale invariant. It can be complicated to program these constraints manually (Steiger, 2002, p. 221). Some SEM computer tools, including SEPATH and RAMONA, allow constrained estimation to be performed automatically by selecting an option. The EQS and Mplus programs can also correctly analyze correlations, but they require raw data files. Constrained estimation can be used on at least three occasions:

1. A researcher is conducting a secondary analysis based on a source wherein correlations are reported, but not standard deviations. The raw data are also not available.
2. There is a theoretical reason to impose equality constraints on standardized estimates, such as when the standardized direct effects of different causes of the same outcome are presumed to be equal. *When a covariance matrix is analyzed, equality constraints are imposed in the unstandardized solution only.*
3. A researcher wishes to report the results of significance tests for the standardized solution. This means that correct standard errors are needed. Note that Mplus and Stata automatically report correct standard errors for the standardized solution in default ML estimation; that is, constrained estimation is not needed to calculate these standard errors.



**FIGURE 11.2.** Quantile plot of standardized residuals for a recursive model of illness generated by LISREL.

## ALTERNATIVE ESTIMATORS

Standard ML estimation works fine in many applications of SEM, but you should be aware of other methods. Some of these alternatives are for continuous endogenous variables with severely non-normal distributions, but others are intended for categorical outcomes, including ordinal or nominal variables. In some disciplines, such as education or epidemiology, categorical outcomes may be analyzed as often as continuous outcomes. The methods described next are generally simultaneous, iterative, full information, and available in many SEM computer programs.

### Other Normal Theory Methods for Continuous Outcomes

Two methods for continuous endogenous variables with multivariate normal distributions include **generalized least squares** (GLS) and **unweighted least squares** (ULS). The ULS method is actually a type of OLS estimation that minimizes the sum of squared differences between sample and predicted covariances. It can generate unbiased estimates across random samples, but it is not as efficient as the ML method (Kaplan, 2009). A drawback of the ULS method is the requirement that all observed variables have the same scale (i.e., the method is neither scale free nor scale invariant). A potential advantage is that, unlike ML, the ULS method does not require a positive definite covariance matrix. It is also robust concerning start values. Thus, ULS estimation could be used to generate user-specified initial estimates for a second analysis of the same model and data but with the ML method.

The GLS method is a member of a larger family of methods known as **fully weighted least squares** (WLS) estimation, and some other methods in this family can be used for severely non-normal data. In contrast to ULS, the GLS estimator is both scale free and scale invariant, and under the assumption of multivariate normality, the GLS and ML methods are asymptotic. The GLS method generally requires less computation time and computer memory, but this potential advantage is not very meaningful today, given fast processors and abundant memory in relatively inexpensive personal computers. In general, ML is preferred over both ULS and GLS.

### Elliptical and Arbitrary Distribution Estimators for Continuous but Non-Normal Distributions

Alternatives to a corrected normal theory method, such as MLR, for continuous but non-normal outcomes are methods that do not assume multivariate normality. For example, a class of estimators based on **elliptical distribution theory** requires only symmetrical distributions (Shapiro & Browne, 1987). These methods estimate the degree of kurtosis in raw data. If all endogenous variables have a common degree of kurtosis, positive or negative, skew is allowed; otherwise, zero skew is assumed. Various elliptical distribution estimators are available in EQS.

The **arbitrary distribution function** (ADF) estimator makes *no* distributional assumptions for continuous outcomes (Browne, 1984). This is because it estimates the degree of both skew and kurtosis in the raw data. Calculations in this method are complex in part because it derives a relatively large **weight matrix** as part of its fit function. The number of rows or columns in this square matrix equals the number of observations, or  $v$  ( $v + 1)/2$ , where  $v$  is the number of observed variables and means are not analyzed. For a model with many observed variables, the weight matrix can be so large that it can be difficult for the computer to derive the inverse. For example, if  $v = 15$ , the dimension of the weight matrix is  $120 \times 120$  for a total of  $120^2 = 14,400$  elements. Very large samples are needed. Bare-bones (i.e., uninteresting) models may require 200 to 500 cases, and thousands may be needed for larger models. These requirements are

often impractical. Results of some computer simulation studies indicate that the ADF method yields overly optimistic values of fit statistics for misspecified models (Olsson et al., 2000).

## Options for Analyzing Categorical Outcomes

Endogenous variables are not always continuous. The most obvious example is a binary outcome, such as relapsed—not relapsed, which may be coded as 0 versus 1 in the data file. There are also ordered-categorical variables with three or more levels that imply a rank order, such as the following item with a Likert scale:

I am happy with my life (1 = *disagree*, 2 = *neutral*, 3 = *agree*)

The numeric scale for this item (1, 2, 3) can distinguish among only three levels of agreement. It would be nigh impossible to argue that the numbers assigned to the three response alternatives make up a continuous scale with equal intervals. There is no “golden rule” concerning the minimum number of levels required before scores on a discrete variable can be approximately normally distributed, but a score range of at least 15 points or more may be needed. Likert scales with 5 to 10 points may be favorable in terms of people’s ability to discriminate between scale values (anchors). But with ten or so anchors on a Likert scale, respondents may arbitrarily choose between adjacent points; thus, it is not practical to somehow “force” a variable with a Likert scale to become continuous by adding levels beyond 10 or so.

Numerical values associated with a particular Likert scale, such as the values “1,” “2,” and “3” for, respectively, *disagree*, *neutral*, and *agree*, are arbitrary; that is, they have no objective numerical or theoretical basis. For example, the values (-1, 0, 1) for the same response options would work just as well as would any other set of three numbers in either ascending or descending order where the distance between successive categories is the same. Accordingly, means, variances, and covariances among Likert-scale items are also arbitrary. Recall that estimation methods in SEM for continuous outcomes generally analyze covariance matrices, but such summary statistics for Likert-scale items are meaningless. Another problem is that covariances include Pearson correlations, which are for continuous variables.

Results from some computer simulation studies indicate that ML estimates may be inaccurate when analyzing ordinal or binary outcomes with relatively few levels or categories, such as five or less (DiStefano, 2002). These simulations generally assume a true population model with continuous indicators. Within generated samples, the indicators are categorized to approximate noncontinuous data. In general, ML estimates and their standard errors may both be too low when the data analyzed are from categorical indicators, and the degree of this negative bias is higher as distributions of item responses become increasingly non-normal. If there is only a single factor in the population but indicators have few categories, one-factor models are rejected too often; that is, categorization can spuriously suggest multiple factors (Bernstein & Teng, 1989). But with

more categories, such as 6–7, and symmetrical distributions, results from ML estimation may be reasonably accurate in large samples (Rhemtulla, Brosseau-Liard, & Savalei, 2012). The message of all these studies and others is that ML estimation is probably not an appropriate method for analyzing noncontinuous variables with few categories or severely asymmetrical distributions.

Summarized next are three options for analyzing models with categorical outcomes that are described in more detail later in the book:

1. The full WLS estimator does not assume any particular distributional form and thus can analyze continuous or noncontinuous variables. (The elliptical and arbitrary estimators for continuous outcomes described earlier are also members of the WLS family.) Full WLS estimation is just as computationally complex as ADF estimation, requires very large samples, and is subject to technical problems in the analysis (Finney & DiStefano, 2006), such as the failure of the computer to derive the inverse of its large weight matrix.
2. Muthén, du Toit, and Spisic (1997) describe **robust WLS** estimation, which uses simpler matrix calculations than does the full WLS method. Specifically, robust WLS methods use only the diagonal in the weight matrix from full WLS estimation. These robust methods also generate corrected standard errors and model test statistics. Other terms for robust WLS estimation include **diagonally weighted least squares** or **modified weighted least squares**. Such methods have generally performed well in computer simulation studies except when the sample size is only about  $N = 200$  or when distributions on categorical indicators are markedly skewed (Muthén et al., 1997; see also Finney & DiStefano, 2013).
3. There is a version of full-information ML estimation for categorical outcomes that analyzes raw data files and is related to methods used in logistic or probit regression. It relies on **numerical integration** to estimate response probabilities in joint multivariate distributions of the latent response variables presumed to underlie observed categorical data. Numerical integration is computationally complex, especially as the number of latent variables increases. Computer implementation may feature the use of methods that randomly sample from probability density functions, such as Markov Chain Monte Carlo (MCMC) routines, or that approximate areas in the distribution with simpler functions, such as **adaptive quadrature**. Large samples may be needed to avoid technical problems in the analysis. Another drawback is a reduction in available measures of global fit compared with other estimation methods, such as robust WLS.

## A HEALTHY PERSPECTIVE ON ESTIMATION

Segal's law states that a person with one clock always knows the time, but a person with two clocks never does. This adage speaks to the challenge of dealing with too much

information. It may also describe how newcomers to SEM may feel after learning about the availability of so many different estimators. Here is some advice about how to cope:

Use the simplest method that you understand that also makes reasonable assumptions. Think about sample size. Certain methods, such as full WLS for ordinal data, need bigger samples in order for the results to be precise. Consider alternatives, such as robust WLS, if the sample size is not large. Remember that the results can be specific to a particular method; that is, it can happen that two different estimators applied to the same model and data generate appreciably different results. This is especially true for the results of significance tests, for which small differences in estimated standard errors can make a big difference in outcome. One should therefore not make hair-splitting distinctions among  $p$  values in SEM. If two alternative estimators that yield appreciably different solutions are both viable choices, then report both sets of results instead of selecting the solution that most favors your hypotheses.

## SUMMARY

Single-equation estimation methods analyze just one endogenous variable at a time and are less efficient than simultaneous methods that estimate all free parameters at once, but simultaneous methods are more susceptible to the effects of the propagation of specification error. If all endogenous variables are continuous, the technique of multiple regression can be used to analyze recursive path models, but some type of instrumental variables regression, such as the two-stage least squares method, is needed for nonrecursive path models. The default method in most SEM computer tools is maximum likelihood estimation, which is a simultaneous, full-information, normal theory, and iterative method for continuous outcomes. Sometimes iterative estimation fails due to poor start values. When this happens, it may be necessary to specify better initial estimates in order to “help” the computer reach a converged solution. Local fit testing involves evaluating conditional independences implied by the model, deriving multiple estimates of the same parameter (when that parameter is identified by multiple sufficient sets of covariates), and inspecting the residuals. Problems in local fit testing should inform global fit testing, the subject of the next chapter.

## LEARN MORE

Bollen (2012) describes instrumental variable estimation in the social sciences, Finney and DiStefano (2013) outline estimators for nonnormal or categorical outcomes, and Lei and Wu (2012) cover estimation principles in SEM and the most widely used methods.

Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38, 37–72.

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: IAP.

Lei, P.-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–179). New York: Guilford Press.

### EXERCISES

1. Interpret the results in Table 11.2 for the direct effect of hardiness on stress.
2. Interpret the results in Table 11.2 for the direct effect of stress on illness.
3. Interpret the results in Table 11.3 for the illness outcome variable.
4. Given the coefficients in Figure 11.1, calculate and interpret the unstandardized and standardized indirect effect of hardiness on illness through stress.
5. Interpret results in Table 11.4 for the indirect effect of hardiness on illness through stress.
6. Calculate the correlation residual of  $-.133$  for the variables stress and fitness (Table 11.8), given the model in Figure 7.5 and the standardized coefficients in Table 11.5.

## **Appendix 11.A**

### Start Value Suggestions for Structural Models

These recommendations concern structural models with continuous variables, whether those models are path models or part of a structural regression model. First, think about the expected direction and magnitude of standardized direct effects. Suppose that a researcher predicts that variable  $Y$  will increase by about one-third of standard deviation, given a change of a full standard deviation in variable  $X$  while controlling for other causes. Then .30 is a reasonable guess for the standardized path coefficient, and the start value for the unstandardized coefficient for the path  $X \rightarrow Y$  would be .30 ( $SD_Y/SD_X$ ). Start values for disturbance variances can be calculated in a similar way, but now think about standardized effect sizes in terms of the proportion of explained variance (i.e.,  $R^2$ ). Suppose that a researcher predicts that all direct causes of  $Y$  will explain about 15% of its variance ( $R^2 = .15$ ). This corresponds to a proportion of unexplained variance of  $1 - .15$ , or .85. Thus, the start value for the disturbance variance would be  $.85 (s_Y^2)$ .

The start value for a disturbance covariance is the product of the square roots of the disturbance variances from the two corresponding endogenous variables and the expected Pearson correlation between their residuals. A positive correlation indicates that a common omitted cause affects both endogenous variables in the same direction, but a negative correlation says just the opposite (one variable increases, the other decreases, given change in the omitted cause). Suppose that  $Y_1$  and  $Y_2$  are two endogenous variables in a structural model and that  $D_1$  and  $D_2$  are, respectively, their disturbances. The model includes the path  $D_1 \rightsquigarrow D_2$ . The start values for the unstandardized variances of  $D_1$  and  $D_2$  are, respectively, 9.0 and 16.0. The expected correlation between the two disturbances is .40. Given these values, the start value for the unstandardized disturbance covariance would be  $.40 (9.0 \times 16.0)^{1/2}$ , or 6.30.

## 12

# Global Fit Testing

---

Introduced next are the two categories of global fit statistics in SEM, model test statistics and approximate fit indexes. Global fit statistics measure only average or overall model-data correspondence, so deciding whether to retain or reject a model should not be based solely on the values of such statistics. Hypothesis testing in SEM where alternative models are fitted to the same data is also considered. The main point in this discussion is that the choice between alternative models should be guided more by rational than statistical considerations. Related topics are power analysis in SEM and evaluation of equivalent or near-equivalent models that fit the same data just as well as the researcher's preferred model or nearly so. Chapter exercises concern the ongoing detailed example.

---

### STATE OF PRACTICE, STATE OF MIND

For at least 40 years the SEM literature has carried an ongoing discussion about the best ways to assess model fit. This is also an active research area, especially computer simulation studies. Discussion and research about this topic is likely to continue because there is no single, black-and-white statistical framework within which we can clearly distinguish correct from incorrect hypotheses in SEM. Nor is there ever likely to be such a thing.

Part of the challenge is a natural tension between what Little (2013) described as the classical and modeling schools in statistics. The *classical school* deals mainly with tests of single hypotheses, and it emphasizes explicit decision rules that should be followed by all. The *modeling school* deals with the evaluation of entire models in a context where the rules are fuzzier and less clear cut. This provides needed flexibility because few statistical models are alike and models must be adapted to different kinds of research questions. Consequently, there is greater ambiguity about rules for testing statistical models.

Another matter is the philosophical question of whether correct statistical models really exist. The recognition of this possibility reflects the view that basically all statistical models are wrong to some degree; that is, they are imperfect approximations that help researchers to structure their thinking about the target phenomenon. If that approximation is too coarse, the model will be rejected, but an overly complex model that closely mirrors the target phenomenon is also of little scientific value. Box (1976) put it like this:

Since all models are wrong the scientist cannot obtain a “correct” one by excessive elaboration. On the contrary . . . [the scientist] should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. (p. 792)

There is also the reality in SEM that fit statistics do not provide a simple yes-or-no answer to the question, should the model be retained? Various guidelines about how to interpret fit statistics as providing something like a yes-or-no answer have been developed over the years, but these rules of thumb are just that. That some of these interpretive guidelines do not apply to the whole range of structural equation models analyzed by researchers is becoming increasingly clear. It is also true that the SEM community has collectively relied too much on unsubstantiated guidelines about what fit statistics say about models.

The previous chapter and the present one describe what I believe to be a rigorous approach to modeling testing that addresses problems seen in too many SEM studies. Not all experts may agree with each and every specific detail of this approach, but I think most experts would concur that authors of SEM studies need to give their readers more information about model specification and fit (MacCallum & Austin, 2000; Shah & Goldstein, 2006). Specifically, I want you to be hardheaded in the sense that you are your model’s toughest critic. But I do not want you to be bullheaded and blindly follow the method described here as though it were *the* path to truth in SEM. Instead, use your good judgment about what makes the most sense in your research area at the same time you follow a rigorous method of hypothesis testing. To paraphrase Millsap (2007), this is SEM made difficult, not easy. The hard part is thinking for yourself in a disciplined way at every step from specification to reporting the results.

## A HEALTHY PERSPECTIVE ON GLOBAL FIT STATISTICS

Dozens of global fit statistics are described in the SEM literature, and new ones are being developed all the time. It is also true that some SEM computer tools print in their output the values of many more fit statistics than are typically reported for the analysis, which presents a few difficulties. One problem is that different fit statistics are reported over studies, and another is that different reviewers of the same manuscript may request

statistics that they know or prefer. It can therefore be difficult for researchers to decide which particular statistics and which values to report. Another possibility is “cherry picking,” or selective reporting of fit statistics with favorable values. A related problem is *fit statistic tunnel vision*, a malady among practitioners of SEM who focus so much on global fit that they overlook other crucial information, such as whether parameter estimates make sense. The cure is to closely inspect the whole computer output—including information about local fit—not just the section on fit statistics.

Before any individual fit statistic is described, it is useful to keep in mind the following limitations of basically *all* global fit statistics in SEM:

1. Values of global fit statistics indicate only average model fit. This is because fit statistics collapse many discrepancies into a single measure (Steiger, 2007). It is thus possible that some parts of the model may poorly fit the data even if the overall value of a global fit statistic seems favorable. In this case, the model may be inadequate despite the values of its fit statistics. This is why I recommend the reporting of more specific diagnostic information about model fit of the type that cannot be directly indicated by fit statistics alone. Tomarken and Waller (2003) discuss potential problems with models that seem to fit the data well based on global fit statistics.
2. Because a single statistic reflects only a particular aspect of fit, a favorable value of that statistic does not by itself indicate acceptable fit. *There is no such thing as a magical, single-number summary that says everything worth knowing about model fit.*
3. Unfortunately, little direct relation exists between values of global fit statistics and the degree or type of misspecification (Millsap, 2007). Researchers can glean relatively little about just where and by how much the model departs from the data from inspecting values of fit statistics. For example, fit statistics cannot tell you whether you have specified the correct directionalities in a structural model or the correct number of factors (3, 4, etc.) in a measurement model. Other kinds of diagnostic information from local fit testing, such as correlation residuals and standardized residuals, can speak more directly to this issue.
4. Values of global fit statistics that seem to suggest adequate fit do not also indicate that the explanatory power of the model is high for individual outcomes as measured by effect sizes such as  $R^2$ . In fact, overall model fit and  $R^2$  for individual outcomes are basically independent. For example, disturbances in structural models with perfect fit can still be large (i.e.,  $R^2$ 's are low), which means that the model accurately captures the relative lack of predictive validity in the data.
5. Fit statistics do not indicate whether the results are theoretically meaningful. For instance, the sign of some path coefficients may be unexpectedly in the opposite direction. Even if values of fit statistics seem reasonable, such anomalous results require explanation.
6. Fit statistics say little about **person-level fit**, or the degree to which the model generates accurate predictions for individual cases. Rensvold and Cheung (1999)

describe procedures to study the impact of the record from each individual case on global model fit.

Described next are the two categories of global fit statistics and the status of interpretive guidelines associated with each. Each category actually represents a different mode or contrasting way of evaluating model fit.

## MODEL TEST STATISTICS

These are the original SEM fit statistics. They are generally chi-square statistics that test the **exact-fit hypothesis** that there is no difference between the covariances predicted by the model, given the parameter estimates, and the population covariance matrix. Rejecting this hypothesis says that (1) the data contain covariance information that speak against the model, and (2) the researcher should explain model–data discrepancies that exceed those expected by sampling error.

The chi-square test just described is an **accept–support test** where the null hypothesis represents the researcher's belief that the model is correct; thus, it is *failure* to reject the null hypothesis, or the *absence* of statistical significance (e.g.,  $p \geq .05$ ), that supports the model. This logic is “backwards” from the more typical **reject–support test** where *rejecting* the null hypothesis (e.g.,  $p < .05$ ) supports the researcher's theory. Of the two, accept–support tests are logically weaker because the failure to disprove an assertion (the exact-fit hypothesis) does not prove that the assertion is true (Steiger, 2007). Low power in accept–support testing means that there is little chance of detecting a false model. This means that analyzing the model in a sample that is too small (i.e., low power) makes it more likely that the model will be retained. In reject–support testing, though, the penalty for low power due to an insufficient sample size is that the researcher's hypotheses are less likely to be supported.

Specifying lower values of  $\alpha$  in reject–support testing (e.g., .001) guards against false claims. This is because a Type I error means in this context that the researcher's theory is wrong. But in accept–support testing, we should worry more about Type II error because false claims in this context arise from *not* rejecting the null hypothesis (Steiger & Fouladi, 1997). So, insisting on low values of  $\alpha$  in accept–support testing may actually facilitate the publication of erroneous claims. Hayduk (1996) reminds us that correct models are just as likely to have a  $p$  value in the .05 region as in the .95 region. This is also true for the .25 region and the .75 region, and hence striving for correctly specified models is striving for models with  $p$  values that should ideally be considerably  $> .05$ . Thus, the convention  $\alpha = .05$  is *not* a golden rule in accept–support testing.

*The binary decision of whether to reject or not reject the exact-fit null hypothesis does not by itself determine whether to reject the model or to retain it.* One reason is power, which means that appreciable differences between model and data could be missed in small samples, but trivial differences could be flagged in large samples. At best, a failed

significance test (the exact-fit hypothesis is rejected) gives preliminary evidence against the model, just as passing the test gives preliminary support for the model. Other information from local fit testing must also be considered. In this way, a model test statistic is like a smoke detector: If the alarm sounds, there may or may not be a fire (serious model–data discrepancy), but it is prudent to treat the alarm seriously (conduct more detailed evaluation of fit).

## APPROXIMATE FIT INDEXES

**Approximate fit indexes** are *not* significance tests, so there is no binary decision about whether to reject or retain a null hypothesis just as there is no demarcation of the limits to sampling error. Instead, these indexes are intended as continuous measures of model–data correspondence. Some are scaled as “badness-of-fit” statistics where higher values indicate worse fit, but others are “goodness-of-fit” measures where higher values indicate better fit. Values of some goodness-of-fit indexes are more or less standardized so that their range is 0–1.0 where a value of 1.0 indicates the best result.

Four categories of approximate fit indexes are described next. These categories are not mutually exclusive because some indexes can be classified under more than one:

**1. Absolute fit indexes** measure how well an a priori model explains the data. That model is the researcher’s model because there is no other point of reference for an absolute fit index. Explaining the data does not by itself say that the model is adequate. This is because any misspecified model can be made to explain the data by adding free parameters to the point where no degrees of freedom remain ( $df_M = 0$ ); that is, most just-identified models will perfectly explain the observed covariances.

**2. Incremental (relative, comparative) fit indexes** measure the *relative* improvement in fit of the researcher’s model over that of a baseline model. The baseline model is usually the **independence (null) model**, which assumes covariances of zero between the endogenous variables. The null model in Mplus uses the sample covariances among the exogenous variables, but other computer programs, such as EQS and LISREL, fix the covariances between pairs of measured exogenous variables to zero, too. Check the definition of the null model in the documentation of your SEM computer tool. The assumption of zero covariances is often implausible. This is why Miles and Shevlin (2007) remarked that incremental fit indexes based on the null model “effectively say, ‘How is my model doing, compared with the worst model there is?’” (p. 870). This means that the null model is a “strawman” argument that is probably false.

**3. Parsimony-adjusted indexes** include in their formulas a correction or “penalty” for model complexity. The same penalty can be viewed as a “reward” for parsimony. Mulaik (2009b, pp. 342–345) defines model parsimony as the fewness of freely estimated parameters relative to the number of observations. Parsimony is related to  $df_M$ ,

but the two are not synonymous. This is because  $df_M$  is not a proportionate measure of the relation between observations and parameters. For example, the value of  $df_M$  can be relatively high when there are many observed variables, but it can still be proportionately small compared with the number of observations when there are many free parameters in a very complex model. Mulaik (2009b) defines the **parsimony ratio** (PR) as the ratio of  $df_M$  from the researcher's model over the degrees of freedom from the null model. The ideal (most parsimonious) model would have a PR of 1.0, which says that the *model* has just as many degrees of freedom as there are available in the *data* (as per the null model). Such models are potentially more disconfirmable with the data than models where the value of the PR is less than 1.0.

4. **Predictive fit indexes** estimate model fit in *hypothetical* replication samples of the same size and randomly drawn from the same population as the original sample. Thus, these indexes may be seen as population based rather than sample based. They may also correct for model degrees of freedom or sample size. There is a specific context for predictive fit indexes that is described later in this chapter, but most applications of SEM do not fall under it.

Formulas of some approximate fit indexes include model test statistics. There is a similar relation in more standard data analyses between test statistics and measures of effect size: many effect sizes can be expressed as functions of test statistics and vice versa (Kline, 2013a). The relation between model test statistics and approximate fit indexes means that both are based on the same distributional assumptions. If these assumptions are untenable, then values of both the approximate fit index and the corresponding test statistic (and its *p* value) may be inaccurate.

A natural question about continuous approximate fit indexes concerns the range of values indicating "acceptable" model fit. There is no simple answer to this question because there is no direct correspondence between values of approximate fit indexes and the seriousness or type of specification error. Most interpretive guidelines originate from computer simulation studies from the 1980s–1990s about the behavior of approximate fit indexes under varying data and model conditions for measurement models of the kind analyzed in the technique of CFA. Gerbing and Anderson (1993) review many of these early studies, and more recent examples include Hu and Bentler (1998) and Marsh, Balla, and Hau (1996). Based on these findings as well as their own simulation studies, Hu and Bentler (1999) proposed a set of thresholds for approximate fit indexes that are the most widely known and cited in the literature. Whether these thresholds are accurate is a critical question.

Hu and Bentler (1999) never intended their rules of thumb for approximate fit indexes to be treated as anything other than just that. One reason is that it is impossible in Monte Carlo studies to evaluate the whole range of models and data analyzed in real studies. Another is that seriously misspecified models are not typically studied in computer simulations. Instead, authors of such studies tend to impose relatively minor

specification errors on known measurement models (e.g., a factor variance is misspecified in generated samples). The case of a more serious specification error, such as the wrong number of factors, may not be studied at all. Finally, thresholds for approximate fit indexes are associated with maximum likelihood (ML) estimation for continuous outcomes, so they may not generalize to other methods or when analyzing categorical outcomes.

Results of several more recent simulation studies cast doubt on the generality of thresholds for approximate fit indexes. Marsh, Hau, and Wen (2004) found that the accuracy of thresholds depends on the particular misspecified model studied in computer simulations. This was especially true for models with approximate fit index values very close to their suggested thresholds. Yuan (2005) studied properties of approximate fit indexes when distributional assumptions were violated. He found that (1) expected values of these indexes had little relation to thresholds; and (2) shapes of their distributions varied as complex functions of sample size, model size, and the degree of misspecification. For measurement models of a type evaluated in personality research, Beauducel and Wittman (2005) found that (1) threshold accuracy was affected by the relative sizes of pattern coefficients and whether measurement was specified as unidimensional versus multidimensional, and (2) different approximate fit indexes did not generally “agree” in their results for the same model.

Given results of the kind just summarized, Barrett (2007) suggests a ban on approximate fit indexes. Hayduk et al. (2007) argue that thresholds for such indexes are so untrustworthy that only model test statistics (and their  $df$  and  $p$  values) should be reported. Relying on thresholds for approximate fit indexes has the consequence that they are treated as though they generate two qualitative outcomes, “acceptable” versus “unacceptable” fit (Markland, 2007). But values of approximate fit indexes reported as point estimates (without confidence intervals) ignore (disregard) sampling error, and values of these indexes for the same model vary across samples. Others argue that approximate fit indexes have a limited role (Mulaik, 2009b), but the consensus is that blind reliance on thresholds is no longer up to standard.

## **RECOMMENDED APPROACH TO FIT EVALUATION**

The method outlined next calls on researchers to report more specific information about model fit than has been true of recent practice. The steps are as follows:

1. If you use a simultaneous estimation method, report the chi-square with its degrees of freedom and  $p$  value. If the model fails the exact-fit test, then tentatively reject the model. Next, diagnose both the magnitude and possible sources of misfit (conduct local fit testing). The rationale is to detect statistically significant but slight model–data discrepancies that explain the failure. This is most likely to happen in a large sample. But if the model passes the exact-fit test, you still have to conduct local fit testing. The rationale is to detect model–data discrepancies that are not statistically

significant but still great enough to cast doubt on the model. This is most likely in a small sample.

2. Report a matrix of residuals, such as correlation residuals, or at least describe the pattern of residuals for a big model. This includes the locations of larger residuals and their signs. Look for patterns that may be of diagnostic value in understanding how the model may be misspecified. *Any report of the results without information about the residuals is incomplete.*

3. If you report values of approximate fit indexes, then include those for the minimal set described next. *But do not try to justify retaining the model by depending solely on discredited thresholds for such fit statistics.* This is especially true if the model failed the exact-fit test and the pattern of residuals suggests a particular kind of specification error that is not trivial.

4. If you respecify the initial model, then explain the rationale for doing so. You should also explain the role that diagnostic statistics, such as residuals, played in the respecification. In other words, point out the connection between the numerical results for the model, relevant theory, and modifications to the original model. If you retain a respecified model that still fails the exact-fit test, then demonstrate that model–data discrepancies are truly slight; otherwise, you have failed to show that there is no appreciable covariance evidence against the model.

5. If no model is retained, then your skills as a scholar are needed to explain the implications for the theory tested in your analysis. At the end of the day, regardless of whether or not you have retained a model, the real honor comes from following to the best of your ability a thorough testing process to its logical end. The poet Ralph Waldo Emerson put it this way: The reward of a thing well done is to have done it.

Listed next is a minimum set of fit statistics that should be reported whenever it is possible to do so. It consists of a model test statistic and three approximate fit indexes:

1. Model chi-square with its degrees of freedom and  $p$  value.
2. Steiger–Lind Root Mean Square Error of Approximation (RMSEA; Steiger, 1990) and its 90% confidence interval.
3. Bentler Comparative Fit Index (CFI; Bentler, 1990).
4. Standardized Root Mean Square Residual (SRMR).

There are many other approximate fit indexes in SEM—so many that they could not all be described here in any real detail. Some older indexes have problems, so it would do little good to describe them because I could not recommend their use; see Kaplan (2009, chap. 6) or Mulaik (2009b, chap. 15) for more information. Besides, you do not really need more than the minimal set just listed, especially if you properly emphasize local fit testing over global fit testing.

## MODEL CHI-SQUARE

Depending on the particular SEM computer tool, the model chi-square is calculated in one of the two different ways listed next:

$$(N - 1) F_{ML} \quad \text{or} \quad N(F_{ML}) \quad (12.1)$$

where  $F_{ML}$  is the value of the fit function minimized in ML estimation. In very large samples, the two products in Equation 12.1 are asymptotic, and, assuming multivariate normality, both products follow a central chi-square distribution with degrees of freedom equal to that of the model, or  $df_M$ . For this reason, either expression in Equation 12.1 is called the **minimum fit function chi-square** or the **likelihood ratio chi-square**. It is designated here as  $\chi^2_M$  for the chi-square for the researcher's model in the ML method. The value of  $\chi^2_M$  for a just-identified model generally equals zero, but technically it is not defined for models with no degrees of freedom. If  $\chi^2_M = 0$ , the model perfectly fits the data (each observed covariance equals its predicted counterpart). If the fit of an over-identified model that is not correctly specified becomes increasingly worse, the value of  $\chi^2_M$  increases, so it is a badness-of-fit statistic.

In large samples and assuming multivariate normality, the hypothesis of exact fit is tested by  $\chi^2_M$  for an overidentified model. For a correctly specified model analyzed over random samples, the expected value of  $\chi^2_M$  equals that of its degrees of freedom,  $df_M$ , regardless of the sample size. This means that in about half the random samples drawn from a population where the model has perfect fit,  $\chi^2_M \leq df_M$ . It is also true that in 19 of 20 random samples the  $p$  value for  $\chi^2_M(df_M)$  will be  $\geq .05$  regardless of the sample size. Thus, the exact-fit hypothesis will be rejected for correct models in less than 1 out of 20 samples when testing at the .05 level.

Another way of looking at  $\chi^2_M$  is that it tests the difference in fit between a given overidentified model and whatever unspecified model would predict a covariance matrix that perfectly corresponds to the data covariance matrix. Suppose for an overidentified model that  $\chi^2_M > 0$  and  $df_M = 5$ . Adding five more free parameters to this model would make it just-identified—thereby making its covariance implications perfectly match the data covariance matrix, even if that model were not correctly specified—and reduce both  $\chi^2_M$  and  $df_M$  to zero.

Model and data are consistent within the limits of sampling error if the exact-fit hypothesis is retained, but whether the model is actually *correct* is unknown. It could be seriously misspecified or it could be one of potentially many other equivalent or near-equivalent models that imply covariance matrices identical or very similar to the observed data (Hayduk et al., 2007). This is why Markland (2007) cautioned that “even a model with a non-significant chi-square test needs to have a serious health warning attached to it” (p. 853). One reason is that  $\chi^2_M$  tends to miss a single large covariance residual or a pattern of smaller but systematic residuals that indicate a problem with the model. Another is that the value of  $\chi^2_M$  is easily reduced by adding free parameters,

which generates more complex models. If those parameters are added without justification, the resulting overparameterized model may have little scientific value.

The value of  $\chi^2_M$  can also be affected by:

**1. Multivariate non-normality.** Depending on the pattern and severity of non-normality, the value of  $\chi^2_M$  can be either increased so that model fit appears worse than it really is or decreased so that model fit looks better than it really is (Hayduk et al., 2007). This is why it is so important to screen your data for severe non-normality when using a normal theory method. You can also report a corrected chi-square that controls for non-normality.

**2. Correlation size.** Bigger correlations among observed variables generally lead to higher values of  $\chi^2_M$  for incorrect models. This happens because larger correlations allow greater potential discrepancies between observed and predicted correlations (and covariances, too).

**3. Unique variance.** Analyzing variables with high proportions of unique variance—which could be due to score unreliability—results in loss of statistical power. This property of  $\chi^2_M$  could potentially “reward” the selection of measures with poor psychometrics because low power in accept–support testing favors the researcher’s model. If there is low power to detect problems, but the model still fails the chi-square test, then those problems may be serious. Thus, the researcher should pay especially careful attention to local fit testing in this case.

**4. Sample size.** For incorrect models that do not imply covariance matrices similar to the sample matrix, the value of  $\chi^2_M$  tends to increase along with the sample size. In “typical” sample sizes in SEM studies ( $N = 200\text{--}300$ ), failing the chi-square test *may* signal a problem serious enough to reject the model. In very large samples, though, much smaller model–data discrepancies could lead to rejection of the exact-fit hypothesis, but you won’t know whether this is true without inspecting the residuals.

Results of simulation studies by Cheung and Rensvold (2002) and Meade, Johnson, and Braddy (2008) suggest that  $\chi^2_M$  is overly sensitive to sample size when testing whether the same factor structure holds across different groups, that is, whether a measurement model is invariant. But values of some approximate fit indexes are less affected by sample size. Mooijaart and Satorra (2009) remind us that  $\chi^2_M$  is generally insensitive to the presence of interaction (moderator) effects because the distribution of  $\chi^2_M$  may not be distorted even when there is severe misspecification of interaction effects. Consequently, they cautioned against concluding that if a model with no product terms that represent interaction effects passes the chi-square test, then the underlying model must truly be without interaction. Approximate fit indexes based on  $\chi^2_M$  would be just as insensitive to interaction misspecification.

Because of the increasing power of  $\chi^2_M$  to detect model–data discrepancy with increasing sample size, it was once common practice for researchers to (1) ignore a

failed chi-square test but then (2) refer to threshold values for approximate fit indexes in order to justify retaining the model. Many published models had statistically significant  $\chi^2_M$  values, but authors paid little attention to this fact while at the same time they interpreted the results of significance tests of individual parameter estimates (Markland, 2007). This poor practice highlights a weird contradiction where researchers ignore covariance evidence against the model (i.e.,  $\chi^2_M$ ) but then get all excited about  $z$  tests of individual parameters that are significant. This form of confirmation bias usually favors the researcher's hypotheses (i.e., the model). It is also a form of evidence disrespect that fosters the publication of false claims (Hayduk, 2014b).

A brief mention of a statistic known as the **normed chi-square** is needed mainly to discourage you from ever using it. In an attempt to reduce the sensitivity of the model chi-square to sample size, some researchers in the past divided this statistic by its expected value, or  $\chi^2_M/df_M$ , which reduced the value of this ratio for  $df_M > 1$  compared with  $\chi^2_M$ . There are three problems with the normed chi-square: (1)  $\chi^2_M$  is sensitive to sample size only for incorrect models; (2)  $df_M$  has nothing to do with sample size; and (3) there were really never any "acceptable" clear-cut guidelines about maximum values of the normed chi-square (e.g.,  $< 2.0?$ — $< 3.0?$ ). Because there is little statistical or logical foundation for the normed chi-square, it should have no role in global fit testing.

### Chi-Squares for Other Estimators

The statistic  $\chi^2_M$  is associated with ML estimation, but model chi-squares have the same general form under different methods. For example, the symbol  $\chi^2_{ADF}$  designates the model chi-square in Browne's (1984) asymptotically distribution free (ADF) estimator. It equals the product of either  $N - 1$  or  $N$  and  $F_{ADF}$ , the value of the fit function minimized in the ADF method. Its degrees of freedom are  $df_M$ .

The test statistic in robust maximum likelihood (MLR) is generally the **Satorra-Bentler scaled chi-square** (Satorra & Bentler, 1994), designated as  $\chi^2_{SB}$ . It is computed by applying a **scaling correction factor**,  $c$ , to the unscaled model chi-square. The specific relation is

$$\chi^2_{SB} = \frac{\chi^2_M}{c} \quad (12.2)$$

The value of  $c$  reflects the average kurtosis in the raw data. Distributions of  $\chi^2_{SB}$  over random samples only approximate central chi-square distributions but have asymptotically correct means (i.e., the expected value is  $df_M$ ).

The **Satorra-Bentler adjusted chi-square** is based on a different scaling correction factor such that its distributions more closely follow central chi-square distributions with asymptotically correct means and variances. This statistic has an estimated degree of freedom that is typically a fractional number (e.g., 15.75) that is used to compute the  $p$  value. The Satorra-Bentler scaled chi-square is more widely known and reported than the adjusted version. Special comment is needed for version 9 of LISREL (Scientific

Software International, 2013), which prints up to five different model chi-squares for the same model and data—see Appendix 12.A.

## RMSEA

The RMSEA, designated here as  $\hat{\epsilon}$  (lowercase Greek letter epsilon), is an absolute fit index scaled as a badness-of-fit statistic where a value of zero indicates the best result. It also generally “rewards” models with more degrees of freedom or models analyzed in larger samples with lower values of  $\hat{\epsilon}$ . It is usually reported in computer output with the 90% confidence interval

$$[\hat{\epsilon}_L, \hat{\epsilon}_U] \quad (12.3)$$

where  $\hat{\epsilon}_L$  is the lower-bound estimate of  $\epsilon$ , the parameter estimated by  $\hat{\epsilon}$ , and  $\hat{\epsilon}_U$  is the upper-bound estimate. If  $\hat{\epsilon} = 0$ , then  $\hat{\epsilon}_L = 0$  and the whole interval is a one-sided confidence interval where  $\hat{\epsilon}_U > \hat{\epsilon}$ . This explains why the confidence level is 90% instead of the more typical 95%, the conventional level for two-sided confidence intervals. If  $\hat{\epsilon} > 0$ , then  $\hat{\epsilon}_L \geq 0$  and the value of  $\hat{\epsilon}$  does not typically fall at the exact center of interval in Equation 12.3.

The model chi-square measures departure from exact or perfect fit, but  $\hat{\epsilon}$  measures departure from close or approximate fit. Thus,  $\hat{\epsilon} = 0$  says only that model–data discrepancy fails to exceed the limit of close fit, not that fit is perfect. The limit of close fit is defined as follows:

$$\hat{\Delta}_M = \max(0, \chi_M^2 - df_M) \quad (12.4)$$

which equals the maximum of either zero or the difference between the model chi-square and its expected value (i.e.,  $df_M$ ) in a central chi-square distribution that assumes perfect fit. If

$$\chi_M^2 \leq df_M$$

then  $\hat{\Delta}_M = 0$ , and there is no departure from close fit; otherwise,  $\hat{\Delta}_M > 0$ , which indicates that the limit of close fit is exceeded by the amount that the value of  $\chi_M^2$  is greater than that of  $df_M$ . The term  $\hat{\Delta}_M$  estimates the noncentrality parameter  $\Delta_M$  in the noncentral chi-square distribution

$$\chi_M^2(df_M, \Delta_M) \quad (12.5)$$

that describes the distribution of  $\hat{\epsilon}$  whenever the model does *not* perfectly fit the population covariance matrix (i.e.,  $\Delta_M > 0$ ). But if that fit is perfect (i.e.,  $\Delta_M = 0$ ), then  $\hat{\epsilon}$  follows a central chi-square distribution with degrees of freedom that equal  $df_M$ .

If  $\hat{\Delta}_M > 0$ , then  $\hat{\epsilon} > 0$  and the formula is

$$\hat{\epsilon} = \sqrt{\frac{\hat{\Delta}_M}{df_M(N-1)}} \quad (12.6)$$

Note in Equation 12.6 that a greater model degrees of freedom or larger sample size reduces the value of  $\hat{\epsilon}$ , but the effect of correcting for  $df_M$  diminishes as  $N$  becomes increasingly large; see Mulaik (2009b, pp. 339–341) for more information. The original threshold is from Browne and Cudeck (1993), who suggested that  $\hat{\epsilon} \leq .05$  may indicate “good fit.” But results of computer simulations by Chen, Curran, Bollen, and Paxton (2008) indicated little support for a universal threshold of .05 (or any other value) regardless of whether  $\hat{\epsilon}$  is used alone or jointly with its 90% confidence interval. Browne and Cudeck (1993) also suggested that  $\hat{\epsilon} \geq .10$  or so may indicate a serious problem, but there is no guarantee. Described in Topic Box 12.1 are options for significance testing based on the RMSEA. Some of these tests are part of power analysis in SEM, a topic addressed in a later section of this chapter.

### TOPIC BOX 12.1

#### Significance Testing Based on the RMSEA

If the lower bound of the 90% confidence interval equals zero ( $\hat{\epsilon}_L = 0$ ), the model chi-square test will *not* reject at the .05 level the exact-fit hypothesis

$$H_0: \epsilon_0 = 0$$

The  $p$  value for an accept–support test of the exact-fit hypothesis equals that of  $\chi^2_M$  ( $df_M$ ) for the same model and data.

Some SEM computer tools also print  $p$  values for the test of the **close-fit hypothesis**, or the one-sided null hypothesis

$$H_0: \epsilon_0 \leq .05$$

Failure to reject the close-fit hypothesis, such as  $p_{\epsilon_0 \leq .05} = .15$  when  $\alpha = .05$ , supports the researcher’s model; otherwise, a model could fail the more stringent exact-fit test but pass the less demanding close-fit test. Hayduk, Pazderka-Robinson, Cummings, Levers, and Beres (2005) describe such models as **close-yet-failing models**. Passing the close-fit test does not justify ignoring a failed exact-fit test. As noted by Hayduk (2014a, p. 920), “close fit, or a small amount of covariance ill fit, does not confidently report that the model is close to being properly causally specified” for the reasons explained in the chapter.

The **not-close-fit hypothesis** is an inversion of the close-fit hypothesis. It is expressed as

$$H_0: \hat{\epsilon}_0 \geq .05$$

If the upper bound of the 90% confidence interval is less than .05 ( $\hat{\epsilon}_U < .05$ ), then the hypothesis of not close fit is rejected, which supports the researcher's model. This means that (1) the test of the not-close-fit hypothesis is a reject–support test, and (2) low power works against the researcher's model. Greater power here implies a higher probability of detecting a reasonably correct model, or at least one that predicts a covariance matrix that approximates the sample data matrix within the limits of close fit.

If the upper bound of the 90% confidence interval equals or exceeds a value that might indicate "poor fit," such as  $\hat{\epsilon}_U \geq .10$ , then the model may warrant less confidence. For example, the test of the **poor-fit hypothesis**

$$H_0: \hat{\epsilon}_0 \geq 10$$

is a reject–support test of whether the researcher's model is just as bad as or even worse than a poor-fitting population model. The test of the poor-fit hypothesis can serve as a kind of reality check against the test of the close-fit hypothesis. The tougher exact-fit test can serve this purpose, too.

Do not expect results for these various significance tests to be consistent for the same model and data. Based on the results listed next for the detailed example (see Table 12.1),

$$\begin{aligned} \chi^2_M(5) &= 11.107, p = 0.049 \\ \hat{\epsilon} &= .057, 90\% \text{ CI } [.003, .103], P_{\hat{\epsilon}_0 \leq .05} = .336 \end{aligned}$$

we can say for the .05 level that the recursive model of illness in Figure 7.5

1. fails the exact-fit test because  $p < .05$  and  $\hat{\epsilon}_L > 0$ ;
2. passes the close-fit test because  $P_{\hat{\epsilon}_0 \leq .05} > .05$ ;
3. fails the not-close-fit test because  $\hat{\epsilon}_U > .05$ ; and
4. fails the poor-fit test because  $\hat{\epsilon}_U > .10$ .

The only way to resolve these apparent contradictions in significance testing is to consider the entire 90% confidence interval, which says for this example that the point estimate of  $\hat{\epsilon} = .057$  is so imprecise that it is just as consistent with the close-fit hypothesis as it is with the poor-fit hypothesis. A larger sample may be needed to obtain more precise results.

Limitations of the RMSEA are summarized next:

1. Interpretation of  $\hat{\epsilon}$  (and the lower and upper bounds of its confidence interval) relative to thresholds requires that it follows noncentral chi-square distributions. There is evidence that casts doubt on this assumption. Olsson, Foss, and Breivik (2004) found in computer simulation studies that empirical distributions of  $\hat{\epsilon}$  for smaller models with relatively less specification error generally followed noncentral chi-square distributions. Otherwise, the empirical distributions did *not* typically follow noncentral chi-square distributions, especially for models with more specification error. These results and others (Chen et al., 2008; Yuan, 2005; Yuan, Hayashi, & Bentler, 2007) question the generality of thresholds for the RMSEA.
2. Nevitt and Hancock (2000) evaluated in Monte Carlo studies the performance of robust forms of the RMSEA corrected for non-normality, one of which is based on the Satorra–Bentler scaled chi-square. Under conditions of non-normality, this robust RMSEA statistic generally outperformed the uncorrected version (Equation 12.6).
3. Breivik and Olsson (2001) found in simulation studies that the RMSEA tends to impose a harsher penalty on smaller models with relatively few variables. This is because smaller models may have relatively few degrees of freedom, but larger models have more “room” for higher  $df_M$  values.

## CFI

The Bentler CFI is an incremental fit index that is also a goodness-of-fit statistic. Its values range from 0 to 1.0 where 1.0 is the best result. The CFI compares the amount of departure from close fit for the researcher's model against that of the independence (null) model. For models where  $\chi^2_M \leq df_M$  (i.e.,  $\hat{\Delta}_M = 0$ ), then  $CFI = 1.0$  (no departure from close fit); otherwise, the formula is

$$CFI = 1 - \frac{\hat{\Delta}_M}{\hat{\Delta}_B} \quad (12.7)$$

where  $\hat{\Delta}_B$  is defined for baseline model as

$$\hat{\Delta}_B = \max(0, \chi^2_B - df_B) \quad (12.8)$$

where  $\chi^2_B$  and  $df_B$  are, respectively, the chi-square and degrees of freedom for the baseline model. The value of  $\chi^2_B$  is often relatively large, and the pattern  $\chi^2_B \leq df_B$  is not usually seen in real data. The result  $CFI = .90$ , for example, says that the fit of the researcher's model is about .90, or 90% better than that of the baseline model.

The CFI is a rescaled version of the Relative Noncentrality Index (McDonald & Marsh, 1990), the values of which can fall outside the 0–1.0 range. A related statistic is the **Tucker–Lewis index** (TLI; Tucker & Lewis, 1973), also called the **non-normed fit**

**index** (NNFI; Bentler & Bonett, 1980). It controls for  $df_M$  from the researcher's model and also for  $df_B$  from the baseline model. Values of the NNFI can also exceed 1.0. The TLI/NNFI imposes a greater relative penalty for model complexity than the CFI, but only one of these two fit statistics should be reported because their values are highly correlated (Kenny, 2014a).

All incremental fit indexes have been criticized when the baseline model is the independence or null model, which is almost always true. The assumption of zero covariances is improbable in many, if not most, studies. It is possible to specify a different, more plausible baseline model than the independence model. For example, Widaman and Thompson (2003) describe a **longitudinal independence model** for panel designs with repeated measures. In this baseline model, all covariances are fixed to zero, but the means and variances of repeated measures variables at time 1 are constrained to equal their counterparts at time 2 (and at time 3, and so on). If observed means or variances change appreciably over time, there is more potential information to be extracted by the researcher's model. But if variances and means do not change much, there is less information to be recovered by the researcher's model (Little, 2013). Calculation of the CFI for a "special" baseline model (i.e., not the default null model) is accomplished by specifying that model in program syntax, fitting it to the data matrix, then recording the values of the model's chi-square and degrees of freedom, and next computing the value of the CFI by hand (Equation 12.7).

Hu and Bentler (1999) suggested using the CFI together with an index based on the correlation residuals described next, the SRMR. Their rationale was that the CFI seemed to be most sensitive to misspecified pattern coefficients, whereas the SRMR seemed to be most sensitive to misspecified factor covariances when testing CFA measurement models. Their **combination rule** for concluding "acceptable fit" based on these indexes was  $CFI \geq .95$  and  $SRMR \leq .08$ . This combination rule was not supported in Monte Carlo studies by Fan and Sivo (2005), who suggested that the original Hu and Bentler (1999) findings about the CFI and SRMR for factor analysis models were artifacts. Results of other simulation studies also do not support the respective thresholds just listed (Yuan, 2005).

## SRMR

The SRMR is an absolute fit index that is a badness-of-fit statistic. It is a standardized version of the **root mean square residual** (RMR), which is a measure of the mean absolute covariance residual. Perfect model fit is indicated by  $RMR = 0$ , and increasingly higher values indicate worse fit. A problem with the RMR is that because it is computed with unstandardized variables, its value and range depend on the metrics of the observed variables. If these metrics are all different, it can be difficult to interpret a given value of the RMR. The SRMR is computed as the square root of the average squared covariance residual in a standardized metric. It is thus a measure of the mean absolute correlation residual, the overall difference between the observed and predicted correlations. Values

of SRMR > .10 may indicate poor fit, but the matrix of correlation residuals should be inspected in any event.

## TIPS FOR INSPECTING RESIDUALS

Correlation residuals are easier to interpret than covariance residuals, and absolute correlation residuals > .10 deserve special attention as *possible* evidence for poor local fit. But you should know that there is actually no dependable or trustworthy connection between the size of the residuals and the type or degree of model misspecification. For example, the degree of misspecification indicated by a low correlation residual may be slight and yet may be severe.

One reason is that values of residuals and other diagnostic statistics, including modification indexes (described later), are themselves affected by misspecification. An analogy in medicine would be a diagnostic test for some disease that is less accurate in patients who actually have that disease. This problem in SEM is a consequence of error propagation when some parts of the model are incorrectly specified. This problem may be even greater when the estimation method is simultaneous (e.g., ML) rather than single equation (e.g., 2SLS). But we do not generally know in advance which parts of the model are incorrect, so it can be difficult to understand exactly what the residuals are telling us.

Inspecting the *pattern* of residuals can sometimes be helpful. Suppose that a pair of variables X and Y where  $r_{XY} > 0$  are connected by indirect pathways only in a structural model. The residual for that pair is positive, which says that the model underpredicts their observed association. In this case, the hypothesis of no direct effect between X and Y may be cast in doubt, and a possible respecification is to add a direct effect between them. Another possibility consistent with the same positive residual is to specify a disturbance correlation. But just which type of effect to add to the model ( $\rightarrow$  vs.  $\curvearrowright$ ) or their directionalities (e.g., X causes Y vs. Y causes X) are not things that residuals can tell you. Just as there is no magic in fit statistics, there is also none in diagnostic statistics, at least none that would relieve researchers from the burden of having to think long and hard about respecification.

## GLOBAL FIT STATISTICS FOR THE DETAILED EXAMPLE

Reported in Table 12.1 are values of global fit statistics computed by LISREL (Scientific Software International, 2013) under default ML estimation for the recursive path model of illness described in the previous chapter (see also Figure 7.5 and Table 4.2). The sample size is  $N = 373$ , and the value of the minimized fit function is  $F_{ML} = .0297787$ . The chi-square in LISREL is computed as

$$\chi^2_M(5) = 373 (.0297787) = 11.107, p = .049$$

The LISREL program also printed the chi-square under Browne's (1984) ADF estimator that assumes normality:

$$\chi^2_{\text{ADF}}(5) = 11.103, p = .049$$

Values of the two test statistics just listed are essentially identical. The model just fails the exact-fit test at the .05 level ( $p = .049$ ), but it passes the less demanding close-fit test at the same level because  $p_{\epsilon_0 \leq .05} = .336$  (see Table 12.1).

Values of approximate fit indexes in Table 12.1 also suggest a mixed picture. The value of the RMSEA is .057, which does not seem terrible, but the upper bound of its 90% confidence interval, .103, is so high that the poor-fit hypothesis cannot be rejected. The fit of the analyzed path model (Figure 7.5) is about 96.2% better than that of the independence model ( $\text{CFI} = .962$ ), which in LISREL assumes zero population covariances for all pairs of measured variables. Neither the value of the CFI nor that of the SRMR, which equals .051 (see the table), indicates a glaring problem. Exercises 1–3 ask you to calculate some of the results in Table 12.1, and for Exercise 4 you should rerun the analysis but specify a larger sample size ( $N = 5,000$ ) in order to observe the effect of increasing just the sample size on values of global fit statistics.

In Chapter 11, we conducted local fit testing for the same model and data, including examination of the residuals (see Tables 11.1, 11.8). Thus, we already know the specific

**TABLE 12.1. Values of Fit Statistics for a Recursive Path Model of Illness**

N	373
$F_{\text{ML}}$	.0297787
$df_M$	5
<u>Model test statistics</u>	
$\chi^2_M$	11.107, $p = .049$
Normal theory $\chi^2_{\text{ADF}}$	11.103, $p = .049$
$p_{\epsilon_0 \leq .05}$	.336
<u>Approximate fit indexes</u>	
RMSEA [90% CI]	.057 [.003, .103]
CFI	.962
SRMR	.051
<u>Independence model</u>	
$\chi^2_B$	172.289
$df_B$	10

*Note.* CI, confidence interval. All results were computed by LISREL.

nature of fit problems, including the fact that the model poorly explains the observed association between fitness and stress, among other problems. Values of some global fit statistics in Table 12.1 signal poor average fit (e.g.,  $\chi^2_M$ ,  $\hat{\epsilon}_U$ ), but not others (e.g.,  $\hat{\epsilon}$ , CFI, SRMR). This example illustrates why local fit testing is so crucial in SEM.

## TESTING HIERARCHICAL MODELS

Next we consider how to test hypotheses about **hierarchical (nested) models** with the same data. Two models are hierarchical or nested if one is a proper subset of the other. For example, if a free parameter is dropped from model 1 (i.e., the parameter is replaced with a fixed value that is usually zero) to form model 2, the two models are hierarchically related (model 2 is nested under model 1). This is the most frequent context for model comparison in SEM.

### Model Trimming and Building

Hierarchical models are compared in one of two different ways. In **model trimming**, the researcher begins with a more complicated model and then simplifies it by eliminating free parameters (paths). This is done by specifying that at least one path that was previously a freely estimated parameter is now constrained to equal zero. The starting point for **model building** is a simpler model to which paths are added (i.e., the initial model must be overidentified). Typically, at least one previously fixed-to-zero path is specified as a free parameter. As a model is trimmed, its overall fit to the data usually becomes worse ( $\chi^2_M$  increases); similarly, model fit generally improves as paths are added ( $\chi^2_M$  decreases). The goal of both trimming and building is to find the model that has a properly specified covariance structure and is theoretically justifiable.

Models can be trimmed or built according to one of two different standards, theoretical or empirical. The first represents tests of specific, *a priori* hypotheses. Suppose that a path model contains the paths

$$X \rightarrow Y_2 \quad \text{and} \quad X \rightarrow Y_1 \rightarrow Y_2$$

If the researcher believes that the effect of  $X$  on  $Y_2$  is purely indirect through  $Y_1$ , then he or she can test this hypothesis by constraining the coefficient for the path  $X \rightarrow Y_2$  to zero. If the fit of the model so constrained is not appreciably worse than the one with  $X \rightarrow Y_2$  as a free parameter, the hypothesis about a purely indirect effect is supported, assuming that the corresponding directionality specifications are correct. The main point, however, is that respecification of a model to test hierarchical versions of it is guided by specific hypotheses.

This is not the case for empirically based respecification, in which free parameters are deleted or added according to statistical criteria. For example, if the sole basis for

trimming paths is that their coefficients are not statistically significant, then respecification is guided by purely empirical considerations. The distinction between theoretically or empirically based respecification has implications for interpreting the results of model trimming or building, which are considered after a model comparison test statistic is introduced.

### Chi-Square Difference Test

The **chi-square difference statistic**,  $\chi^2_D$ , can be used to test the statistical significance of the *decrement* in overall fit as free parameters are eliminated in model trimming or the *improvement* in fit as free parameters are added in model building. As its name suggests,  $\chi^2_D$  is simply the difference between the  $\chi^2_M$  values of two hierarchical models estimated with the same data. Its degrees of freedom,  $df_D$ , equal the difference between the two respective values of  $df_M$ . The  $\chi^2_D$  statistic tests the **equal-fit hypothesis** for two hierarchical models; specifically, smaller values of  $\chi^2_D$  lead to the failure to reject the equal-fit hypothesis, but larger values result in its rejection.

In model trimming, rejection of the equal-fit hypothesis suggests that the model has been simplified too much. But the same result in model building supports retention of the path that was just added. Ideally, the more complex of the two models compared with  $\chi^2_D$  should fit the data reasonably well; if not, it makes little sense to compare the relative fit of two nested models, neither of which adequately explains the data.

Suppose for an overidentified model that

$$\chi^2_{M1}(5) = 18.30$$

A direct effect is added to the model ( $df_M$  is reduced by 1), and the result is

$$\chi^2_{M2}(4) = 9.10$$

Given both results,

$$df_D = 5 - 4 = 1$$

$$\chi^2_D(1) = 18.30 - 9.10 = 9.20, p = .002$$

which says that the overall fit of the new model with an additional path (M2) is statistically better than that of the original model (M1) at the .05 level. In this example, the chi-square difference test is univariate because it concerned a single path ( $df_D = 1$ ). When two hierarchical models that differ by two or more paths are compared ( $df_D \geq 2$ ), the chi-square difference test is a multivariate test of all added (or deleted) paths together. If  $p < .05$  for  $\chi^2_D$  in this case, at least one of the paths may be statistically significant at the .05 level if tested individually, but this is not guaranteed.

**TOPIC BOX 12.2****Scaled Chi-Square Difference Tests**

The Satorra and Bentler (2001) method calculates by hand a scaled chi-square difference statistic when comparing two hierarchical models in MLR estimation. It is assumed next that model 1 is simpler than model 2 (i.e.,  $df_{M1} > df_{M2}$ ), the unscaled test statistics are the ML chi-squares,  $\chi^2_M$  (Equation 12.1), and the scaled test statistics are the Satorra–Bentler scaled chi-squares,  $\chi^2_{SB}$  (Equation 12.2):

1. Calculate the unscaled chi-square difference statistic and its degrees of freedom in the usual way, that is:

$$\chi^2_D = \chi^2_{M1} - \chi^2_{M2} \quad \text{and} \quad df_D = df_{M1} - df_{M2}$$

2. Recover the scaling correction factor,  $c$ , for each model, as follows:

$$c_1 = \frac{\chi^2_{M1}}{\chi^2_{SB1}} \quad \text{and} \quad c_2 = \frac{\chi^2_{M2}}{\chi^2_{SB2}} \quad (12.9)$$

3. Calculate the scaled chi-square difference statistic  $\hat{\chi}^2_D$ , as follows:

$$\hat{\chi}^2_D = \frac{\chi^2_D}{(c_1 df_{M1} - c_2 df_{M2}) / df_D} \quad (12.10)$$

where the probability for  $\hat{\chi}^2_D(df_D)$  in a central chi-square distribution is the  $p$  value for the scaled chi-square difference test.

The “difftest” option in Mplus automatically computes values of scaled chi-square difference statistics. The freely available program SBDIFF.EXE by Crawford (2007) for Windows platform computers is another option. There is also a calculating webpage for the scaled chi-square difference test.\* It can happen in small samples or when the simpler model is very wrong that the denominator in Equation 12.10 is  $< 0$ , which invalidates the test. Satorra and Bentler (2010) described a new scaled chi-square difference test that avoids negative values, but it requires information that is not in standard output from SEM computer programs. Bryant and Satorra (2012) provide syntax for EQS, Mplus, and LISREL 8 for implementing the new difference test. A Microsoft Excel spreadsheet by Bryant and Satorra (2013) that automatically calculates the new test for EQS, Mplus, and LISREL 8-9 can be freely downloaded.

\*[www.uoguelph.ca/~scolwell/difftest.html](http://www.uoguelph.ca/~scolwell/difftest.html)

Note that the difference between the Satorra–Bentler scaled chi-squares of two hierarchical models cannot generally be interpreted as a statistic that tests the equal-fit hypothesis. This is because such differences do not follow chi-square distributions. Described in Topic Box 12.2 are methods to calculate a **scaled chi-square difference statistic**, which follows approximate chi-square distributions. Exercise 5 asks you to conduct the scaled chi-square difference test for a pair of hierarchical models. The researcher in robust estimation can always compare the relative fit for two hierarchical models, each fitted to the same data based on each model's set of approximate fit indexes (RMSEA, CFI, SRMR, etc.), but these comparisons are not significance tests. If the simpler model has obviously worse correspondence with the data based on values of approximate fit indexes, the more complex model would be preferred. This assumes that the fit of the more complex model is acceptable based on local fit testing.

### Empirical versus Theoretical Respecification

The interpretation of  $\chi_D^2$  as a test statistic depends in part on whether the new model is derived empirically or theoretically. For example, if individual paths that are not statistically significant are dropped from the model, it is likely that  $\chi_D^2$  will not be statistically significant. But if the deleted path is also predicted in advance to be zero, then  $\chi_D^2$  is of utmost interest. If respecification is driven entirely by empirical criteria such as statistical significance, the researcher should worry—a lot, actually—about capitalization on chance. This is because a path coefficient may be significant due only to chance variation, and its inclusion in the model would be akin to a Type I error. Similarly, a path coefficient that corresponds to a true nonzero causal effect may not be significant, and its exclusion from the model would be essentially a Type II error. A buffer against the problem of sample-specific results, though, has a greater role for theory in respecification.

The issue of capitalization on chance is especially relevant when the researcher uses an “automatic modification” option available in some computer tools such as LISREL. These wholly exploratory procedures drop or add paths according to empirical criteria such as statistical significance at the .05 level of a **modification index**, which is calculated for every path that is fixed to zero. A modification index is actually a univariate **Lagrange Multiplier** (LM) (after the Italian mathematician and astronomer J.-L. Lagrange), which is expressed as a chi-square statistic with a single degree of freedom, or  $\chi^2(1)$ . It *approximates* the amount by which  $\chi_M^2$  would decrease if a particular fixed-to-zero parameter were freely estimated; that is, a modification index *estimates*  $\chi_D^2(1)$  for adding a single path. Thus, the greater the value of a modification index, the better the predicted improvement in overall fit if that path were added to the model. Similarly, a multivariate LM estimates the effect of allowing a set of constrained-to-zero parameters to be freely estimated. Some computer programs, such as Amos and EQS, allow the user to generate modification indexes for specific parameters, which lends a more a priori sense to this statistic.

Note three cautions about modification indexes:

1. The computer may print the value of a modification index for an “illegal” parameter, such as a covariance between a measured exogenous variable and a disturbance. If you actually tried to add that parameter, the analysis would fail.
2. Modification indexes may be printed for parameters that, if actually added to the model, would make the respecified model nonidentified.
3. Each individual modification index assumes that the model is correctly specified, except for the fixed-to-zero parameter associated with that index. These assumptions are actually contradictory over a whole set of modification indexes for the same model.

The first and second cautions just listed are explained by the fact that modification indexes merely estimate  $\chi^2_D(1)$  values. These estimates are *not* derived by the computer actually adding the parameter to the model and rerunning the analysis. Instead, the computer uses a shortcut method based on linear algebra that “guesses” at the value of  $\chi^2_D(1)$ , given the covariance matrix and estimates for the more restricted (original) model.

The **Wald W statistic** (after mathematician Abraham Wald) is used in model trimming. A univariate Wald W estimates the amount by which the overall  $\chi^2_M$  would *increase* if a particular freely estimated parameter were fixed to zero (trimmed); that is, a univariate Wald W estimates  $\chi^2_D(1)$  for dropping the path. A value of a univariate Wald W that is not statistically significant at, say, the .05 level predicts a decrement in overall model fit that is not significant at the same level. Model trimming that is empirically based would thus delete paths with Wald W statistics that are not significant, but actually doing so just capitalizes on chance. A multivariate Wald W approximates the value of  $\chi^2_D$  for trimming two or more paths from the model.

All the test statistics just described are sensitive to sample size. Accordingly, even a trivial change in overall model fit due to adding or dropping a free parameter could be statistically significant in a very large sample. In addition to noting the statistical significance of a modification index, the researcher should also consider the absolute magnitude of the change in the coefficient for the parameter if it is allowed to be freely estimated, or the **expected parameter change**. If the expected change (i.e., from zero) is small, the statistical significance of the modification index may reflect more the sample size than it does the magnitude of the corresponding effect (Kaplan, 2009, pp. 124–126).

## Specification Searches

Results of two early computer simulation studies of **specification searches** by MacCallum (1986) and Silvia and MacCallum (1988) are eye opening. They took known structural equation models, imposed specification errors on them, and evaluated the erroneous models using data generated from populations in which the known mod-

els were true. In MacCallum's (1986) study, models were respecified using empirically based methods, such as modification indexes. Most of the time the changes suggested by empirically based methods were *wrong*, which means that they typically did not recover the true model. This pattern was even more apparent for small samples (e.g.,  $N = 100$ ). It is not hard to figure out why: Purely empirical respecification chases sampling error and accordingly changes the model, but covariance patterns in one sample do not precisely mimic those in the population. Silvia and MacCallum (1988) followed a similar procedure except that the application of automatic modification was guided by theory, which improved the chances of recovering the true model. The lesson of these studies is clear: Learn from your data, but your data should not be your teacher.

Marcoulides and Ing (2012) describe automated, yet "intelligent," specification search methods based on heuristics that try to optimize respecification compared with "dumb" specification searches (e.g., automatic modification based on LM statistics). These algorithms are based on principles of data mining or machine learning. For example, genetic search methods evaluate models through successive "generations" from parent to child model. Ant colony optimization methods aim to maximize fit by converging on the correct model in ways that mimic how ants forage for food by accumulating pheromones on the shortest route, stimulating other ants to take the same path. These methods tend to work best in very large data sets that allow the possibility for replication. They also perform better when guided by good hypotheses about presumed causes. These kinds of algorithms are not yet implemented in SEM computer programs, but I am skeptical of any specification search method, "intelligent" or otherwise, that is not guided by reason.

## **Example of Model Building**

Recall that the recursive path model of illness for the detailed example (Figure 7.5) does not have acceptable fit (e.g., Tables 11.8, 12.1). Reported in Table 12.2 from the LISREL analysis are values of modification indexes for all paths that could be added to the original model, yet remain recursive. Also reported in the table for each path is  $\chi^2_D(1)$ , or the actual reduction in the model chi-square after each path is individually added to the original model. Although modification indexes only estimate the corresponding  $\chi^2_D(1)$  values, that approximation for this example is close.

Note in Table 12.2 that the modification indexes for three omitted paths,

$$\text{Stress} \rightarrow \text{Fitness} \quad \text{Fitness} \rightarrow \text{Stress} \quad \text{and} \quad D_{\text{Fi}} \curvearrowright D_{\text{St}}$$

are each significant at the .05 level. The values of these indexes for two of these paths, from stress to fitness and the reverse, are similar, respectively, 5.357 and 5.096 (see the table). The addition of either path to the model would result in about the same expected decrease in  $\chi^2_M$  for the respecified models. A different respecification that would reduce  $\chi^2_M$  by a smaller amount (3.896) is to allow the disturbances of fitness and stress to covary.

**TABLE 12.2. Modification Indexes and Actual Chi-Square Difference Statistics for a Recursive Path Model of Illness**

Path	MI	p	$\chi^2_D(1)$
Stress → Fitness	5.357	.021	5.424
Fitness → Stress	5.096	.024	5.170
$D_{Fi} \curvearrowleft D_{St}$	3.896	.048	3.972
Hardiness → Fitness	2.931	.087	2.950
Hardiness → Illness	2.459	.117	2.477
Exercise → Stress	1.273	.259	1.278
Exercise → Illness	.576	.448	.578

Note. MI, modification index. All results were computed by LISREL.

Now, which respecified model just mentioned is correct—if any? It does make sense that fitness could affect the experience of stress: People who are in better physical shape may better withstand stress (Fitness → Stress). But is it not also plausible that stress could affect measured fitness (Stress → Fitness), or that fitness and stress may have common unmeasured causes (correlated disturbances)? Without theory as a guide, there is no way to select among the three alternative respecifications (among others). None of the remaining modification indexes in Table 12.2 is statistically significant. They were all calculated for respecified models with no additional paths between fitness and stress, but the omission of a path between these variables could be a specification error. This is a limitation of any modification index: Specification errors elsewhere in the model could affect its accuracy. (The same is true of covariance, correlation, standardized, and normalized residuals.)

## COMPARING NONHIERARCHICAL MODELS

Sometimes researchers compare alternative models based on the same variables and fitted to the same data matrix that are not hierarchically related.<sup>1</sup> The values of  $\chi^2_M$  from two nonhierarchical models can be informally compared, but the difference between them cannot be interpreted as a test statistic; namely, the chi-square difference test (unscaled or scaled) does not apply. This is where the family of predictive fit indexes comes in handy.

Perhaps the best known statistic in this category under ML estimation is the Akaike Information Criterion (AIC) (Akaike, 1974). It is based on an information the-

<sup>1</sup>It is also theoretically possible to compare nonhierarchical models based on different subsets of variables measured in the same sample, but such comparisons become less and less meaningful as the common set of variables gets smaller to the point where no variables are shared.

ory approach to data analysis that combines statistical estimation and model selection in a single framework. It also controls for  $df_M$  in that it may favor simpler models. Confusingly, two different formulas for the AIC are presented in the SEM literature. The first is

$$AIC_1 = \chi_M^2 + 2q \quad (12.11)$$

where  $q$  is the number of free model parameters. Equation 12.11 thus *increases* the model chi-square by a factor of twice the number of freely estimated parameters. The second formula is

$$AIC_2 = \chi_M^2 - 2df_M \quad (12.12)$$

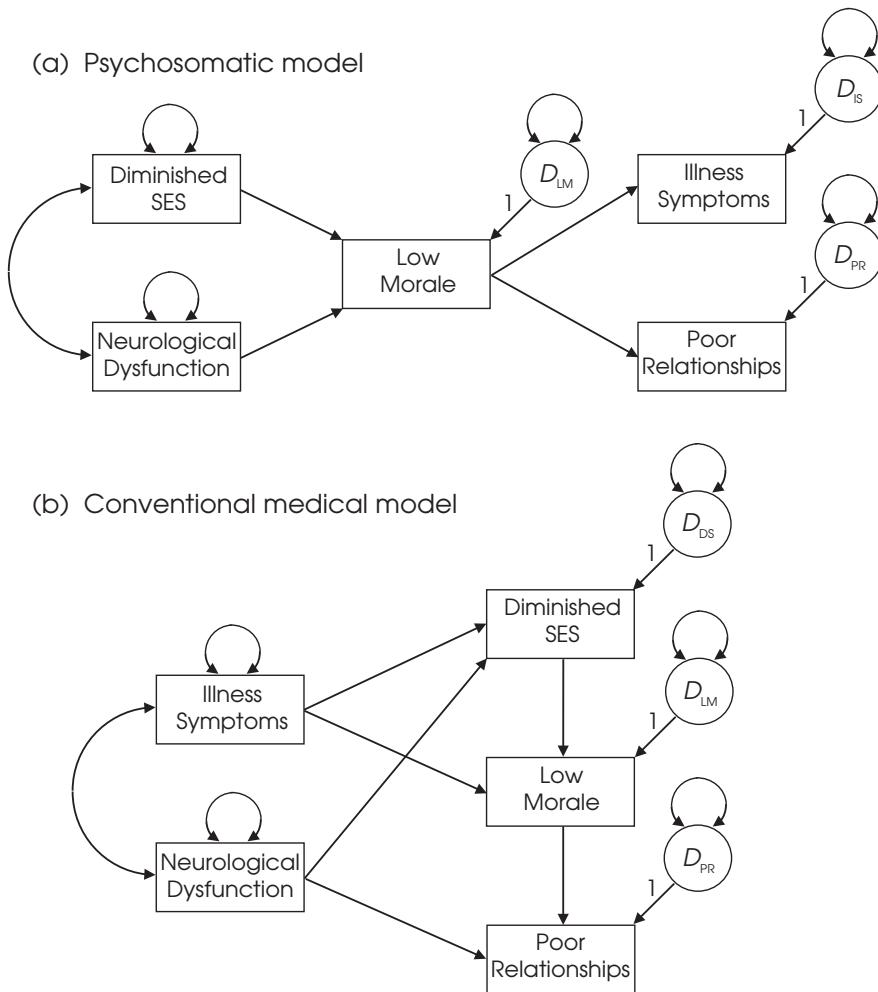
which *decreases* the model chi-square by a factor of twice the model degrees of freedom. Although the two formulas are different, the key is that the relative change in the AIC is the same in both versions, and this change is a function of model complexity. Note that the relative correction for complexity of the AIC becomes smaller and smaller as the sample size increases.

A predictive fit index that takes more direct account of sample size is the Bayes Information Criterion (BIC) (Raftery, 1995). The formula is

$$BIC = \chi_M^2 + q \ln(N) \quad (12.13)$$

where  $q$  is the number of free parameters and  $\ln(N)$  is the natural logarithm (base  $e$ , approximately 2.7183) of the sample size. The AIC and BIC are generally used to select among competing nonhierarchical models; specifically, the model with the *smallest* value of the particular predictive fit index is chosen as the one most likely to replicate. This model has relatively better fit and fewer free parameters than competing models. More complex models with comparable overall fit may be less likely to replicate owing to greater capitalization on chance. An example follows.

Presented in Figure 12.1 are two different recursive path models of recovery after cardiac surgery evaluated by Romney, Jenkins, and Bynner (1992). The *psychosomatic model* of Figure 12.1(a) represents the hypothesis that patient morale transmits the effects of neurological dysfunction and diminished socioeconomic status (SES) on physical symptoms and social relationships. The *conventional medical model* of Figure 12.1(b) represents a different pattern of causal relations among the same variables. Reported in Table 12.3 are the correlations among the observed variables for a sample of 469 patients. Unfortunately, Romney et al. (1992) did not report standard deviations, and the analysis of a correlation matrix with default ML estimation is not recommended. I used the SEPATH module of STATISTICA Advanced (StatSoft, 2013) to fit each model to the correlation matrix in Table 12.3 using the method of constrained ML estimation. Both analyses converged to admissible solutions.



**FIGURE 12.1.** Alternative nonhierarchical recursive path models of adjustment after cardiac surgery.

**TABLE 12.3. Input Data (Correlations) for Analysis of Nonhierarchical Recursive Path Models of Recovery after Cardiac Surgery**

Variable	1	2	3	4	5
1. Low Morale	1.00				
2. Illness Symptoms	.53	1.00			
3. Neurological Dysfunction	.15	.18	1.00		
4. Poor Relationships	.52	.29	-.05	1.00	
5. Diminished SES	.30	.34	.23	.09	1.00

Note. These data are from Romney et al. (1992);  $N = 469$ .

Values of selected fit statistics for the two alternative Romney et al. (1992) path models are reported in Table 12.4. It is no surprise that the global fit of the more complex *conventional medical model* ( $df_M = 3$ ) is better than that of the simpler *psychosomatic model* ( $df_M = 5$ ). But the fit advantage of the more complex model is enough to offset the penalty for having more free parameters imposed by  $AIC_1$  and also the penalty weighted by a factor of the sample size levied by the BIC. For example,  $AIC_1 = 27.238$  for the *conventional medical model*, but for the *psychosomatic model*,  $AIC_1 = 60.402$  (see the table). Exercise 6 asks you to verify these results, which altogether say that the *conventional medical model* is preferred.

Results of computer simulations by Preacher and Merkle (2012) indicate that model selection decisions based on the BIC are subject to sampling error—perhaps so much so that claims of model superiority may not hold up. (This caution also applies to the AIC and related information criteria.) One reason is that unlike most statistics, variation in the BIC actually *increases* with sample size. This is because values of the  $\chi^2_M$  component of the BIC increase with sample size for false models that are not just-identified. Also, **model selection uncertainty**, or sampling variations in the rank order of models based on BIC values, did *not* generally decrease with increasing sample size. Within all sample sizes studied by Preacher and Merkle (2012) ( $N = 80 - 5,000$ ), there was considerable variation in model rankings. Their results suggest that caution should be exercised about declaring a particular model selected by the BIC or other predictive fit indexes as the clear “winner” over rival models.

**TABLE 12.4. Values of Selected Fit Statistics for Two Nonhierarchical Recursive Path Models of Adjustment after Cardiac Surgery**

Statistic	Model	
	Psychosomatic model (Figure 12.1(a))	Conventional medical model (Figure 12.1(b))
$\chi^2_M$	40.402	3.238
$df_M$	5	3
$p$	< .001	.356
$q$	10	12
$AIC_1$	60.402	27.238
BIC	101.908	77.045
RMSEA [90% CI]	.120 [.086, .156]	.016 [0, .080]
CFI	.913	.999
SRMR	.065	.016

Note.  $q$ , number of free parameters; CI, confidence interval. All results except  $AIC_1$  and BIC were computed by SEPATH in STATISTICA.

## POWER ANALYSIS

Researchers can estimate statistical power at one of two different levels in SEM. The first concerns the power to detect an individual effect (parameter). A method for estimating the power of single-*df* tests is described by Saris and Satorra (1993). A drawback of this method is that it must be repeated for every individual parameter for which an estimate of power is desired. A more contemporary option is to use a Monte Carlo method such as the ones implemented in Mplus, EQS, or PRELIS of LISREL, which estimates the proportion of generated samples where the null hypothesis that some parameter of interest is zero is correctly rejected (Bandalos & Leite, 2013). An alternative in a simulation is to estimate the minimum sample sizes needed in order to attain power levels that equal or exceed a target value.

An approach to power analysis at the model level by MacCallum, Browne, and Sugawara (1996) and Hancock and Freeman (2001) is based on the RMSEA and noncentral chi-square distributions for tests of the exact-fit hypothesis ( $\epsilon_0 = 0$ ), the close-fit hypothesis ( $\epsilon_0 \leq .05$ ), and the not-close-fit hypothesis ( $\epsilon_0 \geq .05$ ) (Topic Box 12.1). The analysis for any of the null hypotheses just listed is conducted by specifying  $N$ ,  $\alpha$ ,  $df_M$ , and a suitable value of  $\epsilon$  under the alternative hypothesis, or  $\epsilon_1$ . For example,  $\epsilon_1$  could be specified for the close-fit hypothesis as .08, which exceeds the threshold of .05 for “close fit” but not the threshold of .10 for “poor fit.” For the not-close-fit hypothesis,  $\epsilon_1$  could be specified as .01, a value that represents even better approximate fit than  $\epsilon_1 = .05$ . A variation is to determine the minimum sample size needed to reach a target level of power, given  $\alpha$ ,  $df_M$ ,  $\epsilon_0$ , and  $\epsilon_1$ . (Recall the earlier discussion about whether thresholds for the RMSEA are meaningful.)

Estimated power or minimum samples sizes can be obtained by consulting special tables in MacCallum et al. (1996) or Hancock and Freeman (2001) for the not-close-fit hypothesis or through the use of a computer tool. In an appendix, MacCallum et al. (1996) gives SAS/STAT syntax for power analysis based on the methods just outlined. Friendly (2009) describes the *csmpower* macro for SAS/STAT that carries out a MacCallum–Browne–Sugawara power analysis that can be freely downloaded.<sup>2</sup> A webpage by Preacher and Coffman (2006) generates R code that conducts the same type of model-level power analysis, calculates the minimum  $N$  needed to obtain a target level of power, and estimates power for testing differences between two nested models.<sup>3</sup> Another webpage by Gnambs (2013) generates syntax for R and SPSS for power analysis based on the RMSEA and other approximate fit indexes.<sup>4</sup> The *semTools* package for R (Pornprasertmanit, Miller, Schoemann, Rosseel, et al., 2014) can also estimate power based on the RMSEA for both the close fit and not-close-fit hypotheses.

Another option is the Power Analysis module by J. Steiger in STATISTICA Advanced, which can estimate power for structural equation models over ranges of  $\epsilon_1$  (with  $\epsilon_0$

---

<sup>2</sup>[www.datavis.ca/sasmac/csmpower.html](http://www.datavis.ca/sasmac/csmpower.html)

<sup>3</sup>[www.quantpsy.org/rmsea/rmsea.htm](http://www.quantpsy.org/rmsea/rmsea.htm)

<sup>4</sup><http://timo.gnambs.at/en/scripts/powerforsem>

fixed to its specified value),  $\alpha$ ,  $df_M$ , and  $N$ . The Power Analysis module also allows the researcher to specify the values of both  $\epsilon_0$  and  $\epsilon_1$ . This feature is handy if there are theoretical reasons not to use the values of these parameters suggested by MacCallum et al. (1996). The ability to generate and inspect power curves as functions of sample size and other assumptions is useful for planning a study, especially when granting agencies demand a priori power estimates.

I used the Power Analysis module in STATISTICA (StatSoft, 2013) to estimate power for tests of both the close-fit hypothesis and the not-close-fit hypothesis for the recursive path model of illness (Figure 7.5) for the original sample size,  $N = 373$ . Also estimated for this model are the minimum sample sizes needed in order to attain a target level of power  $\geq .80$  for each of the two preceding hypotheses. The results are summarized in Table 12.5. The estimated power for the test of the close-fit hypothesis for the original sample size is .317. If this model actually does *not* have close fit in the population, then the estimated probability that we can reject this incorrect model is just greater than 30%, given the other assumptions for this analysis (see the table). For the same model, the estimated power for the test of the not-close-fit hypothesis for the original sample size is .229, so there is only about a 23% chance of detecting a model with “good” approximate fit in the population. The minimum sample sizes needed in order for power to be at least .80 for tests of the close-fit hypothesis and the not-close-fit hypothesis are, respectively, 1,465 and 1,220 cases. Presented in Figure 12.2 is the curve for the analysis just described that shows the relation between sample size and power for the test of the close-fit hypothesis for this example.

These results reflect a general trend that power at the model level may be low when there are few degrees of freedom (here,  $df_M = 5$ ), even for a sample size that is large enough ( $N = 373$ ) for reasonable statistical precision of the parameter estimates. For models with only one or two degrees of freedom, sample sizes in the thousands may be required in order for power to be  $\geq .80$  (MacCallum et al., 1996, p. 144). Sample

**TABLE 12.5. Power Analysis Results for a Recursive Path Model of Illness**

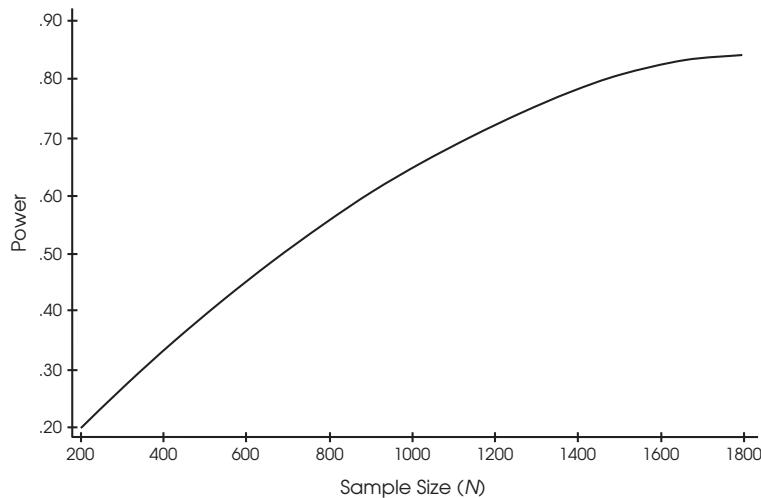
<u>Power at <math>N = 373</math></u>	
Close fit <sup>a</sup>	.317
Not close fit <sup>b</sup>	.229
<u>Minimum <math>N</math> for power <math>\geq .80^c</math></u>	
Close fit	1,465
Not close fit	1,220

Note.  $df_M = 5$ ,  $\alpha = .05$ . All results were computed by Power Analysis in STATISTICA.

<sup>a</sup> $H_0: \epsilon_0 \leq .05, H_1: \epsilon_1 = .08$ .

<sup>b</sup> $H_0: \epsilon_0 \geq .05, H_1: \epsilon_1 = .01$ .

<sup>c</sup>Sample size rounded up to closest multiple of 5.



**FIGURE 12.2.** Power as a function of sample size for the test of the close-fit hypothesis  $\epsilon_0 \leq .05$  against the alternative hypothesis  $\epsilon_1 = .08$  at  $\alpha = .05$  and  $df_M = 5$  for a recursive path model of illness.

size requirements for the same level of power drop to some 300 to 400 cases when  $df_M$  is about 10. Even smaller sample sizes may be needed for a target power of  $\geq .80$  if  $df_M > 20$ , but the sample size should not be less than 100 or so. If an analysis in global fit testing has a low probability of rejecting a false model, this fact should temper the researcher's enthusiasm for his or her preferred model.

Following is a brief summary of other developments in power estimation at the model level. Kim (2005) studied a total of four approximate fit indexes, including the RMSEA and CFI, in relation to power estimation and the determination of sample size requirements for minimum desired levels of power. Kim (2005) found that estimates of power and minimum sample sizes varied as a function of choice of the index, number of observed variables and model degrees of freedom, and magnitude of covariation among the variables. This result is not surprising considering that (1) different fit statistics reflect different aspects of model fit and (2) little direct correspondence exists between values of various fit statistics and degrees of freedom or types of model misspecification. As Kim (2005) notes, a value of .95 for the CFI does not necessarily indicate the same misspecification as a value of .05 for the RMSEA. See Hancock and French (2013) for more information on power analysis in SEM.

## EQUIVALENT AND NEAR-EQUIVALENT MODELS

After a final model is selected from hierarchical or nonhierarchical alternatives, equivalent versions should be considered. Equivalent models have the same degrees of freedom (they are equally complex) but feature a different configuration of paths among the

same variables. The most general form is **observational equivalence**, which says that one model generates every probability distribution that can be generated by another model (Hershberger & Marcoulides, 2013). A particular form for linear models fitted to covariance matrices is **covariance equivalence**, which means that every covariance matrix predicted by one model can also be generated by another model. Two covariance equivalent models also generate the same residuals and conditional independences, or sets of vanishing correlations (Pearl, 2009b). The latter property refers to **d-separation equivalence**.

The **Lee–Hershberger replacing rules** (Lee & Hershberger, 1990), summarized in Table 12.6, are the best known rules in the SEM literature for generating equivalent structural models. Note that substituting equality-constrained reciprocal direct effects for other types of paths would make the model nonrecursive, but it is assumed that the new model is identified. Some applications of the replacing rules may be implausible owing to the nature of the variables or the timing of their measurement. For example, a model with a direct effect from an acculturation variable to chronological age would be illogical, and the measurement of  $Y_1$  before  $Y_2$  in a longitudinal design is inconsistent with the specification that  $Y_2$  causes  $Y_1$ .

There are some problems with the replacing rules: They are not guaranteed to be transitive (Hershberger, 2006). Reapplying the replacing rules to equivalent versions generated by previous applications of the rules is not guaranteed to generate even more versions that are all equivalent. A greater concern is that the replacing rules can generate a new structural model that predicts a different set of conditional independences than the original model; that is, applying the replacing rules can sometimes create or destroy an implied d-separation in the respecified model compared with the original model (Pearl, 2012). I am not aware of a general law that predicts the occurrence of this problem, but I suspect that applications of the replacing rules that change the status of variables as colliders are more susceptible to this anomaly.

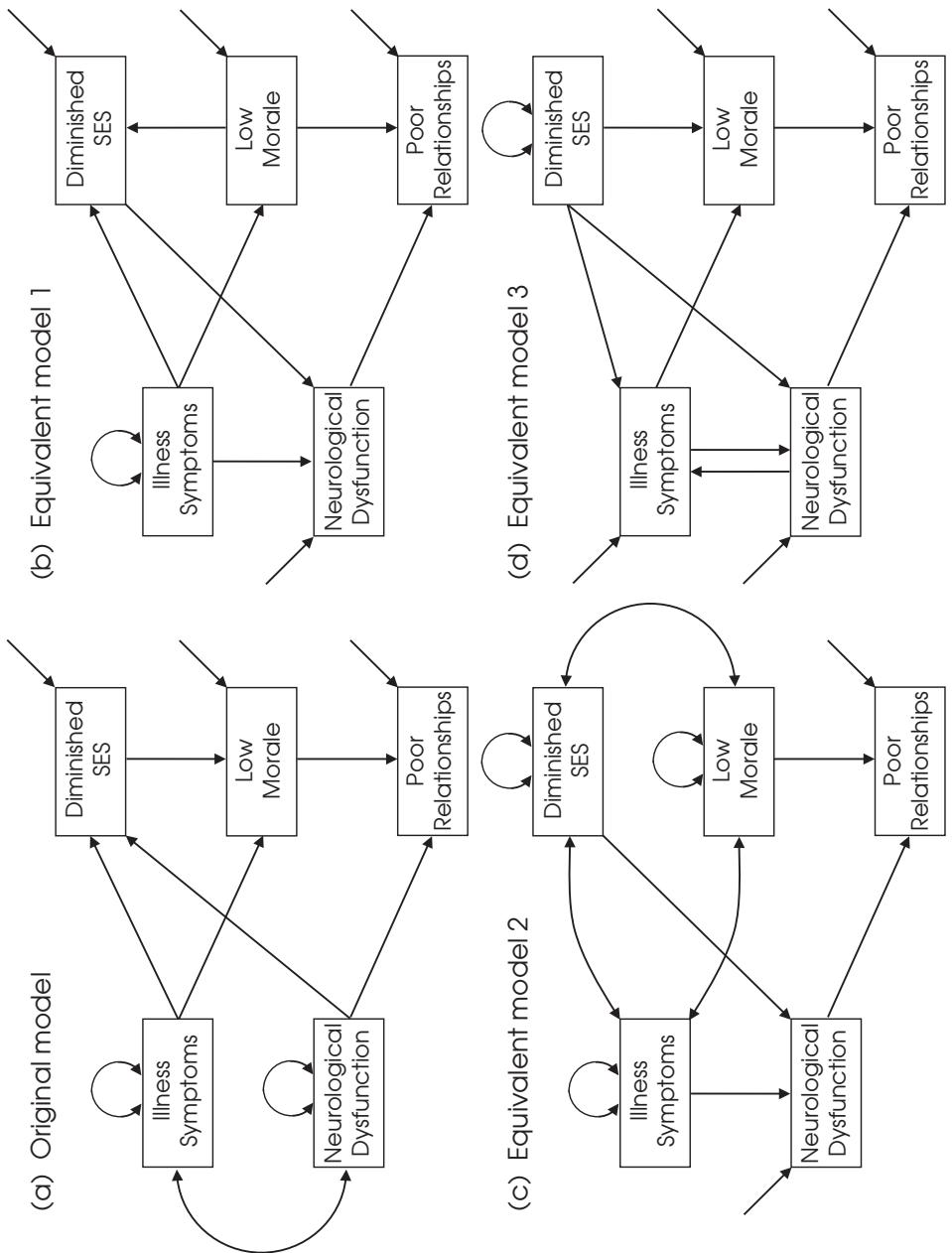
Let's consider an uncomplicated example where the replacing rules do not alter conditional independences. Presented in Figure 12.3(a) is Romney and associates' original *conventional medical model* shown with compact symbolism for disturbances. The other

**TABLE 12.6. Lee–Hershberger Replacing Rules for Structural Models**

Within a block of variables at the beginning of a structural model that is just-identified and with unidirectional relations to subsequent variables, direct effects, correlated disturbances, and equality-constrained reciprocal effects<sup>a</sup> are interchangeable. For example,  $Y_1 \rightarrow Y_2$  may be replaced by  $Y_2 \rightarrow Y_1$ ,  $D_1 \rightsquigarrow D_2$ , or  $Y_1 \rightleftarrows Y_2$ . If two variables are specified as exogenous, then an unanalyzed association can be substituted, too. (Rule 12.1)

At subsequent places in the model where two endogenous variables have the same causes and their relations are unidirectional, all of the following may be substituted for one another:  $Y_1 \rightarrow Y_2$ ,  $Y_2 \rightarrow Y_1$ ,  $D_1 \rightsquigarrow D_2$ , and the equality-constrained reciprocal effect  $Y_1 \rightleftarrows Y_2$ . (Rule 12.2)

<sup>a</sup>The two unstandardized direct effects are constrained to be equal.



**FIGURE 12.3.** Four equivalent path models of adjustment after cardiac surgery shown with compact symbolism for disturbances.

three models in the figure are generated from the original model using the replacing rules in Table 12.6. For example, the equivalent model of Figure 12.3(b) substitutes a direct effect for a covariance between illness symptoms and neurological dysfunction. It also reverses the direct effects between diminished SES and low morale and between diminished SES and neurological dysfunction. The equivalent model of Figure 12.3(c) replaces two of three direct effects that involve diminished SES with covariances. The equivalent model of Figure 12.3(d) replaces the covariance between illness symptoms and neurological dysfunction with an equality-constrained reciprocal effect. It also reverses the direct effect between illness symptoms and diminished SES and between neurological dysfunction and diminished SES. All four models in the figure have the same fit to the data (i.e.,  $\chi^2_M(3) = 3.238$  for each model). They also imply the same conditional independences. (You should verify this statement.) This is the best case when applying the replacing rules.

Now we consider a more difficult example. Presented in Figure 12.4(a) is an original path model with a direct effect from  $Y_1$  to  $Y_2$  and a disturbance correlation between  $Y_2$  and  $Y_3$ . There are three paths between  $X$  and  $Y_3$ :

$$\begin{aligned} X &\rightarrow Y_2 \leftarrow U \rightarrow Y_3 \\ X &\rightarrow Y_1 \rightarrow Y_2 \leftarrow U \rightarrow Y_3 \\ X &\rightarrow Y_1 \rightarrow Y_3 \end{aligned}$$

where  $U$  represents an unmeasured cause of  $Y_2$  and  $Y_3$  and thus replaces their disturbance correlation. The first two paths just listed are blocked by the collider  $Y_2$ , and controlling for  $Y_1$  blocks the open third path. Thus, Figure 12.4(a) implies

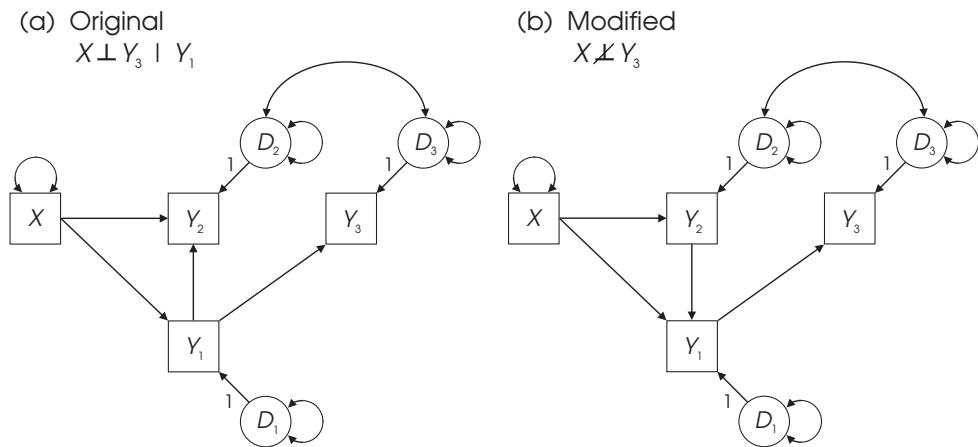
$$X \perp Y_3 \mid Y_1$$

which for continuous variables in a linear model with no interactions predicts that  $\rho_{XY_3 \cdot Y_1} = 0$ .

Because  $Y_1$  and  $Y_2$  in Figure 12.4(a) have the same cause,  $X$ , we can apply Rule 12.2 (see Table 12.6) and reverse the path between them. The model so respecified is presented in Figure 12.4(b), and the new paths between  $X$  and  $Y_3$  are listed next:

$$\begin{aligned} X &\rightarrow Y_2 \leftarrow U \rightarrow Y_3 \\ X &\rightarrow Y_1 \leftarrow Y_2 \leftarrow U \rightarrow Y_3 \\ X &\rightarrow Y_1 \rightarrow Y_3 \end{aligned}$$

The first path just listed is blocked by the collider  $Y_2$ . The second path is also blocked by a collider, but that collider in the respecified model is  $Y_1$ , not  $Y_2$ . Controlling for  $Y_1$  will block the open third path, but doing so will open the second path. Controlling for both  $Y_1$  and  $Y_2$  will close the second path, but doing so will open the first path. Therefore, variables  $X$  and  $Y_3$  in Figure 12.4(b) cannot be d-separated. Figures 12.4(a) and 12.4(b)



**FIGURE 12.4.** An original path model (a) and a modified version generated by applying the replacing rules that is not d-separation equivalent (b).

are not d-separation equivalent, even though one model generates the other using the replacing rules. Pearl (2009b, pp. 145–149) describes graphical criteria for generating d-separation equivalent models.

Relatively simple structural models may have few equivalent versions, but more complicated ones may have hundreds or even thousands (MacCallum, Wegener, Uchino, & Fabrigar, 1993). In general, more parsimonious structural models tend to have fewer equivalent versions. You will learn in the next chapter that CFA measurement models can have infinitely many equivalent versions; thus, it is unrealistic that researchers consider all possible equivalent models. As a compromise, researchers should generate at least a few substantively meaningful equivalent versions. Unfortunately, even this limited step is usually neglected. Few authors of SEM studies even acknowledge the existence of equivalent models (MacCallum & Austin, 2000). This type of confirmation bias threatens most published SEM studies.

Pearl (2009b) notes that the existence of equivalent models is inevitable if we agree that causal relations cannot be inferred from data alone, that is, without assumptions. There are few statistical bases for preferring one equivalent model over another (Hershberger & Marcoulides, 2013, pp. 30–33). There may be stronger grounds for a preference in theory or results of empirical studies when presumed causal variables in the model are manipulated. So remember that any retained structural model may be just one exemplar from a larger equivalence class of models that all explain the same data equally well. Specifying models that are simpler is one way to eliminate some equivalent versions. Other ways include using designs with time precedence in measurement or applying strong theoretical knowledge about the directionalities of causal effects—see Williams (2012) for more information.

The problem of equivalent models explains the need for time precedence in studies where the goal is to estimate mediation. Suppose that variables  $X$ ,  $M$ , and  $Y$  are, respec-

tively, a presumed cause, a mediator, and an outcome. With no time precedence in their measurement and also no other variables in the model, any rearrangement of direct effects between  $X$ ,  $M$ , and  $Y$  generates an equivalent version (six in total). For each of the equivalent models just mentioned, substitution of an equality-constrained reciprocal effect for a direct effect between variables with the same cause generates another equivalent model (six in total). There are even more possibilities for equivalent models that do not include indirect effects, such as the model with the paths listed next:

$$Y \rightarrow X \quad Y \rightarrow M \quad \text{and} \quad D_X \curvearrowleft D_M$$

where a direct effect between two variables with the same cause is replaced by a disturbance correlation (6 in total). Altogether, there are 18 equivalent versions of path models for variables  $X$ ,  $M$ , and  $Y$  if there is no time precedence in measurement that would rule out certain versions.

In addition to equivalent models, there may also be **near-equivalent models** that do not generate the exact same predicted covariances or conditional independences, but nearly so. For instance, the recursive path model of illness in Figure 7.5 but with (1) a direct effect from fitness to stress or (2) a direct effect from stress to fitness are near-equivalent models (see Table 12.2). There is no specific rule for generating near-equivalent models. Instead, such models would be specified according to theory. In some cases, near-equivalent models may be more numerous than truly equivalent models and thus a more serious research threat than the equivalent models.

## SUMMARY

Although optimal strategies for global fit testing are still being debated in the SEM literature, there is consensus that some routine practices are inadequate. One practice involves ignoring a failed exact-fit test in samples that are not large, and another is the claim for “good” fit based on values of approximate fit indexes that exceed—or, in some cases, fall below—suggested thresholds based on prior simulation studies. Instead, researchers should explicitly diagnose possible sources of misspecification by describing patterns of residuals or values of modification indexes with a basis in theory. Researchers often seek to select a structural equation model from a set of alternative models all analyzed with the same data. The most frequent context is when hierarchical models are compared where a simpler model is nested within a more complex model. The chi-square difference test in this case evaluates the equal-fit hypothesis. There is no significance test for comparing nonhierarchical models, but predictive fit indexes can be used to evaluate such models. If a model is retained, then it is important to estimate statistical power and generate at least a few plausible equivalent or near-equivalent versions. If the power to reject a false model is low or there are few grounds to choose among equivalent versions, then little support for the researcher’s preferred model is indicated. The next chapter deals with the analysis of reflective measurement models in the CFA technique.

**LEARN MORE**

Tomarken and Waller (2003) consider examples of poor explanatory power for models with apparently “good” fit based on values of global fit statistics, and the special issue of the journal *Personality and Individual Differences* on SEM concerns the roles of test statistics and approximate-fit statistics in global fit testing (Vernon & Eysenck, 2007). West, Taylor, and Wu (2012) address the general challenge of model selection in SEM.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with “well-fitting” models. *Journal of Abnormal Psychology*, 112, 578–598.

Vernon, P. A., & Eysenck, S. B. G. (Eds.). (2007). Structural equation modeling [Special issue]. *Personality and Individual Differences*, 42(5).

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. 209–246). New York: Guilford Press.

**EXERCISES**

1. Calculate the value of the RMSEA in Table 12.1.
2. Calculate the value of the CFI in Table 12.1.
3. Compare the value of the SRMR in Table 12.1 with the average absolute correlation residual in Table 11.8 for the same model and data.
4. Fit the path model in Figure 7.5 to the data in Table 4.2 but specify  $N = 5,000$ . What results in Table 12.1 do you expect to change compared with the original analysis where  $N = 373$ ?
5. Calculate the scaled chi-square difference statistic given the results listed next for two hierarchical models (see Topic Box 12.2):

Model 1:  $df_{M1} = 17$ ,  $\chi^2_{M1} = 57.50$ ,  $\chi^2_{SB1} = 28.35$   
Model 2:  $df_{M2} = 12$ ,  $\chi^2_{M2} = 18.10$ ,  $\chi^2_{SB2} = 11.55$
6. Calculate the values of the  $AIC_1$  and  $BIC$  based on the results in Table 12.4 for the non-nested path models in Figure 12.1.

## Appendix 12.A

### Model Chi-Squares Printed by LISREL

Under ML estimation assuming multivariate normality, when the observed (uncorrected) covariance matrix is analyzed, LISREL 9 (Scientific Software International, 2013) prints two model chi-squares. One is the product  $N(F_{ML})$  (i.e.,  $\chi^2_M$ ), which is labeled *maximum likelihood ratio chi-square* and  $C_1$  in program documentation. The other is labeled Browne's (1984) ADF *chi-square* and  $C_2(\mathbf{W}_{NT})$  in documentation. The latter is based on the weight matrix in ADF estimation that assumes multivariate normality. If this assumption is tenable, the values of these two statistics should be similar. I recommend reporting  $C_1$  instead of  $C_2(\mathbf{W}_{NT})$  in order to more closely match the results generated by other SEM computer tools for the same model and data.

When the covariance matrix is asymptotic—that is, it is estimated in PRELIS and then fitted to the model (e.g., robust estimation)—LISREL prints three additional chi-squares. One is  $C_2(\mathbf{W}_{NNT})$ , where the weight matrix is from the ADF estimator that does not assume normality. This matrix may be unstable unless the sample size is very large. A better alternative for smaller samples is labeled Satorra–Bentler (1988) *scaled chi-square* and  $C_3$  in documentation.<sup>1</sup> It is defined by Equation 12.2. The statistic labeled Satorra–Bentler (1988) *adjusted chi-square* and  $C_4$  in documentation has its own degrees of freedom that is typically a fractional number. In LISREL 8 (Scientific Software International, 2006), the statistics  $C_2(\mathbf{W}_{NT})$  and  $C_2(\mathbf{W}_{NNT})$  were called C2 and C4, respectively—see Jöreskog (2004) for more information.

---

<sup>1</sup>The expression “Satorra–Bentler (1988)” refers to a conference presentation by Satorra and Bentler (1988). The corresponding journal article is Satorra and Bentler (1994).

## 13

# Analysis of Confirmatory Factor Analysis Models

---

This is the first of two chapters about the analysis of core latent-variable models in SEM, in this case reflective measurement models as evaluated in CFA. Topics include hypothesis-testing strategies, respecification, and equivalent CFA models. Special types of models are described, including hierarchical models and bifactor models. The detailed example concerns a model with continuous indicators, but methods for analyzing ordinal indicators, such as Likert-scale items, are also described and demonstrated in a second example. If you know how to analyze CFA models, it is easier to learn about the analysis of SR models, which is covered in the next chapter.

---

## FALLACIES ABOUT FACTOR OR INDICATOR LABELS

The specification of CFA models was introduced earlier using diagrams where factors are designated with letters, such as  $A$  and  $B$  (e.g., Figure 9.3). In LISREL notation, factors are designated with Greek letters, such as  $\xi_1$  and  $\xi_2$  (Appendix 9.A). In actual analyses, researchers usually assign meaningful names to factors, such as “sequential processing” (Figure 9.7). It is important to avoid three logical errors concerning factor names.

The first of these errors is the **naming fallacy**: Just because a factor is named does not mean that the hypothetical construct is understood or even correctly labeled. Factors require designation, though, if for no other reason than communication of the results. Although verbal labels are more “reader friendly” than more abstract symbols, they are conveniences, not substitutes for critical thinking. The second error is that of **reification**: the belief that a factor *must* correspond to a real thing. For example, a general cognitive ability factor, often called  $g$ , may not actually correspond to any particular genetic or neurological substrate. To automatically consider  $g$  as real instead of a statistical abstraction (i.e., a factor) is a potential error of reification. Along these lines, Gardner

(1993) reminded educators not to reify *g* by assuming that “intelligence” corresponds to a single dimension (i.e., educators should also emphasize artistic, social, athletic, and other domains).

Jingle–jangle fallacies concern names for indicators and are the third kind of error to avoid. The **jingle fallacy** is the assumption that *because* different things are called by the same name, they must be the same thing. The **jangle fallacy** refers to the belief that things must be different from each other *because* they are called by different names. In measurement, the jingle fallacy is indicated when low intercorrelations are observed among tests claimed to measure the *same* construct. In this case, no single test can be relied on as actually reflecting the target domain. The jangle fallacy is apparent when very high intercorrelations are observed among tests that are supposed to measure *different* constructs. The lesson of jingle–jangle fallacies is that interpretations of test scores should not be based on test names; instead, researchers should rely on more rigorous methods, including CFA, to establish convergent validity and discriminant validity.

## ESTIMATION OF CFA MODELS

This discussion assumes continuous indicators, which are most likely to happen when each indicator is a scale that generates a total score over a set of items. A later section of this chapter deals with the analysis of Likert-scale items as indicators (the data are ordinal).

### Interpretation of the Estimates

Parameter estimates in CFA are interpreted as follows:

1. Pattern coefficients are interpreted as regression coefficients. For example, if an unstandardized pattern coefficient is 4.0, then we expect a 4-point increase in the indicator, given an increase of 1 point in the factor. Coefficients fixed to 1.0 to scale a factor remain so in the unstandardized solution and have no standard errors, and thus have no significance (*z*) test.
2. In a standardized solution where all variables have unit variance (1.0), standardized pattern coefficients for simple indicators (they depend on a single factor) are estimated Pearson correlations. In this case, squared standardized pattern coefficients are proportions of explained variance. If a standardized coefficient is .80, for example, then the factor explains  $.80^2 = .64$ , or 64.0% of the observed variance of that simple indicator. An ideal result in CFA is that the model explains the majority of the variance ( $> .50$ ) in every continuous indicator.
3. For complex indicators that depend on two or more factors, standardized pattern coefficients are interpreted as standardized regression coefficients (beta weights) that control for correlated causes (factors). Because standardized coefficients for complex

indicators are not correlations, one cannot generally square their values to derive proportions of explained variance. Fortunately, many SEM computer tools print  $R^2$  for each indicator, whether simple or complex.

4. The ratio of an unstandardized error variance over the observed variance of the corresponding indicator equals the proportion of unexplained variance. Suppose that the variance of an indicator is 25.0 and that in CFA its error variance is 9.0. The proportion of unexplained variance is 9.0/25.0, or .36, and the proportion of explained variance is  $R^2 = 1 - .36$ , or .64.

The estimated Pearson correlation between an indicator and a factor is a **structure coefficient**. The standardized pattern coefficient for a simple indicator is a structure coefficient, but not for a complex indicator. Graham, Guthrie, and Thompson (2003) remind us that the specification that a pattern coefficient is zero does *not* mean that the correlation between that indicator and factor must be zero; that is, a zero pattern coefficient does *not* generally imply a zero structure coefficient. This is because factors in CFA models are assumed to covary, which generally implies nonzero correlations between each indicator and all factors. But indicators should have higher estimated correlations with the factors they are specified to measure.

## Types of Standardized Solutions

Some SEM computer tools print multiple standardized solutions for CFA models. In LISREL, for example, the variances of just the factors are unity in its *standardized solution* (output option “SS”), which means that estimates of pattern coefficients are still in the original metrics of the indicators. This solution in LISREL is analogous to the STD solution in Mplus, where just the factors are standardized. Either standardized solution may be preferred if raw scores on the indicators are meaningful, such as performance time in seconds. This is because the original metric is lost when a variable is standardized. Both factors and indicators are standardized in LISREL’s *completely standardized solution* (output option “SC”) and in the STDYX solution in Mplus.<sup>1</sup> Both standardized solutions are analogous to the only standardized solution generated in EQS, where all latent and observed variables are standardized. Inform your readers about the particular standardized solution you selected.

## Problems

Failure of iterative estimation in CFA can be caused by poor start values; see Appendix 13.A for suggestions. Inadmissible solutions include Heywood cases such as negative variance estimates or estimated absolute correlations  $> 1.0$ . Nonconvergence or

---

<sup>1</sup>Option STDY in Mplus standardizes all variables except for covariates, but there are no predictors in CFA models, so results for options STDYX and STDY are identical when all indicators are continuous.

improper solutions are more likely when there are only two indicators per factor or the sample size is less than 100–150 cases (Marsh & Hau, 1999). Marsh and Hau give the following suggestions for analyzing CFA models when the sample size is not large:

1. Use indicators with good psychometrics that will each also have standardized coefficients  $> .70$  or so. Such models are less susceptible to Heywood cases (Wothke, 1993).
2. Imposing equality constraints on the unstandardized coefficients of indicators of the same factor all based on the same score metric may help to prevent inadmissible solutions. This tactic makes more sense if the corresponding indicators share the same metric.
3. If the indicators are items, it may be better to analyze them in groups (parcels) rather than individually. The technique of parceling is discussed later in this chapter.

Solution inadmissibility can also occur at the parameter matrix level. The computer estimates in CFA a factor covariance matrix and an error covariance matrix (Appendix 9.A). If any element of either matrix is out of bounds, then the whole matrix is nonpositive definite. The causes of nonpositive definite parameter matrices include the following (Wothke, 1993):

1. The data provide too little information (e.g., small sample, two indicators per factor).
2. The model is overparameterized (too many free parameters).
3. The sample has outliers or severely non-normal distributions (poor data screening).
4. There is empirical underidentification concerning factor covariances (e.g., Figure 9.5(c)).
5. The measurement model is misspecified.

## **Empirical Checks for Identification**

It is theoretically possible for the computer to generate a converged, admissible solution for a model that is not really identified, yet print no warning or error message. This is most likely to happen in CFA when analyzing a nonstandard model with both correlated error terms and complex indicators for which application of heuristics cannot prove identification. Described next are empirical tests for solution uniqueness that can be applied when analyzing any type of structural equation model. These tests concern necessary but insufficient requirements; that is, if any test is failed, the solution is not unique, but passing it does not prove identification:

1. A second analysis of the same model should be done, but different start values than in the first analysis should be used. If estimation converges to a different solution working from different initial estimates, the original solution is not unique and the model is not identified.
2. This check applies to overidentified models only: Use the predicted covariance matrix from the first analysis as the input data for a second analysis of the same model. If the second analysis does not generate the same parameter estimates as the first, the model is not identified.
3. Some SEM computer programs optionally print the matrix of estimated correlations among the parameter estimates. Although parameters are fixed values in the population, their estimates vary randomly over samples. Estimates of their covariances are based on the **Fisher information matrix**, which is associated with simultaneous estimation methods. If the model is identified, this matrix has an inverse, which is the matrix of covariances among the parameter estimates. Correlations among the parameter estimates are derived from this matrix. A problem is indicated if any of these absolute correlations is close to 1.0, which indicates extreme linear dependency. Bollen and Bauldry (2010) describe additional empirical tests for identification.

## DETAILED EXAMPLE

This example concerns the analysis of the CFA model for the first edition of the Kaufman Assessment Battery for Children (KABC-I) introduced in Chapter 9 (see Figure 9.7). Briefly, three subtests are specified to measure one factor (sequential processing) and five subtests are specified to measure a second factor (simultaneous processing).<sup>2</sup> The data for this analysis, summarized in Table 13.1, are from the test's standardization sample for 10-year-old children ( $N = 200$ ). I used Mplus (Müthen & Müthen, 1998–2014) for the analyses described next. Variances are calculated in Mplus as  $S^2$  with  $N$  in the denominator, not as  $s^2$  with  $N - 1$  in the denominator. You can download from this book's website all computer files for this analysis in Mplus and also in Amos, EQS, lavaan, LISREL, SPSS, and Stata.

### Single-Factor Model

If the target model has two or more factors, the first model analyzed in CFA is often a single-factor model. If a single-factor model cannot be rejected, there is little point in evaluating models with more factors. I submitted to Mplus the covariance matrix based on the data in Table 13.1 for ML estimation of a single-factor model. The unstandardized coefficient of the Hand Movements subtest was fixed to 1.0 to scale the single fac-

---

<sup>2</sup>Keith (1985) suggested alternative factor names, including “short term memory” instead of “sequential processing” and “visual-spatial reasoning” instead of “simultaneous processing.”

**TABLE 13.1. Input Data (Correlations, Standard Deviations) for Analysis of a Two-Factor Model of the KABC-I**

Indicator	1	2	3	4	5	6	7	8
<u>Sequential scale</u>								
1. Hand Movements	1.00							
2. Number Recall	.39	1.00						
3. Word Order	.35	.67	1.00					
<u>Simultaneous scale</u>								
4. Gestalt Closure	.21	.11	.16	1.00				
5. Triangles	.32	.27	.29	.38	1.00			
6. Spatial Memory	.40	.29	.28	.30	.47	1.00		
7. Matrix Analogies	.39	.32	.30	.31	.42	.41	1.00	
8. Photo Series	.39	.29	.37	.42	.58	.51	.42	1.00
SD	3.40	2.40	2.90	2.70	2.70	4.20	2.80	3.00

Note. KABC-I, Kaufman Assessment Battery for Children, first edition. Input data are from Kaufman and Kaufman (1983), N = 200.

tor. Exercise 1 asks you to verify that  $df_M = 20$  for the single-factor model. Estimation in Mplus converged to an admissible solution. Values of selected fit statistics for the single-factor model, reported in the second column of Table 13.2, suggest poor global fit. For example, the model fails both the exact-fit and close-fit tests ( $p < .001$  for both), and the lower bound of the RMSEA's 90% confidence interval, or .119, exceeds .10, a value that may suggest poor fit. Exercise 2 asks you to generate and inspect the residuals for the single-factor model, but I can tell you that local fit is poor, too.

A single-factor CFA model is nested under any other CFA model with two or more factors for the same indicators. This is because a one-factor model is just a restricted version of any model with multiple factors where, conceptually, all absolute pairwise factor correlations are fixed to 1.0. If so, then all factors are identical, which is the same as replacing multiple factors with just one. This also means that the chi-square difference test can be conducted to directly compare the relative fit to the same data of single- versus multiple-factor CFA models.

To demonstrate the chi-square difference test for this example, I fitted the two-factor model of the KABC-I in Figure 9.7 to the data in Table 13.1 using the ML method in Mplus. The analysis converged normally to an admissible solution. The test statistic for the two-factor model is

$$\chi^2_M(19) = 38.325, p = .006$$

**TABLE 13.2. Values of Selected Fit Statistics for One- and Two-Factor Models of the KABC-I**

Statistic	Model	
	One factor	Two factors
$\chi^2_M$	105.427	38.325
$df_M$	20	19
$p$	< .001	.006
RMSEA [90% CI]	.146 [.119, .174]	.071 [.038, .104]
$p_{\epsilon_0 \leq .05}$	< .001	.132
CFI	.818	.959
SRMR	.084	.072

Note. KABC-I, Kaufman Assessment Battery for Children, first edition; CI, confidence interval. All results were computed by Mplus.

The two-factor model fails the chi-square test, so we tentatively reject it, but we will revisit this model soon. The same results for the single-factor model are

$$\chi^2_M(20) = 105.427, p < .001$$

Given both sets of results, we calculate

$$\chi^2_D(1) = 105.427 - 38.325 = 67.102, p < .001$$

which says that the fit of the model with two factors is statistically better than that of the model with a single factor. The implication of this result is not yet clear because the two-factor model is tentatively rejected. In general, the outcome of the chi-square difference test is most meaningful when the more complex of the two models being compared has acceptable fit.

Kenny (1979) noted that the test for a single factor is also relevant for path models. The inability to reject a single-factor model in this context means that the variables measure only one domain (they do not show discriminant validity). I conducted this test for the five variables of the recursive path model of illness in Figure 7.5 by fitting a single-factor CFA model to the data in Table 4.2 with the ML method in LISREL (Scientific Software International, 2013). Values of fit statistics reported next say that a single-factor model has poor global fit (local fit is poor, too), which provides a “green light” to proceed with the evaluation of a path model:

$$\begin{aligned} \chi^2_M(5) &= 60.549, p < .001 \\ \text{RMSEA} &= .173, 90\% \text{ CI } [.135, .213], p_{\epsilon_0 \leq .05} < .001 \\ \text{CFI} &= .644; \text{ SRMR} = .096 \end{aligned}$$

## Two-Factor Model

Reported in the last column of Table 13.2 are values of fit statistics for the two-factor model of the KABC-I. As mentioned, the exact-fit test is failed at  $p < .001$ . Results based on the RMSEA are mixed: The close-fit test is passed because  $\hat{\epsilon}_L = .038$  is  $\leq .05$  ( $p_{\epsilon_0 \leq .05} = .138$ ). But the not-close-fit test is failed at the .05 level because  $\hat{\epsilon}_U = .104$  is  $\geq .05$ , and the poor-fit test is also failed at the same level because  $\hat{\epsilon}_U = .104$  is  $\geq .10$  (see the table). Values of the CFI and SRMR are, respectively, .959 and .072, and neither result is clearly problematic.

The power of the tests for the close-fit hypothesis and the not-close-fit hypothesis estimated in the Power Analysis procedure of STATISTICA Advanced (StatSoft, 2013) are both low, respectively, .440 and .302.<sup>3</sup> Minimum sample sizes of over twice that of the actual size for this analysis ( $N = 200$ ) would be needed in order for power to be at least .80. For the close-fit hypothesis, this minimum sample size is about 455 cases, and the minimum sample size for the not-close-fit hypothesis is about 490 cases, both for a target level of power at .80.

Presented in Table 13.3 are parameter estimates for the two-factor model with standard errors for both solutions. Note that the unstandardized pattern coefficients for reference variables equal 1.0 and have no standard errors. All other results are statistically significant, but of greater interest are the standardized pattern coefficients, which are also structure coefficients, so their squared values are proportions of explained variance. For example, the standardized coefficient for the Hand Movements task is .497, so its factor explains  $.497^2$ , or about .247 (25%) of its variance. This is a poor result for a continuous indicator. Altogether the two-factor model fails to explain the majority ( $> .50$ ) of variance for a total of four out of eight indicators (see the table), which indicates poor convergent validity. Conversely, the estimated factor correlation (.557) is only moderate in size, which suggests reasonable discriminant validity.

Reported in Table 13.4 are structure coefficients for the two-factor model. Values presented in boldface are also standardized pattern coefficients for the corresponding indicators. For example, Hand Movements is not specified to measure simultaneous processing (Figure 9.7). Structure coefficients for this task are .497 and .277. The .497 coefficient equals the standardized pattern coefficient for this indicator of the sequential processing factor (Table 13.3). The .277 coefficient is the predicted correlation between the Hand Movements task and the simultaneous processing factor. It is calculated by applying the tracing rule to the path

Hand Movements  $\longleftrightarrow$  Sequential Processing  $\curvearrowright$  Simultaneous Processing

which is estimated as  $.497(.577) = .277$ , where the second operand is the factor correlation (Table 13.3). Exercise 3 asks you to calculate structure coefficients for the other indicators in Table 13.4 using the tracing rule. These results indicate that structure

---

<sup>3</sup>Close-fit hypothesis,  $H_0: \epsilon_0 \leq .05, \epsilon_1 = .08$ ; not-close-fit hypothesis,  $H_0: \epsilon_0 \geq .05, \epsilon_1 = .01; \alpha = .05$  for both.

**TABLE 13.3. Maximum Likelihood Estimates for a Two-Factor Model of the KABC-I**

Parameter	Unstandardized		Standardized	
	Estimate	SE	Estimate	SE
<u>Pattern coefficients</u>				
<u>Sequential factor</u>				
Hand Movements	1.000	—	.497	.062
Number Recall	1.147	.181	.807	.046
Word Order	1.388	.219	.808	.046
<u>Simultaneous factor</u>				
Gestalt Closure	1.000	—	.503	.061
Triangles	1.445	.227	.726	.044
Spatial Memory	2.029	.335	.656	.050
Matrix Analogies	1.212	.212	.588	.055
Photo Series	1.727	.265	.782	.040
<u>Error variances</u>				
Hand Movements	8.664	.938	.753	.061
Number Recall	1.998	.414	.349	.075
Word Order	2.902	.604	.347	.075
Gestalt Closure	5.419	.585	.747	.061
Triangles	3.425	.458	.472	.064
Spatial Memory	9.998	1.202	.570	.065
Matrix Analogies	5.104	.578	.654	.065
Photo Series	3.483	.537	.389	.063
<u>Factor variances and covariance</u>				
Sequential	2.839	.838	1.000	—
Simultaneous	1.835	.530	1.000	—
Sequential ↗ Simultaneous	1.271	.324	.557	.067

Note. KABC-I, Kaufman Assessment Battery for Children, first edition. Standardized estimates for error variances are proportions of unexplained variance. All results were computed by Mplus for which the standardized solution is STDYX.

**TABLE 13.4. Structure Coefficients for a Two-Factor Model of the KABC-I**

Indicator	Factor	
	Sequential	Simultaneous
Hand Movements	<b>.497</b>	.277
Number Recall	<b>.807</b>	.449
Word Order	<b>.808</b>	.450
Gestalt Closure	.280	<b>.503</b>
Triangles	.405	<b>.726</b>
Spatial Memory	.365	<b>.656</b>
Matrix Analogies	.327	<b>.588</b>
Photo Series	.435	<b>.782</b>

Note. KABC-I, Kaufman Assessment Battery for Children, first edition. All results were computed by Mplus.

coefficients are not typically zero for corresponding zero pattern coefficients when the factors covary.

The two-factor model in Figure 9.7 implies that (1) each pair of indicators of the same factor is d-separated by that factor and (2) each pair of indicators where one member measures one factor but the other measures the other factor is d-separated by both factors. Reported in Table 13.5 are values of partial correlations that correspond to the basis set (Rule 8.2) just described. Absolute partial correlations  $> .10$  are shown in boldface in the table. Most of the results just mentioned are associated with two indicators, Hand Movements and Number Recall, both of which are specified to measure sequential processing. For both indicators, the model poorly explains their associations with four out of five indicators of simultaneous processing.

Correlation residuals are reported in the top part of Table 13.6, and standardized residuals are reported in the bottom part. Absolute correlation residuals  $\geq .10$  and standardized residuals that are significant at the .05 level ( $> 1.96$ ) are shown in boldface. Many of the results just mentioned concern Hand Movements and most of the indicators of the other factor. All these residuals are positive, which says that the model underestimates the corresponding sample associations. Given all the results considered so far, the two-factor model in Figure 9.7 is rejected.

## RESPECIFICATION OF CFA MODELS

In the face of adversity, the protagonist of Kurt Vonnegut's novel *Slaughterhouse-Five* often remarks, "So it goes." And so it often goes in CFA that the initial model is rejected. Respecification is more challenging for a CFA model than for path models because there are more possibilities for change. Thus, respecification in CFA should be guided

**TABLE 13.5. Partial Correlations for Conditional Independences of a Basis Set for a Two-Factor Model of the KABC-I**

Indicator	1	2	3	4	5	6	7	8
<u>Sequential scale</u>								
1. Hand Movements	—							
2. Number Recall	-.022	—						
3. Word Order	<b>-.101</b>	.052	—					
<u>Simultaneous scale</u>								
4. Gestalt Closure	.094	<b>-.227</b>	<b>-.130</b>	—				
5. Triangles	<b>.199</b>	<b>-.139</b>	-.092	.025	—			
6. Spatial Memory	<b>.334</b>	-.009	-.033	-.046	-.012	—		
7. Matrix Analogies	<b>.324</b>	<b>.118</b>	.075	.020	-.012	.040	—	
8. Photo Series	<b>.321</b>	<b>-.165</b>	.051	.049	.029	-.006	-.079	—

Note. KABC-I, Kaufman Assessment Battery for Children, first edition. These results were computed in SPSS.

as much as possible by substantive considerations; otherwise, respecification could put the researcher in the same situation as the sailor in this adage attributed to Leonardo da Vinci: One who loves practice without theory is like a sailor who boards a ship without a rudder and compass and never knows where he or she may be cast.

The first category of possible changes involves the indicators. Sometimes an indicator fails to have a substantial pattern coefficient for the factor it is supposed to measure. One option is to specify that the indicator measures a different factor. Inspection of the residuals can help to identify the other factor to which the indicator's correspondence may be switched. Suppose that the residuals between an indicator of factor A and those for indicators of factor B are large and positive. This suggests that the indicator may measure factor B more than it does factor A. Note that an indicator can have a relatively high pattern coefficient for its own factor but also have high residuals between it and the indicators of another factor. This pattern suggests that the indicator may measure both factors. Another prospect is that these indicators share something that is unique to them, such as stimuli or informants (e.g., parents), but unrelated to the factors. Shared unique variance is represented by specifying correlated errors.

The second class of possible changes involves the factors. The researcher may have specified the wrong number of factors. For example, poor discriminant validity is evidenced by very high factor correlations (there are too many factors), but poor convergent validity within sets of indicators for the same factor suggests that the model may have too few factors. Changing the number of factors is a more radical change than adjusting factor-indicator correspondence or error correlations; that is, the original hypotheses about measurement were very wrong.

A starting point for respecification often includes inspection of the residuals and modification indexes. Earlier we examined the residuals in Table 13.6 for the two-factor

model of the KABC-I. Most of the larger and positive residuals are between Hand Movements and other tasks specified to measure the other factor. Because the standardized pattern coefficient of Hand Movements is at least moderate (.497; Table 13.3), it is possible that this task may measure both factors. Reported in Table 13.7 are the 10 largest modification indexes for unstandardized pattern coefficients and error covariances fixed to zero in the original model (Figure 9.7). Note in the table that the  $\chi^2(1)$  statistics for the paths

Simultaneous  $\rightarrow$  Hand Movements and  $E_{NR} \curvearrowleft E_{WO}$

**TABLE 13.6. Correlation Residuals and Standardized Residuals for a Two-Factor Model of the KABC-I**

Indicator	1	2	3	4	5	6	7	8
<u>Correlation residuals</u>								
<u>Sequential scale</u>								
1. Hand Movements	0							
2. Number Recall	-.011	0						
3. Word Order	-.052	.018	0					
<u>Simultaneous scale</u>								
4. Gestalt Closure	.071	<b>-.116</b>	-.066	0				
5. Triangles	<b>.119</b>	-.056	-.037	.015	0			
6. Spatial Memory	<b>.218</b>	-.005	-.015	-.030	-.007	0		
7. Matrix Analogies	<b>.227</b>	.056	.035	.014	-.007	.024	0	
8. Photo Series	<b>.174</b>	-.061	.018	.027	.012	-.003	-.040	0
<u>Standardized residuals</u>								
<u>Sequential scale</u>								
1. Hand Movements	—							
2. Number Recall	-.595	—						
3. Word Order	<b>-3.803</b>	1.537	—					
<u>Simultaneous scale</u>								
4. Gestalt Closure	1.126	<b>-2.329</b>	-1.315	—				
5. Triangles	<b>2.046</b>	-1.558	-1.001	.427	—			
6. Spatial Memory	<b>3.464</b>	-.112	-.354	-.785	-.268	—		
7. Matrix Analogies	<b>3.505</b>	1.129	.727	.323	-.246	.664	—	
8. Photo Series	<b>2.990</b>	<b>-2.001</b>	.524	.909	.676	-.144	<b>-1.978</b>	—

Note. KABC-I, Kaufman Assessment Battery for Children, first edition. The correlation residuals were computed by EQS, and the standardized residuals were computed by Mplus.

**TABLE 13.7. The Ten Largest Modification Indexes for a Two-Factor Model of the KABC-I**

Path	MI	<i>p</i>
Simultaneous → Hand Movements	20.091	< .001
$E_{NR} \curvearrowright E_{WO}$	20.042	< .001
$E_{HM} \curvearrowright E_{WO}$	7.015	.008
Simultaneous → Number Recall	7.010	.008
$E_{HM} \curvearrowright E_{SM}$	4.847	.028
$E_{HM} \curvearrowright E_{MA}$	3.799	.051
Sequential → Matrix Analogies	3.247	.072
$E_{NR} \curvearrowright E_{PS}$	3.147	.076
Sequential → Gestalt Closure	2.902	.089
$E_{MA} \curvearrowright E_{PS}$	2.727	.099

Note. KABC-I, Kaufman Assessment Battery for Children, first edition; MI, modification index; NR, Number Recall; WO, Word Order; HM, Hand Movements; SM, Spatial Memory; MA, Matrix Analogies; PS, Photo Series. All results were computed by Mplus.

are nearly identical, respectively, 20.091 and 20.042. This means that allowing Hand Movements to also depend on the simultaneous processing factor or adding an error correlation between Number Recall and Word Order would reduce  $\chi^2_M$  by about 20 points. Among other changes suggested by the modification indexes, two have nearly the same value: Allow the errors of the Hand Movements and Word Order tasks to covary (7.015) or to respecify Number Recall as a complex indicator (7.010). Based on my knowledge of the KABC-I (Kline, Snyder, & Castellanos, 1996) and results of other factor-analytic studies (e.g., Keith, 1985), specifying that Hand Movements measures both factors is plausible. Kline (2012) describes the analysis of the unrestricted two-factor model of the KABC-I for the same data in EFA.

## SPECIAL TOPICS AND TESTS

Score reliability coefficients for observed variable analyses were reviewed in Chapter 4. Coefficients that estimate the reliability of factor measurement in SEM are described in Topic Box 13.1. Exercise 4 asks you to calculate a reliability coefficient for the simultaneous processing factor in Figure 9.7 based on the results in Table 13.3.

The choice between analyzing factors in unstandardized versus standardized form usually does not affect model fit. Steiger (2002) describes an exception called **constraint interaction** that can occur for CFA models where some factors have only two indicators and a **cross-factor equality constraint** is imposed on the coefficients for indicators of different factors. In some cases the value of  $\chi^2_D(1)$  for the test of the equality constraint depends on how the factors are scaled. Constraint interaction probably does not occur

### TOPIC BOX 13.1

#### Reliability of Factor Measurement

Raykov (2004) and Hancock and Mueller (2001) describe coefficients that estimate the reliability of factor measurement in CFA models or SR models where reflective measurement is specified. These coefficients are generally better alternatives than Cronbach's alpha (Equation 4.7), which does not directly measure whether the indicators depend on a single factor. The **composite reliability** (CR), also called the **factor rho coefficient**, is the ratio of explained variance over total variance. For factors with no error correlations between its indicators, the composite reliability is estimated in the unstandardized solution as

$$CR = \frac{(\sum \hat{\lambda}_i)^2 \hat{\phi}}{(\sum \hat{\lambda}_i)^2 \hat{\phi} + \sum \hat{\theta}_{ii}} \quad (13.1)$$

where  $\sum \hat{\lambda}_i$  is the sum of the unstandardized pattern coefficients among indicators of the same factor,  $\hat{\phi}$  is the estimated factor variance, and  $\sum \hat{\theta}_{ii}$  is the sum of the unstandardized error variances. A different formula is needed for when indicators share at least one error covariance:

$$CR = \frac{(\sum \hat{\lambda}_i)^2 \hat{\phi}}{(\sum \hat{\lambda}_i)^2 \hat{\phi} + \sum \hat{\theta}_{ii} + 2 \sum \hat{\theta}_{ij}} \quad (13.2)$$

where  $\sum \hat{\theta}_{ij}$  is the sum of the nonzero unstandardized error covariances. Other variations of these equations are described by Raykov (2004) and Hancock and Mueller (2001).

A computationally simpler alternative is the **average variance extracted** (AVE), which is just the average of the squared standardized pattern coefficients for indicators that depend on the same factor but are specified to measure no other factors (they are simple indicators). Because AVE is based on standardized coefficients, its values may not be directly comparable for the same indicators across different samples. In this case, CR would be preferred because its equation requires unstandardized coefficients that are more directly comparable over samples.

Next we calculate CR for the three indicators of the sequential processing factor in the two-factor model of Figure 9.7, given the results in Table 13.3. There are no error correlations in the model, so we need Equation 13.1:

$$\begin{aligned} \sum \hat{\lambda}_i &= 1.000 + 1.147 + 1.388 = 3.535 \\ \hat{\phi} &= 2.839 \end{aligned}$$

$$\sum \hat{\theta}_{ii} = 8.664 + 1.998 + 2.902 = 13.564$$

$$CR = \frac{3.535^2(2.839)}{3.535^2(2.839) + 13.564} = .723$$

which is not a terrible result, but still the evidence for convergent validity among the three indicators of this factor is dubious (see Table 13.3). Bentler (2009) describes coefficients within the CFA framework that are suitable for ordinal data, such as when the indicators are binary or polytomous items.

in most CFA applications, but you should know about this phenomenon in case it ever crops up in your own work—see Appendix 13.B.

Remember that equality-constrained pattern coefficients are equal in the unstandardized solution, but the corresponding standardized coefficients are typically unequal. *Thus, it usually makes no sense to compare standardized estimates for equality-constrained pattern coefficients.* If it is really necessary to constrain a pair of standardized coefficients to equality, then one option is to analyze a correlation matrix using the method of constrained estimation.

Whether indicators are congeneric, tau-equivalent, or parallel can be tested by comparing hierarchical models with the chi-square difference test. **Congeneric indicators** measure the same construct but not necessarily to the same degree. The congenericity model imposes no constraints except the specification that a set of indicators depends on the same factor. If this model is retained, next test for tau equivalence. **Tau-equivalent indicators** are congeneric and have equal true score variances. This hypothesis is tested by fixing all unstandardized pattern coefficients to 1.0. If the fit of the tau equivalence model is not appreciably worse than that of the congenericity model, next test for parallelism. **Parallel indicators** have equal error variances. If the fit of the model with equality-constrained errors is not appreciably worse than that of the tau-equivalence model, the indicators may be parallel. All these models assume independent errors and must be fitted to a covariance matrix, not a correlation matrix; see Brown (2015, chap. 7) for examples.

Merging all factors in a multifactor model generates a single-factor model that is nested under the original. The comparison with the chi-square difference test just described is the **test for redundancy**. A variation is to fix the covariances between multiple factors to zero, which provides a **test for orthogonality**. This procedure is unnecessary for models with two factors because the significance test of the factor covariance in the unconstrained model provides the same information. For models with three or more factors, the test for orthogonality is akin to a multivariate test for whether all factor covariances together differ from zero. Each factor should have at least three indicators for the redundancy test; otherwise, the constrained model may not be identified.

**Vanishing tetrads** are a kind of overidentifying restriction for factors with at least four continuous indicators and no error correlations. In his work on factor analysis, Spearman (1904) showed that differences in products of covariances or correlations between certain pairs of indicators must be zero, if all indicators depend on the same factor. For indicators  $X_1$ – $X_4$  of the same factor, there are three vanishing tetrads in a correlation metric:

$$\begin{aligned}\rho_{12}\rho_{13} - \rho_{13}\rho_{24} &= 0 \\ \rho_{12}\rho_{34} - \rho_{14}\rho_{23} &= 0 \\ \rho_{13}\rho_{24} - \rho_{14}\rho_{23} &= 0\end{aligned}\tag{13.3}$$

where the symbol  $\rho_{ij}$  represents the population correlation between indicators  $X_i$  and  $X_j$ , and so on. If any two of the restrictions hold in Equation 13.3, the third must also be true, so there are actually just two independent overidentifying restrictions. For each factor with  $k \geq 4$  indicators, there are a total of  $k(k-3)/2$  independent overidentifying restrictions. Kenny and Milan (2012) describe additional types of between- or within-factor vanishing tetrads for CFA models.

Spirites, Glymour, and Scheines (2001) describe **exploratory tetrad analysis** (ETA), a computer-based search algorithm that tries to locate unidimensional, single-factor measurement models based on observed vanishing tetrads among four or more indicators. The freely available computer program TETRAD V (Glymour, Scheines, Spirtes, Ramsey, 2014) implements related search methods.<sup>4</sup> Bollen and Ting (1993) describe **confirmatory tetrad analysis** (CTA) which, unlike ETA, requires a priori measurement models. Both local and global fit can be evaluated by examining vanishing tetrads for normal or non-normal variables. Because vanishing tetrads are estimated from observed variables, the technique of CTA may be useful for analyzing measurement models that are not identified in the CFA/SEM framework.

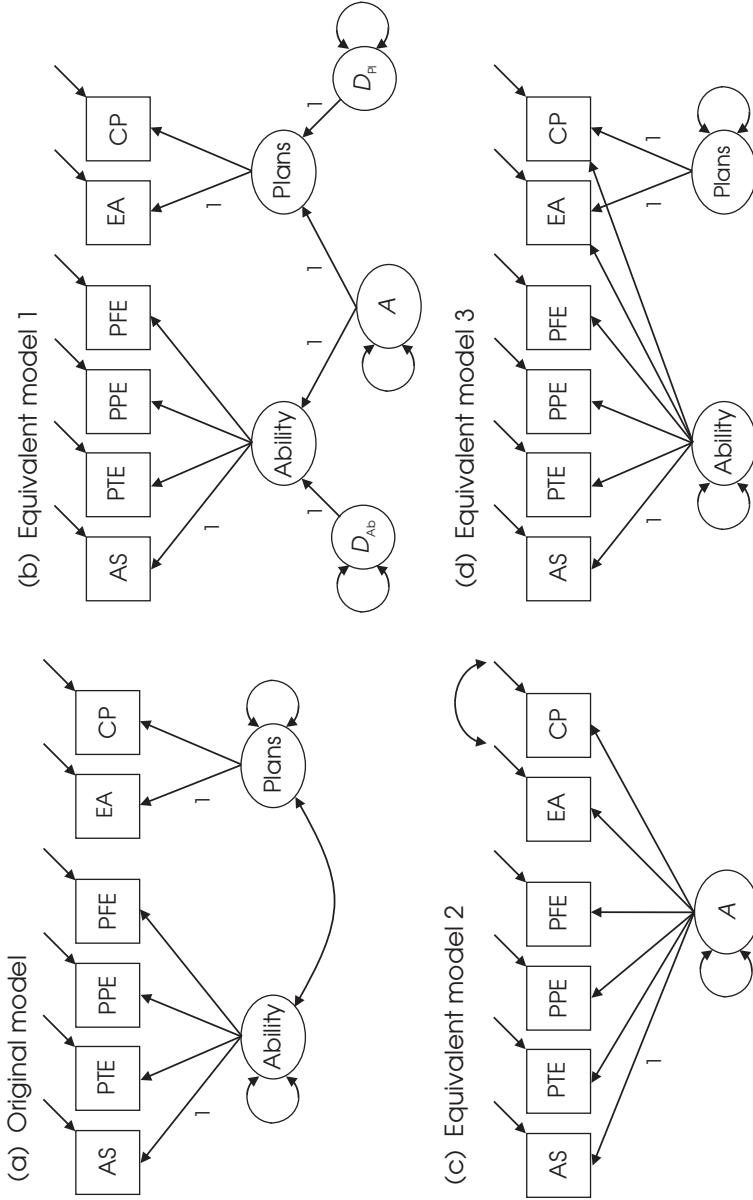
## EQUIVALENT CFA MODELS

There are two sets of principles for generating equivalent CFA models—one for models with multiple factors and another for single-factor models. As an example of the first set, consider the two-factor model of self-perception of ability and achievement by Kenny (1979) presented in Figure 13.1(a) with compact symbolism for error terms. I used the method of constrained ML estimation in the SEPATH procedure of STATISTICA Advanced (StatSoft, 2013) to fit this model to the correlation matrix reported in a sample of 556 Grade 8 students (presented in Table 13.8). Values of selected fit statistics indicate reasonable global fit:

$$\chi^2_M(8) = 9.256, p = .321$$

---

<sup>4</sup>[www.phil.cmu.edu/tetrad](http://www.phil.cmu.edu/tetrad)



**FIGURE 13.1.** Four equivalent measurement models of self-perceived ability and educational plans shown with compact symbolism for error terms. AS, Ability Self-Concept; PTE, Perceived Teacher Evaluation; PPE, Perceived Parental Evaluation; PFE, Perceived Friends' Evaluation; EA, Educational Aspiration; CP, College Plans.

RMSEA = .012, 90% CI [.017, .054]

CFI = .999; SRMR = .012

The other three CFA models in Figure 13.1 are equivalent versions of the original model that yield the same values of fit statistics and predicted correlations. Figure 13.1(b) is a hierarchical CFA model in which the covariance between the factors of the original model is replaced by a second-order factor (*A*), which has no indicators and is presumed to directly affect the first-order factors (Ability, Plans). This specification provides an account of why the two first-order factors (which are endogenous in this model) covary. Because the second-order factor has only two indicators, it is necessary to constrain its unstandardized direct effects on the first-order factors to be equal. The other equivalent versions are unique to models wherein some factors have only two indicators. Figure 13.1(c) features the substitution of the Plans factor with a correlation between the errors of its indicators. Figure 13.1(d) features replacement of the factor correlation with the specification that two indicators depend on both factors, which explains the sample correlations just as well as the original model. The pattern coefficients for the indicators of the Plans factor must be equal in order for this model to be identified.

The situation regarding equivalent versions of CFA models with multiple factors is even more complex than suggested by the last example. It is possible to apply the replacing rules (Table 12.6) to substitute factor correlations with direct effects, which makes some factors endogenous. The resulting model is an SR model, but it will fit the data equally well. For example, substitution of the factor correlation in Figure 13.1(a) with a direct effect generates an equivalent SR model. Raykov and Marcoulides (2001) show that there are infinitely many equivalent versions of standard CFA models. For each equivalent model, the factor correlations are eliminated (orthogonality is specified) and replaced by one or more factors not represented in the original model with fixed unit pattern coefficients (1.0) for all indicators. These models with additional factors explain the data just as well as the original.

Equivalent versions of single-factor CFA models can be derived using Hershberger and Marcoulides's (2013) **reversed indicator rule**, where one indicator is specified as

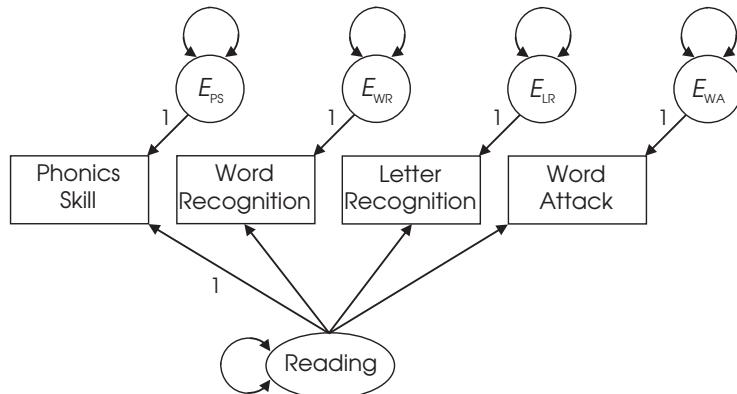
**TABLE 13.8. Input Data (Correlations) for Analysis of a Two-Factor Model of Perceived Ability and Educational Plans**

Variable	1	2	3	4	5	6
1. Ability Self-Concept	1.00					
2. Perceived Parental Evaluation	.73	1.00				
3. Perceived Teacher Evaluation	.70	.68	1.00			
4. Perceived Friends' Evaluation	.58	.61	.57	1.00		
5. Education Aspiration	.46	.43	.40	.37	1.00	
6. College Plans	.56	.52	.48	.41	.71	1.00

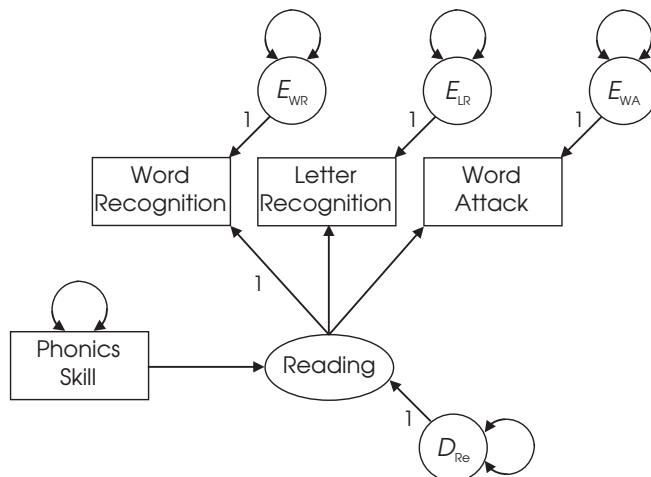
*Note.* Input data are from Kenny (1979); *N* = 556.

causal and the others remain as effect indicators. Consider the CFA model of reading in Figure 13.2(a). The effect indicators represent phonics, word and letter recognition, and word attack. The equivalent version in Figure 13.2(b) specifies phonics as a *cause* of the reading factor, which is now endogenous and thus has a disturbance. In addition, the causal phonics indicator is exogenous. Figure 13.2(b) is actually an SR model. A total of three other equivalent models could potentially be generated, one with each of the remaining effect indicators respecified as causes. Not all of the equivalent versions may be theoretically plausible, but the one with phonics as a causal indicator is logical. Figure 13.2(b) has a **multiple-indicators and multiple-causes** (MIMIC) factor with both causal indicators and effect indicators. Models with MIMIC factors are SR models, not

(a) Original model with effect indicators



(b) Equivalent model with a causal indicator

**FIGURE 13.2.** Application of the reversed indicator rule to generate an equivalent one-factor model of reading.

CFA models, because MIMIC factors are always endogenous. The analysis of SR models with MIMIC factors is covered in the next chapter.

## SPECIAL CFA MODELS

This section describes three special kinds of CFA models. The first two types are for hierarchically structured constructs, and the third concerns models with features that represent method effects.

### Hierarchical Models

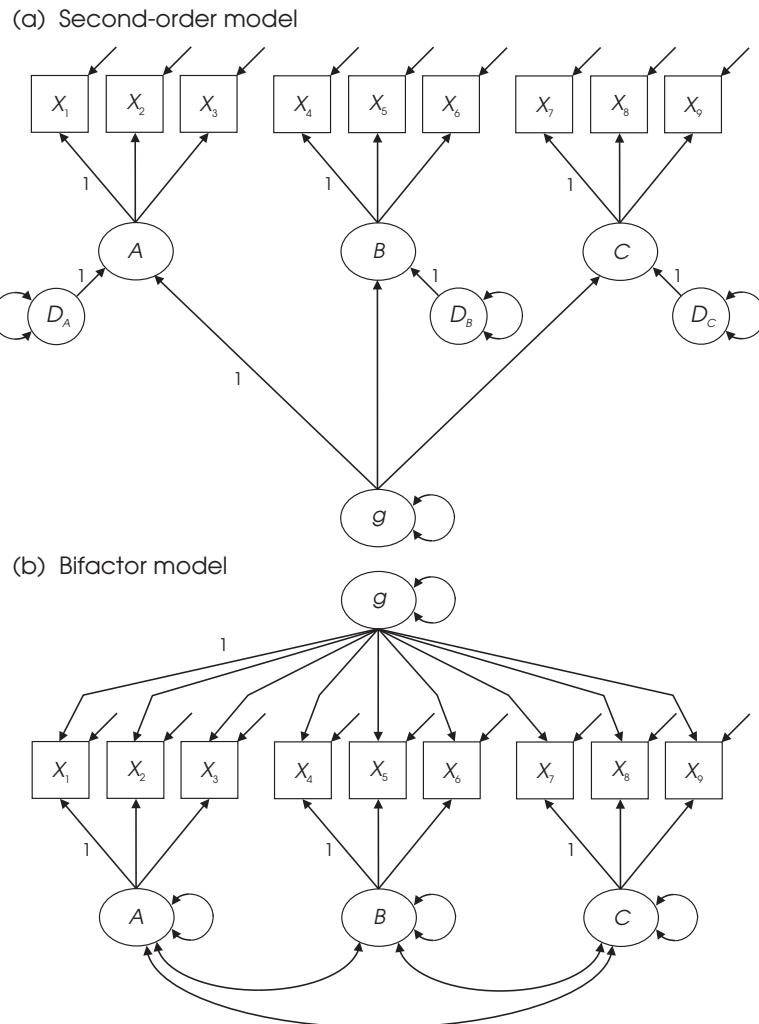
The **second-order model** in Figure 13.3(a) shown with compact symbolism for error terms represents the hypothesis that a general **second-order factor**,  $g$ , causes each of three **first-order factors**, A–C. The first-order factors have indicators, but the general factor has none; that is, the second-order factor is measured only indirectly through the indicators of the first-order factors. The specification of  $g$  as a common cause of A–C in the figure implies that associations between the first-order factors are spurious (i.e.,  $g$  explains why the factors covary). The other direct cause of each first-order factor is a disturbance, which represents variation not explained by  $g$ .

To identify a second-order CFA model, there must be at least three first-order factors or their disturbance variances may be underidentified. Each first-order factor should have at least two indicators, but having more is better. Figure 13.3(a) satisfies both of these requirements. The general factor in the figure is scaled by fixing the direct effect of  $g$  on A to 1.0. Another option is to fix the variance of  $g$  to 1.0 (standardize it). This approach leaves all three direct effects of  $g$  on the first-order factors as free parameters. Either way of scaling  $g$  in a single-sample analysis is probably fine, but it is usually inappropriate to standardize factors in multiple-samples analyses. There are examples of second-order CFA models in assessment research where  $g$  is conceptualized as a superordinate ability factor that affects more specific skills such as verbal reasoning or memory (Williams, McIntosh, Dixon, Newton, & Youman, 2010). Second-order CFA models are also analyzed in studies of personality and quality of life, among others.

### Bifactor Models

Bifactor models—also called **nested-factor models** or **general-specific models**—may be less familiar than hierarchical models, but they also concern situations where several correlated specific constructs make up a more general construct of interest (Chen, West, & Sousa, 2006). The main difference between a second-order model and a bifactor model is that the general factor in the bifactor model directly affects the indicators but is orthogonal to the specific factors. Bifactor models where the general factor covaries with the specific factors may not be identified.

Presented in Figure 13.3(b) is a bifactor model where the general factor ( $g$ ) causes all indicators and is unrelated to the specific factors A–C. This model partitions indicator variance into three nonoverlapping sources, the specific factors, the general factor, and error. Because the specific and general factors in a bifactor model are orthogonal, it is actually the disturbances in a second-order model that resemble the specific factors in a bifactor model. For example, terms  $D_A$ – $D_C$  in the second-order model of Figure 13.3(a) correspond to the specific factors A–C in the bifactor model of Figure 13.3(b) in that both sets of variables are independent of the general factor. Also, specific factors in a bifactor model are not intervening variables, but first-order factors in second-order models are always specified as mediators.



**FIGURE 13.3.** A second-order model (a) and a bifactor model (b) shown with compact symbolism for indicator error terms.

Chen et al. (2006) described a bifactor model of life quality where the indicators reflect both general quality of life and adjustment in more specific domains, such as mental health and physical health. The general quality-of-life factor is unrelated to the domain-specific factors. In contrast, a second-order model would assume that a general quality-of-life factor causes each of the domain-specific (first-order) factors. Thus, the two models make very different assumptions about whether the general factor is unrelated to the other factors (bifactor model) or covaries with—and actually causes—those other factors (second-order model).

Because the general factor in a second-order model has no indicators, it may be more difficult to interpret than the general factor in a bifactor model, where all observed variables are indicators of the general factor (Gignac, 2008). Another advantage of bifactor models is that the predictive validity of specific factors, which are independent of the general factor, can be directly estimated by “embedding” a bifactor model in a larger SR model where outcomes are predicted by specific versus general factors (Chen et al., 2006). It is trickier to do so for second-order models where the general and first-order factors overlap; see Chen et al. (2006) and Gignac (2008) for more information.

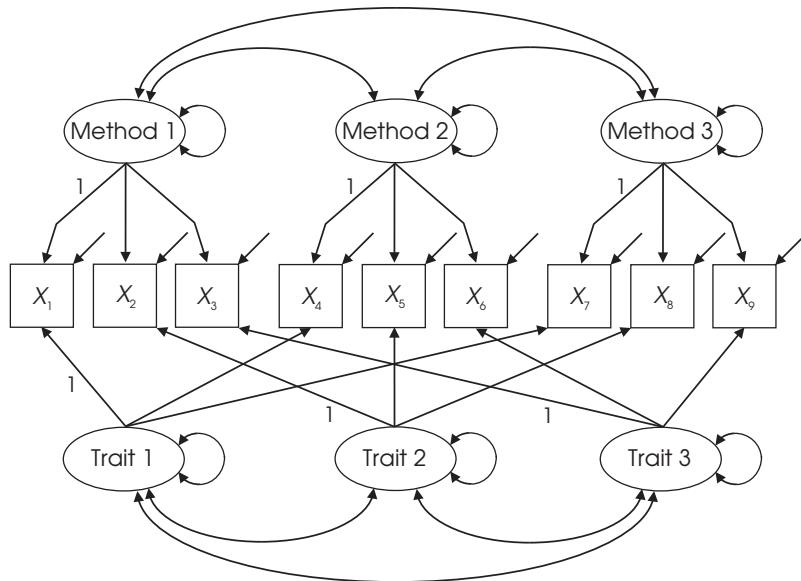
## Models for Multitrait–Multimethod Data

The method of CFA can also be used to analyze data from a **multitrait–multimethod (MTMM) study**, the logic of which was first articulated by Campbell and Fiske (1959). In such a study, two or more traits are measured with two or more methods. Traits are hypothetical constructs about stable characteristics, such as cognitive abilities, and methods refer to multiple-test forms, occasions, ways of collecting data (e.g., self-report), or informants (e.g., teachers). The goals are to (1) evaluate the convergent validity and discriminant validity of tests that vary in their measurement method and (2) derive separate estimates of the effects of traits versus methods on the observed scores.

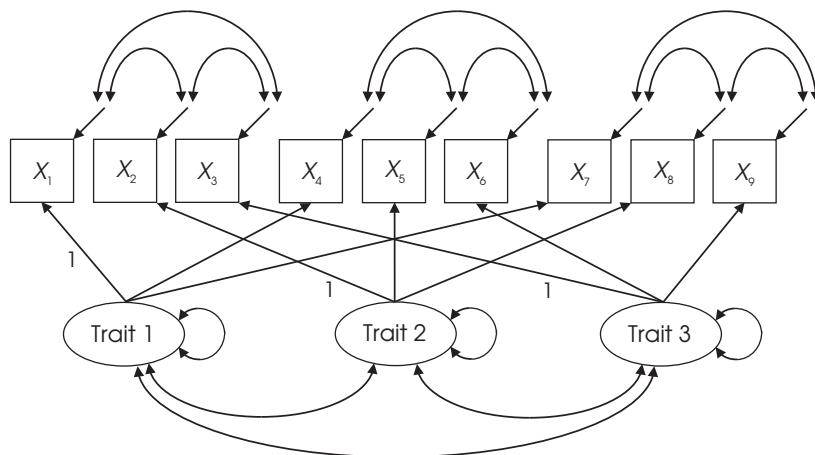
The earliest procedure for analyzing data from an MTMM study involved inspection of the correlation matrix for all variables. For example, convergent validity would be indicated by the observation of high correlations among variables that supposedly measure the same trait but with different methods. If correlations among variables that should measure different traits but use the same method are relatively high, then common method effects are indicated. This implies that correlations between different variables based on the same method may be relatively high even if they measure unrelated traits.

When CFA was first applied to the problem in the 1970s, researchers typically specified models like the one presented in Figure 13.4(a), a **correlated trait–correlated method (CTCM) model**. Such models have separate trait and method factors that are assumed to covary, but method factors are assumed to be independent of trait factors. In the figure, indicators  $X_1$ – $X_3$  are based on one method,  $X_4$ – $X_6$  are based on another method, and  $X_7$ – $X_9$  are based on a third method. This model also specifies that the set of indicators ( $X_1$ ,  $X_4$ ,  $X_7$ ) measures one trait but that each of the other two sets, ( $X_2$ ,  $X_5$ ,  $X_8$ ) and ( $X_3$ ,  $X_6$ ,  $X_9$ ), measures a different trait. Given these specifications, relatively high

(a) Correlated trait–correlated method model



(b) Correlated–uniqueness model



**FIGURE 13.4.** A correlated trait–correlated method model (a) and a correlated–uniqueness model (b) for multitrait–multimethod data shown with compact symbolism for indicator error terms.

pattern coefficients for trait factors would suggest convergent validity, high coefficients for method factors would indicate common method effects, and moderate correlations (not too high) between the factors would indicate discriminant validity.

There are reports of “successful” analyses of CTCM models in the literature, but others have found that such analyses tend to yield inadmissible or unstable solutions. For example, Marsh and Bailey (1991) found in computer simulations that illogical estimates were derived about 75% of the time for CTCM models. Kenny and Kashy (1992) noted part of the problem: CTCM models are not identified if the pattern coefficients on the trait or method factors are equal. If the pattern coefficients are different but similar in value, then CTCM models may be empirically underidentified.

Some simpler alternatives to CTCM models have been proposed, including those with multiple but uncorrelated method factors and a model such as the one in Figure 13.4(b), which is a **correlated-uniqueness (CU) model** (Marsh & Grayson, 1995). This model has error correlations among indicators based on the same method instead of separate method factors. Method effects are assumed to be a property of each indicator, and relatively high correlations among their residuals are taken as evidence for common method effects. Saris and Alberts (2003) describe alternatives that could account for correlated residuals in CU models, including models that represent response biases, effects due to relative answers (when respondents compare their answers), and method effects; see Eid et al. (2008) for more information.

Suppose that all indicators rely on a common measurement method and there is concern about common method variance. Antonakis et al. (2010) remind us that specifying what is essentially a bifactor model where all indicators depend on a general “method” factor in addition to specific factors that represent substantive latent variables does *not* generally control for common method variance. This is because it is actually impossible to estimate the exact effect of a common method without measuring markers of the common source variable and including them in the model. Such markers are theoretically unrelated to at least one substantive latent variable but may be affected by common method variance. An example of a marker for social desirability bias is the Marlowe–Crowne Social Desirability Scale (Crowne & Marlowe, 1960). Multiple markers should reflect various facets of common method effects, and in CFA they are specified as the indicators of a common method factor; see Richardson, Simmering, and Sturman (2009) for more information.

## ANALYZING LIKERT-SCALE ITEMS AS INDICATORS

Estimation methods for continuous variables are not the best choice when the indicators are Likert-scale items with a relatively small number of categories (e.g., five or fewer) or response distributions are severely asymmetrical. Described next are alternative estimators for CFA models with ordered-categorical (ordinal) indicators. The first method is robust weighted least squares (WLS) estimation, which is a computationally simpler version of full WLS estimation. This method makes no distributional assumptions but

requires very large samples. The robust WLS method is available in several SEM computer tools, and its use with noncontinuous outcomes is increasingly being described in the literature.

### Robust WLS Estimation

The logic of WLS estimation (full or robust) of CFA models with ordinal indicators is based on an approach to estimating structural equation models with any combination of ordinal, nominal, or continuous outcomes known as **continuous/categorical variable methodology** (Muthén, 1984). In this framework, each ordinal indicator is associated with a **latent response variable**, which is the underlying amount of a continuous and normally distributed continuum that is required to respond in a certain way on the corresponding indicator. For dichotomous items, this amount or **threshold** is the point on the latent variable where one response option is given (e.g., true), if the threshold is exceeded. It is also the point where the other response option is given (e.g., false), if the threshold is not exceeded.

Dichotomous items have a single threshold, but the number of thresholds for polytomous items with three or more response categories equals the number of categories minus one. Suppose that item  $X$  has the 3-point Likert response format listed next:

$$1 = \text{disagree}, 2 = \text{neutral}, 3 = \text{agree}$$

The scale just described can be viewed as a crude categorization of  $X^*$ , the underlying latent response variable. Item  $X$  has two threshold parameters,  $\tau_1$  and  $\tau_2$ , where  $\tau$  is the lowercase Greek letter tau. When  $X^*$  has a mean of 0 and a variance of 1.0, thresholds are values of the normal deviate ( $z$ ) that divide a normal distribution into categories and thus relate discrete responses on  $X$  to continuous  $X^*$  values. Specifically, the observed data are considered to be generated as follows:

$$X = \begin{cases} 1, & \text{if } X^* \leq \tau_1; \\ 2, & \text{if } \tau_1 < X^* \leq \tau_2; \\ 3, & \text{if } X^* > \tau_2. \end{cases} \quad (13.4)$$

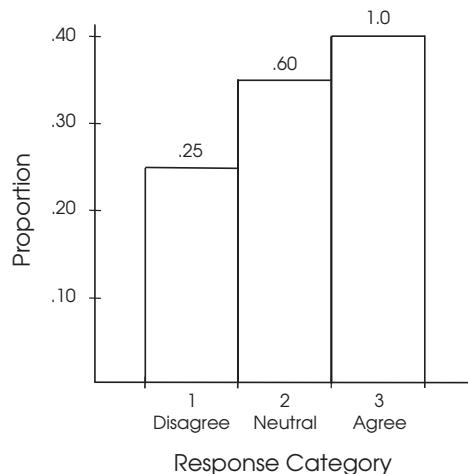
In words, an observed response of “1” (*disagree*) is expected if the level of  $X^*$  is less than that of  $\tau_1$  in standard deviation units. For levels of  $X^*$  greater than  $\tau_1$  but less than or equal to  $\tau_2$ , the predicted response is “2” (*neutral*), and  $X^* > \tau_2$  corresponds to a response of “3” (*agree*) on item  $X$ . An example follows.

Presented in Figure 13.5(a) is the histogram of responses to hypothetical item  $X$  with a 3-point Likert scale. Cumulative probabilities over the three categories are also shown in the figure. Thresholds are estimated based on cumulative response probabilities. For example, the cumulative probability for endorsing “1” (*disagree*) is .25, so  $\hat{\tau}_1 = -.67$ , which is the value of the normal deviate that falls at about the 25th percentile and is shown in Figure 13.5(b). The same result also says that responses to item  $X$  are

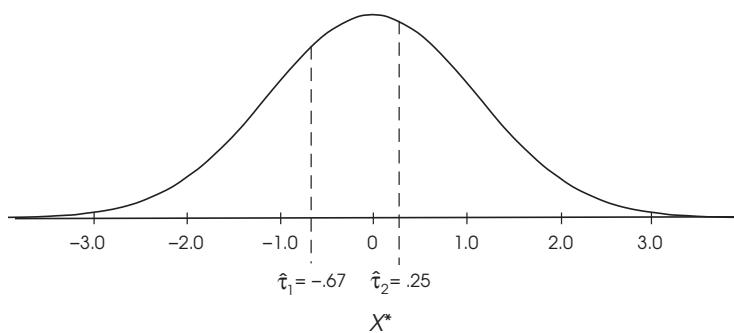
expected to switch from “1” (*disagree*) to “2” (*neutral*) when the level of  $X^*$  increases to about two-thirds below the mean. Exercise 5 asks you to interpret the result  $\hat{\tau}_2 = .25$  in Figure 13.5(b), given the cumulative probability of .60 over the response categories of “1” (*disagree*) and “2” (*neutral*) together.

For a set of items, estimated thresholds and observed cross tabulations of item responses are used by the computer to estimate the matrix of Pearson correlations between the latent response variables. For a pair of dichotomous items, this estimated correlation is a tetrachoric correlation; otherwise, the estimated correlation is a polychoric correlation. The polychoric correlation is the most general estimated Pearson correlation, so just the term *polychoric correlation* is used from this point. Next, the computer generates the **asymptotic covariance matrix** (i.e., information matrix) of the polychoric correlations, the inverse of which is the weight matrix in full WLS estimation. Diagonal elements in the asymptotic covariance matrix estimate the variance of

(a) Histogram of item  $X$  responses with cumulative probabilities



(b) Latent response variable with threshold estimates



**FIGURE 13.5.** Histogram for observed responses on hypothetical item  $X$  with a 3-point Likert scale (a) and the corresponding latent response variable  $X^*$  with estimated thresholds,  $\hat{\tau}$  (b).

the polychoric correlations over random samples, and the off-diagonal elements represent the covariances between these estimates. Robust WLS estimation uses just the diagonal of the whole asymptotic covariance matrix (i.e., the error variances) in its fit function, which also includes the polychoric correlation matrix and thresholds depending on the computer program.<sup>5</sup>

The measurement model analyzed consists of the latent response variables as the continuous indicators of the common factors. Both the number of factors and the correspondence between indicators and factors are specified by the researcher just as in “regular” CFA for a linear measurement model (i.e., when the indicators are continuous observed variables). Although relations between factors and latent response variables are assumed to be linear, associations between the latent response variables and the indicators (Likert-scale items) are nonlinear (e.g., Figure 4.4). Parameters estimated in robust WLS estimation are derived by the computer such that the correspondence between the observed polychoric correlations and those predicted by the model is as close as possible, given both the model and the parameter estimates.

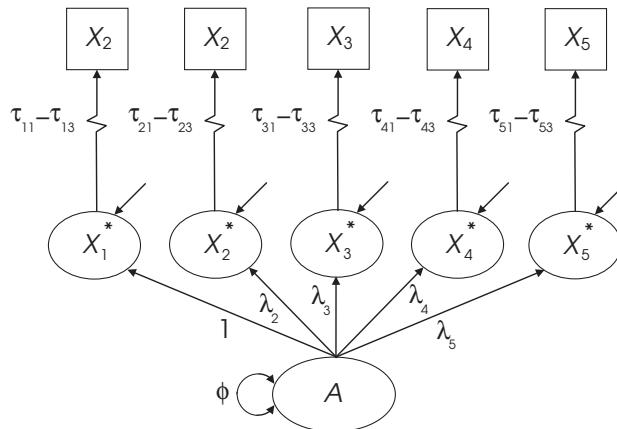
Values of standard errors or model test statistics based on just the diagonal of the asymptotic covariance matrix would be biased over random samples. This is why robust WLS estimation uses information from the full asymptotic covariance matrix (but not its inverse) to calculate robust standard errors and corrected model test statistics, such as the Satorra–Bentler scaled chi-square. These adjusted results in robust WLS estimation for ordinal indicators are analogous to those derived in MLR estimation for continuous-but-non-normal indicators. See Finney and DiStefano (2013) and Edwards, Wirth, Houts, and Xi (2012) for more information about robust WLS estimation and other methods for analyzing ordinal data in SEM.

Presented in Figure 13.6 is the diagram for a single-factor CFA model that represents application of the WLS estimator to a set of five items  $X_1$ – $X_5$ . All items have the same four-point Likert scale. Each path in the figure from a latent response variable  $X^*$  to its corresponding item has a “zigzag” that represents the set of thresholds (three in total) for each item (Edwards et al., 2012). This **threshold structure** associates the  $X^*$  and  $X$  variables. The common factor is designated as  $A$ , and its indicators are the latent response variables, which have error terms. The items in the figure all have the same Likert scale, but it is no special problem to analyze items with different scales (e.g., some items are true–false, others have three or more response options). The total number of thresholds for each item is just one less than the number of its response categories.

Some SEM computer tools, such as Mplus and the lavaan package for R, offer a choice of two different ways to scale latent response variables. Thresholds are free parameters in both methods. In **delta scaling (parameterization)**, the *total* variance of the latent response variables is fixed to 1.0. This metric is consistent with that of the polychoric correlations, which assume variances of 1.0 for the latent response variables. In the standardized solution of delta scaling (i.e., common factor variances are also fixed to 1.0), (1) pattern coefficients estimate the amount of standard deviation change in a

---

<sup>5</sup>The Mplus program analyzes the thresholds, polychoric correlations, and asymptotic covariance matrix, and the LISREL program analyzes just the polychoric correlations and asymptotic covariance matrix.



**FIGURE 13.6.** Single-factor model for five Likert-scale items ( $X$ ), each with four response categories shown with latent response variables ( $X^*$ ), compact symbolism for error terms, and free parameters in delta scaling.  $\tau$ , thresholds;  $\lambda$ , pattern coefficients;  $\phi$  factor variance. Error variances are not free parameters in delta scaling.

latent response variable, given a change of one standard deviation in the common factor; and (2) thresholds are normal deviates that correspond to cumulative areas of the curve to the left of a particular category (Finney & DiStefano, 2013). These interpretations are familiar and straightforward.

Specification of **theta scaling (parameterization)** instead of delta scaling does not change model fit in a single sample analysis. The main difference is that in theta scaling the *residual* variance of each latent response variable is fixed to 1.0. This metric for the error variance is consistent with the scaling in probit regression. In the unstandardized solution, (1) pattern coefficients estimate the amount of change in probit (normal deviate) units in the latent response variable for every 1-point change in the factor; and (2) thresholds are predicted normal deviates for the next lowest response category where the latent response variable is *not* standardized (its variance is *not* 1.0). Interpretation of unstandardized estimates in theta scaling can be difficult, given that the total variance of each latent response variable is not 1.0 (Finney & DiStefano, 2013). Fortunately, the completely standardized solution in theta scaling is identical to the corresponding solution in delta scaling, which is easier to interpret.

Two kinds of robust WLS estimators are available in Mplus: **mean-adjusted least squares** (WLSM) and **mean- and variance-adjusted weighted least squares** (WLSMV). Parameter estimates and robust standard errors from these two methods are the same, but WLSMV estimation can make somewhat different adjustments to the model chi-square or degrees of freedom compared with WLSM estimation. For example, the WLSMV method estimates the degrees of freedom in order to better approximate a chi-square distribution, and they can vary over samples for the same model (Lei & Wu, 2012). Thus, values of the same fit statistics can differ over the two methods due to different chi-square or degrees of freedom values used in the calculations. Results of computer simulation studies generally favor the WLSMV estimator over the WLSM

estimator (Finney & DiStefano, 2013), and WLSMV may be preferred when the number of observed variables is relatively small.

The robust WLS method in LISREL is referred to as **robust diagonally weighted least squares** (RDWLS). In LISREL 8 (Scientific Software International, 2006), the researcher had to input the raw data to PRELIS in order to estimate item thresholds, the polychoric correlation matrix, and the asymptotic covariance matrix. The latter two matrices were then input to LISREL for analysis of the measurement model with ordinal indicators. Both steps can now be performed automatically in LISREL 9 (Scientific Software International, 2013). Namely, use of PRELIS to prepare the data is no longer required, although doing so remains an option.

### **Example Analysis with the Robust WLS Estimator**

Described next is the analysis of the single-factor model in Figure 13.6 with the WLSMV estimator and delta scaling in Mplus (Müthen & Müthen, 1998–2014). The data for this example are the responses of 2,004 white men to five items (nos. 1, 2, 7, 11, and 20) from the Center for Epidemiologic Studies Depression (CES-D) scale (Radloff, 1977). These items concern somatic complaints or reports of reduced activity related to depression. Their 4-point Likert scale indicates the degree to which each symptom was experienced during the prior week:

$$0 = < 1 \text{ day}, 1 = 1\text{--}2 \text{ days}, 2 = 3\text{--}4 \text{ days}, 3 = 5\text{--}7 \text{ days}$$

These data were collected as part of the National Health and Nutrition Examination Survey (NHANES) 1982–1984 Epidemiological Follow-up (Cornoni-Huntley et al., 1983; Madans et al., 1986). The raw data are available from the website for the Inter-university Consortium for Political and Social Research (ICPSR).<sup>6</sup> Through permission of the ICPSR, these raw data are also available on the website for this book.

The number of observations in this analysis with  $v = 5$  indicators include  $5(4)/2$ , or 10 polychoric correlations plus 15 thresholds (3 per item) for a total of 25. The total number of free parameters is 20, including 15 thresholds, 4 pattern coefficients for the latent response variables, and the variance of the common factor (see Figure 13.6), so  $df_M = 25 - 20 = 5$ . Because the numbers of sample versus estimated thresholds are the same (15), each observed threshold will equal its predicted counterpart; that is, all **threshold residuals** will be zero. Estimation in Mplus converged to an admissible solution. Values of selected fit statistics are reported next where the model chi-square is the scaled Satorra–Bentler statistic:

$$\begin{aligned}\chi^2_{SB}(5) &= 17.904, p = .003 \\ \text{RMSEA} &= .036, 90\% \text{ CI } [.019, .055] \\ \text{CFI} &= .994\end{aligned}$$

---

<sup>6</sup>[www.icpsr.umich.edu](http://www.icpsr.umich.edu)

Note that Mplus does not compute the standardized root mean square residual (SRMR) for this type of analysis. The exact-fit test is failed, so the model is tentatively rejected. Results on other fit statistics do not seem problematic, but note that interpretive guidelines for approximate fit indexes based on the analysis of continuous data may not apply when analyzing ordinal data. In a moment we will inspect the residuals, but I can tell you now that no obvious problems are indicated, so the single-factor model with ordinal indicators is retained.

Parameter estimates for delta scaling are reported in Table 13.9. The unstandardized pattern coefficients estimate the amount of change in each latent response variable, given a 1-point change in their common factor, or A (depression). Each standardized pattern coefficient estimates the Pearson correlation between the depression factor and each latent response variable. These same coefficients also estimate the amount of the change in standard deviation units in the latent response variables, given a change of one full standard deviation in the depression factor. The squares of these coefficients indicate the proportions of explained variance ( $R^2$ ), but these values concern the latent response variables, *not* the original items (observed variables).

Reported in the top part of Table 13.10 are the correlation residuals, or differences between the estimated (sample) polychoric correlations and values predicted by the model. None of the correlation residuals exceed .10 in absolute value. The standardized residuals are reported in the bottom part of Table 13.10, and only one of these values (shown in boldface) is statistically significant at the .05 level, but the corresponding correlation residual (.024) is not large. Given all these results, I would retain the model because it closely reproduces the sample correlations. The two failed significance tests

**TABLE 13.9. Robust Weighted Least Squares Estimates for a Single-Factor Model with Ordinal Indicators**

Parameter	Unstandardized		Standardized		
	Estimate	SE	Estimate	SE	$R^2$
Pattern coefficients					
$A \rightarrow X_1^*$	1.000	—	.609	.028	.370
$A \rightarrow X_2^*$	1.070	.065	.651	.029	.424
$A \rightarrow X_3^*$	1.285	.065	.782	.020	.612
$A \rightarrow X_4^*$	1.004	.056	.611	.023	.373
$A \rightarrow X_5^*$	1.266	.065	.771	.021	.594
Factor variance					
A	.370	.034	1.000	—	—

*Note.* Factor A refers to a depression factor. Thresholds:  $X_1$ , .772, 1.420, 1.874;  $X_2$ , 1.044, 1.543, 1.874;  $X_3$ , .541, 1.152, 1.503;  $X_4$ , .288, 1.000, 1.500;  $X_5$ , .558, 1.252, 1.712. All results were computed with Mplus in delta parameterization and STDYX standardization.

**TABLE 13.10. Correlation Residuals and Standardized Residuals for a Single-Factor Model with Ordinal Indicators**

Indicator	$X_1^*$	$X_2^*$	$X_3^*$	$X_4^*$	$X_5^*$
<u>Correlation residuals</u>					
$X_1^*$	—				
$X_2^*$	.041	—			
$X_3^*$	-.005	-.029	—		
$X_4^*$	.030	.020	-.024	—	
$X_5^*$	-.046	-.013	.024	-.005	—
<u>Standardized residuals</u>					
$X_1^*$	—				
$X_2^*$	1.331	—			
$X_3^*$	-.213	-1.193	—		
$X_4^*$	1.110	.679	-1.230	—	
$X_5^*$	-1.935	-.511	<b>2.370</b>	-.282	—

Note. The correlation residuals were computed by Mplus, and the standardized residuals were computed by LISREL.

(model chi-square, standardized residual) are probably more due to the relatively large sample size ( $N = 2,004$ ) than to gross specification error. In the LISREL analysis for the same model and data, I used the PRELIS program to generate the matrix of polychoric correlations and the asymptotic covariance matrix that were subsequently analyzed in LISREL. You can download from this book's website all Mplus, LISREL, and PRELIS syntax, data, and output files for this example.

## Other Estimation Methods

Described next are some additional options for analyzing CFA models with ordinal indicators. The EQS program uses a two-stage method by Lee, Poon, and Bentler (1995) for analyzing models with any combination of continuous or categorical endogenous variables. In the first stage, a special form of ML estimation is used to estimate correlations between the latent response variables. In the second stage, an asymptotic covariance matrix is computed, and the model is analyzed with a method that in EQS is referred to as **arbitrary generalized least squares** (AGLS), which is full WLS estimation.

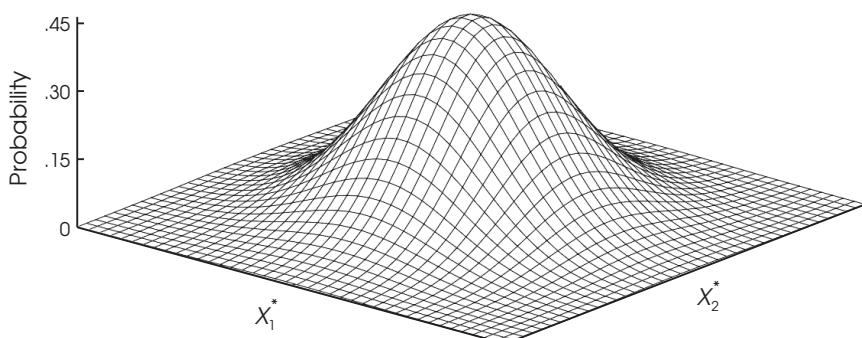
The Amos program takes a Bayesian approach to the analysis of ordinal data. It generates posterior distributions of parameter estimates and provides the user with different kinds of graphical displays about precision, but knowledge of Bayesian methods is required. Forero, Maydeu-Olivares, and Gallardo-Pujol (2009) described an unweighted least squares (ULS) estimator for ordinal data. It performed well against robust WLS

in computer simulations, but the ULS method is not scale invariant and it requires the same response format for all indicators.

A full-information version of ML (FIML) estimation for noncontinuous indicators is becoming increasingly available in SEM computer tools, including LISREL, Mplus, Stata, and others. It does not fit the model to a bivariate correlation structure as in the limited-information WLS method. Instead, it directly analyzes the raw data and estimates the latent response variables using methods for numerical integration. This means that the computer attempts to estimate probabilities of the data within the probability integral for the multivariate normal distribution of the latent response variables. For example, presented in Figure 13.7 is a bivariate normal distribution for two variables. The whole distribution is defined by a double probability integral,<sup>7</sup> but it is difficult to get computers to solve probability integrals for even a single dimension, much less over multiple dimensions.

Instead, computer algorithms for FIML estimation rely on approximate methods that basically select random samples from the joint distribution of the latent response variables. One method is the Markov Chain Monte Carlo (MCMC) method, and another is adaptive quadrature, among other methods of sampling from multidimensional probability integrals using simpler geometric structures such as rectangles or chains. Computational requirements greatly increase as the number of dimensions increases. Another drawback is a reduction in information about fit that is available in the output. For example, Mplus prints the values of a small number of fit statistics, such as the AIC and BIC, for models with more than a relatively small number of ordinal indicators, and no residuals are available, depending on analysis options. A potential advantage of the FIML estimator is presumably better precision compared with limited-information methods, but very large samples are needed—see Edwards et al. (2012).

**Parceling** is an older method for analyzing items in CFA, but it is less relevant nowadays, given the increasing availability of methods for ordinal indicators. Briefly, a



**FIGURE 13.7.** A bivariate normal distribution for two latent response variables.

<sup>7</sup>For example, see <http://mathworld.wolfram.com/BivariateNormalDistribution.html>

**parcel** is an average or total score across a set of homogeneous items, each with a Likert scale. Little (2013) recommends using the average instead of the total score because the average will be in the same metric as the original items. If the number of items going into each parcel differs, then parcels based on total scores will have different metrics, but their averages will have similar metrics. Parcels are generally treated as continuous variables, and score reliabilities of parcels (average or total scores) tend to be greater than that for individual items. If the distributions of all parcels are normal, then default ML estimation could be used to analyze the data. Parcels are then typically specified as continuous indicators of substantive latent variables in a measurement model, such as in a CFA model or when analyzing an SR model.

There are two potential drawbacks to parceling. First, there are many different ways to parcel items, including random assignment and grouping items based on rational grounds, among others, and the choice can affect the results. Second, parceling is *not* recommended if unidimensionality cannot be assumed. Specifically, parceling should not be part of an analysis aimed at determining whether a set of items is unidimensional. This is because it is possible that parceling can mask a multidimensional factor structure in such a way that a seriously misspecified model may nevertheless fit the data. Yang, Nay, and Hoyle (2010) describe situations where parceling as a form of data aggregation may be helpful when analyzing questionnaires with a large number of items. Parceling may also be advantageous in small-sample analyses where many individual items are replaced with a smaller number of parcels.

The technique of EFA is also an alternative to CFA for analyzing items. The advantage is that EFA analyzes unrestricted measurement models, where items are allowed to depend on all factors. Items in EFA often have relatively high pattern coefficients over multiple factors, and forcing such items to depend on a single factor in a restricted measurement model, as in CFA, may be too demanding (i.e., the CFA model may be rejected). Estimation methods for ordinal data, such as robust WLS and FIML, are increasingly available in computer procedures for EFA. This includes SEM computer tools, such as LISREL and Mplus, with capabilities for EFA. The technique of EFA is more “forgiving” than CFA in item-level analyses, and sometimes this is exactly what is needed when theory is weak.

## ITEM RESPONSE THEORY AS AN ALTERNATIVE TO CFA

Wirth and Edwards (2007) show that parameter estimates in item-level CFA analyses, such as pattern coefficients and thresholds, can be rescaled as item discriminations and item difficulties in two-parameter item response theory (IRT) models (e.g., Figure 4.4), and vice versa. Some SEM computer tools can optionally print IRT-type estimates, too. These sets of estimates (IRT, CFA) are simple transformations of each other, but CFA has relatively few advantages over IRT-based analyses of item-level data. One is that it is easy to specify correlated errors in CFA, but IRT models generally assume local independence. Another is that indicators can be specified as complex or multidimensional

in CFA, but doing so in IRT is more difficult. Measurement models analyzed in SEM can include predictors (covariates). Such models are SR models, not CFA models, but it is easy to include covariates in SEM.

The IRT approach offers more flexibility than CFA in a few key areas. One is the capability to develop **tailored tests**, or subsets of items that may optimally assess a particular person based on the correctness of his or her previous responses. If the examinee fails initial items, then the computer presents easier ones. Testing stops when more difficult items are consistently failed. A reliability coefficient can be estimated for each person, given the particular items administered. In contrast, CFA analyzes static measurement models fitted to data from whole samples, not individual cases. There are specific methods in IRT for equating different tests for difficulty and for shortening an extant questionnaire in a way that ensures optimal precision of the new scores compared with the old ones.

## SUMMARY

Many types of hypotheses about reflective measurement can be tested in CFA. The analysis of a model with multiple factors that specifies unidimensional measurement allows precise tests of convergent validity and discriminant validity. Respecification of a measurement model can be challenging because many possible changes could be made to a given model. Another problem is that of equivalent measurement models. The only way to deal with both of these challenges is to rely more on substantive knowledge than on statistical considerations in model evaluation. There are special estimation methods for analyzing ordinal data. Limited-information estimators, such as robust weighted least squares, fit the model to a correlation structure for the latent response variables, and full-information methods based on maximum likelihood directly analyze the raw data. These methods for items with relatively few response categories or severely asymmetrical distributions are more appropriate than default ML estimation for continuous data.

## LEARN MORE

Books by Brown (2015) and Harrington (2009) are excellent resources for CFA. Edwards, Wirth, Houts, and Xi (2012) describe the analysis of ordinal data in SEM.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.

Edwards, M. C., Wirth, R. J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 195–208). New York: Guilford Press.

Harrington, D. (2009). *Confirmatory factor analysis*. New York: Oxford University Press.

**EXERCISES**

1. Show that  $df_M = 20$  for the single-factor model of the KABC-I (see Figure 9.7).
2. Use an SEM computer tool to fit the single-factor model of the KABC-I to the data in Table 13.1. Evaluate the residuals for this model.
3. Reproduce the values of the structure coefficients in Table 13.4 using the tracing rule with the parameter estimates in Table 13.3.
4. Calculate the composite reliability (CR) for the simultaneous processing factor, given the results in Table 13.3.
5. Interpret  $\hat{\tau}_2 = .25$  in Figure 13.5(b), given the information in Figure 13.5(a).

## **Appendix 13.A**

### Start Value Suggestions for Measurement Models

These recommendations concern reflective measurement models, whether these models are CFA models or part of an SR model. Unstandardized variables, including the factors, are assumed. In the reference (marker) variable method of scaling factors, initial estimates of factor variances should probably not exceed 90% of that of the observed (sample) variance of the corresponding reference variable. In the effects coding method, all indicators of the same factor have the same metric, so the start value for the factor variance should be less than 90% of the average observed variance over all the indicators. Start values for factor covariances follow the initial estimates of their variances; that is, they are the product of each factor's standard deviation (the square root of the initial estimates of their variances) and the expected correlation between them.

If indicators of the same factor have similar variances to that of the reference variable, then initial estimates of their pattern coefficients can also be 1.0. If the reference variable is, say, one-tenth as variable as another indicator, the initial estimate of the other indicator's pattern coefficient could be 10.0. Conservative start values for indicator error variances could be 90% of the observed variance of the associated, which assumes that only 10% of the variance will be explained. Bentler (2006) suggests that it is probably better to overestimate the variances of factors and error variances than to underestimate them.

## Appendix 13.B

### Constraint Interaction in CFA Models

Suppose that a researcher specifies a standard two-factor CFA model where the indicators of factors  $A$  are  $X_1$  and  $X_2$  and the indicators of factor  $B$  are  $X_3$  and  $X_4$ . The sample covariance matrix where the order of the variables is  $X_1-X_4$  and  $N = 200$  is as shown here:

$$\begin{bmatrix} 25.00 & & & \\ 7.20 & 9.00 & & \\ 3.20 & 2.00 & 4.00 & \\ 2.00 & 1.25 & 1.20 & 4.00 \end{bmatrix} \quad (\text{I})$$

It is believed that the unstandardized pattern coefficients of  $X_2$  and  $X_4$  are equal. To test this hypothesis, an equality constraint is imposed on the unstandardized estimates, or

$$A \rightarrow X_2 = B \rightarrow X_4$$

Next, the model so restricted is compared to the one without this constraint. Ideally, the value of  $\chi^2_D(1)$  for this comparison should not depend on how the factors are scaled, but this ideal is not realized for this example. If  $X_1$  and  $X_3$  are the reference variables for their respective unstandardized factors, then  $\chi^2_D(1) = 0$ . But if instead the factor variances are fixed to 1.0 (the factors are standardized), then  $\chi^2_D(1) = 14.087$  calculated in the student version of LISREL for the same comparison. Try it!

This unexpected result is an example of constraint interaction, which means that the value of the chi-square difference statistic for the test of the equality constraint depends on how the factors are scaled. It happens in this example because the imposition of the cross-factor equality constraint has the unintended consequence of making unnecessary one of the two identification constraints that scale the factors. But removing the unnecessary identification constraint from the model with the equality constraint would result in two nonhierarchical models with equal degrees of freedom. In other words, we could not conduct the chi-square difference test.

Steiger (2002) describes this test for constraint interaction: Obtain  $\chi^2_M$  for the model with the cross-factor equality constraint. If the factors are unstandardized, fix the pattern coefficient

of the reference variable to a different constant, such as 2.0. If the factors are standardized, fix the variance of one of these factors to a constant other than 1.0. Fit the model so respecified to the same data. If the value of  $\chi^2_M$  for the respecified model is not identical to that of the original, constraint interaction is present. If so, the choice of how to scale the factor should be based on substantive grounds. If no such grounds exist, the test results for the equality constraint may not be meaningful. Gonzalez and Griffin (2001) describe how the estimation of standard errors in SEM is not always invariant to how the factors are scaled.

## 14

# Analysis of Structural Regression Models

---

Described next are two different strategies for analyzing fully latent SR models where every substantive variable in the structural model is a factor measured by multiple indicators. These strategies address the problem of how to locate the source(s) of specification error by separating the evaluation of the measurement part of the model from the analysis of the structural part. The detailed example for this chapter follows the two-step approach just mentioned. Also discussed in this chapter is (1) the analysis of partially latent SR models while controlling for measurement error in single indicators and (2) the estimation of formative measurement models in SEM. The advanced techniques covered in the next part of this book extend the rationale of SR models to other kinds of analyses.

---

### TWO-STEP MODELING

Suppose that a researcher specified the fully latent, three-factor SR model in Figure 10.4(a). The data are collected and the researcher uses **one-step modeling** to estimate this model, which means that its measurement and structural components are analyzed simultaneously in a single analysis. The results indicate poor fit. Now, where is the model misspecified? the measurement part? the structural part? or both? With one-step modeling, it is hard to precisely locate the source of the problem. **Two-step modeling** by Anderson and Gerbing (1988) parallels the two-step heuristic (Rule 10.1) for the identification of SR models:

1. In the first step, a fully latent SR model is respecified as a CFA measurement model, which is then analyzed in order to determine whether it fits the data. If the fit of this CFA model is poor, then not only may the researcher's hypotheses about measurement be wrong, but also the fit of the original SR model may be even worse if its structural model is overidentified. Suppose that the fit of the three-factor CFA model in Figure 10.4(b) is poor. This model has *three* paths among the factors that represent all

possible covariances. In contrast, the structural part of the original SR model in Figure 10.4(a) has only *two* paths that represent direct effects. If the fit of the CFA model with three paths among the factors is poor, then the fit of the SR model with only two paths may be even worse. The first step thus involves finding an adequate measurement model (follow the suggestions in the previous chapter).

2. Given an acceptable measurement model, the second step is to compare the fits of the original SR model (with modifications to its measurement part, if any) and those with different structural models to one another and to the fit of the CFA model with the chi-square difference test. (This assumes that hierarchical models are compared.) Here is the procedure: If the structural part of an SR model is just-identified, the fits of the SR model and the CFA respecification of it are identical because these two models are equivalent. For example, if the path  $A \rightarrow C$  were added to the SR model of Figure 10.4(a), then it would have just as many free parameters as does the CFA model of Figure 10.4(b). The original SR model of Figure 10.4(a) with its overidentified structural part is thus nested under the CFA model of Figure 10.4(b). But it may be possible to trim a just-identified structural part of an SR model without appreciable deterioration in fit. Structural portions of SR models are respecified according to the same principles as in path analysis.

Given a retained CFA model, one should observe only slight changes in the pattern coefficients as SR models with alternative structural components are tested. If so, then the assumptions about measurement may be invariant to changes in the structural part of an SR model. But if values of the pattern coefficients change markedly when the different structural models are specified, the measurement model is clearly not invariant. This phenomenon may lead to **interpretational confounding** (Burt, 1976); that is, the empirical definitions of the constructs depend on hypotheses about causal effects between them. It is generally easier to detect interpretational confounding in two-step modeling than in one-step modeling.

## FOUR-STEP MODELING

Four-step modeling (Hayduk & Glaser, 2000; Mulaik & Millsap, 2000) for fully latent SR models extends the two-step method to even more precisely diagnose misspecification in the measurement model. The researcher tests a sequence of at least four hierarchical models. In order for these models to be identified, each factor should have at least four indicators. As in two-step modeling, if the fit of a model in four-step modeling with fewer constraints is poor, then a model with even more constraints should not even be considered. The steps are outlined next:

1. The least restrictive model specified at the first step is an EFA model where each indicator depends on all factors and the number of factors is the same as that in the target SR model. This EFA model should be analyzed with the same estimation method,

such as robust WLS when the data are ordinal, as used to analyze the final SR model (at the fourth step). This first step is intended to test the provisional correctness of the hypothesis regarding the number of factors, but it cannot *confirm* that hypothesis if model fit is adequate (Hayduk & Glaser, 2000).

**2.** The second step of four-step modeling corresponds to the first step of two-step modeling: A CFA model is specified where some pattern coefficients are fixed to zero. If the fit of the CFA model at this step is reasonable, one goes on to test the original SR model; otherwise, the measurement model should be revised.

**3.** The third step involves testing the SR model with the same set of zero pattern coefficients as represented in the measurement model from the second step but where at least one factor covariance from the second step is respecified as a direct causal effect.

**4.** The last step involves tests of *a priori* hypotheses about parameters free from the onset of model testing. These tests typically involve imposing zero constraints or dropping a path from the structural model. The third and fourth steps of four-step modeling are basically a more specific statement of activities that would fall under the second step of two-step modeling.

Which approach to analyzing fully latent SR models is best, two-step or four-step modeling? Both methods have their critics and defenders, and both capitalize on chance variation when hierarchical models are tested and respecified using the same data. The two-step method is simpler, and it does not require four or more indicators per factor. Both two-step and four-step modeling are better than one-step modeling, where there is no separation of measurement issues from structural issues. Neither is a “gold standard” for testing SR models, but there is no such thing (Bentler, 2000). Bollen (2000) describes additional methods for testing SR models.

## **INTERPRETATION OF PARAMETER ESTIMATES AND PROBLEMS**

Interpretation of estimates from the analysis of an SR model should not be difficult if one knows something about path analysis and CFA. For example, path coefficients are interpreted for SR models as regression coefficients between factors. Total effects between factors can be decomposed into direct and total indirect effects, just as in path analysis. Pattern coefficients are interpreted as regression coefficients for direct effects of factors on indicators, just as in CFA.

Many SEM computer tools print estimated  $R^2$  values for each endogenous variable. This includes, for SR models, the indicators in the measurement model and the factors specified as outcomes in the structural model. Values of  $R^2$  are usually computed for indicators in the unstandardized solution as one minus the ratio of the estimated error variance over the sample variance of that indicator. Although variances of endogenous

factors are not model parameters, they nevertheless have predicted variances; therefore, values of  $R^2$  are usually calculated for endogenous factors as one minus the ratio of the estimated disturbance variance over the predicted variance for that factor. Watch for Heywood cases that suggest a problem with the data, specification, sample size, number of indicators per factor, or identification status of the model. If iterative estimation fails due to poor start values set automatically by the computer, the guidelines in Appendix 11.A can be followed for generating your own start values for the structural model or the guidelines in Appendix 13.A for the measurement model.

Most SEM computer tools calculate a standardized solution for SR models by first finding the unstandardized solution with unit loading identification (ULI) constraints for endogenous factors and then transforming it to standardized form. Steiger (2002) notes that this method assumes that ULI constraints function only to scale the endogenous variables. In other words, there is no constraint interaction. See Appendix 14.A for more information about constraint interaction in SR models.

Depending on the SEM computer tool, more than one standardized solution may be printed in the output for the same SR model and data. For example, all variables are standardized in LISREL's *completely standardized solution* and in the STDYX solution of Mplus, but just the factors are standardized in LISREL's *standardized solution* and in Mplus's STD solution. For fully latent SR models with no measured exogenous variables in the structural model, the Mplus STDYX and STDY solutions will be identical. But in partially latent SR models with measured exogenous variables in the structural model (e.g., Figure 10.2(a)), the two solutions may differ. This is because the STDY solution in Mplus does *not* use the variances of measured exogenous variables (covariates) in its standardized solution.

## DETAILED EXAMPLE

This example of the two-step analysis of a fully latent SR model of job satisfaction was introduced in Chapter 10. Briefly, Houghton and Jinkerson (2007) measured within a sample of 263 full-time university employees three indicators each of constructive thinking, dysfunctional thinking, subjective well-being, and job satisfaction. They hypothesized that constructive thinking reduces dysfunctional thinking, which leads to an enhanced sense of well-being, which in turn results in greater job satisfaction. They also predicted that dysfunctional thinking directly affects job satisfaction. Their SR model is presented in Figure 10.6. We will first analyze whether its measurement part is consistent with the data summarized in Table 14.1. All results described next are from converged, admissible solutions.

I submitted the correlations and standard deviations in Table 14.1 to Stata (Stata-Corp, 1985–2015) for analysis with the `sem` command. Variances are calculated with  $N - 1$  in the denominator, not  $N$ . The first model analyzed with ML estimation was a standard one-factor CFA model with 12 indicators. Values of selected fit statistics for this

**TABLE 14.1. Input Data (Correlations, Standard Deviations) for Analysis of a Structural Regression Model of Thought Strategies and Job Satisfaction**

Variable	1	2	3	4	5	6	7	8	9	10	11	12
<u>Job satisfaction</u>												
1. Work <sub>1</sub>	1.00											
2. Work <sub>2</sub>	.668	1.00										
3. Work <sub>3</sub>	.635	.599	1.00									
<u>Subjective well-being</u>												
4. Happy	.263	.261	.164	1.00								
5. Mood <sub>1</sub>	.290	.315	.247	.486	1.00							
6. Mood <sub>2</sub>	.207	.245	.231	.251	.449	1.00						
<u>Dysfunctional thinking</u>												
7. Perform <sub>1</sub>	-.206	-.182	-.195	-.309	-.266	-.142	1.00					
8. Perform <sub>2</sub>	-.280	-.241	-.238	-.344	-.305	-.230	.753	1.00				
9. Approval	-.258	-.244	-.185	-.255	-.255	-.215	.554	.587	1.00			
<u>Constructive thinking</u>												
10. Beliefs	.080	.096	.094	-.017	.151	.141	-.074	-.111	.016	1.00		
11. Self-Talk	.061	.028	-.035	-.058	-.051	-.003	-.040	-.040	-.018	.284	1.00	
12. Imagery	.113	.174	.059	.063	.138	.044	-.119	-.073	-.084	.563	.379	1.00
M	3.96	4.12	4.13	3.97	3.61	3.30	2.13	1.63	1.99	3.86	3.62	3.50
SD	.939	1.017	.937	.562	.760	.524	.585	.609	.731	.711	1.124	1.001

Note. Input data are from Houghton and Jinkerson (2007); N = 263.

**TABLE 14.2. Values of Selected Fit Statistics for Two-Step Testing of a Structural Regression Model of Thought Strategies and Job Satisfaction**

Model	$\chi^2_{\text{M}}$	$df_{\text{M}}$	$p$	$\chi^2_{\text{D}}$	$df_{\text{D}}$	$p$	RMSEA (90% CI)	CFI	SRMR
<u>Measurement model</u>									
One factor	566.797	54	<.001	—	—	—	.190 (.176-.204)	.498	.133
Four factor	62.468	48	.078	504.329	6	<.001	.034 (0-.056)	.986	.037
Four factor, $E_{\text{Ha}} \rightsquigarrow E_{\text{Mo}_2}$	56.662	47	.158	5.806	1	.016	.028 (0-.052)	.991	.035
<u>Structural regression model</u>									
Six paths	56.662	47	.158	—	—	—	.028 (0-.052)	.991	.035
Four paths	60.010	49	.135	3.348	2	.188	.029 (0-.052)	.989	.040

Note. CI, confidence interval. All results were computed by Stata.

initial measurement model are reported in Table 14.2. The fit of the one-factor CFA model is clearly poor. For example, the exact-fit test is failed,  $\chi^2_M(54) = 566.797, p < .001$ , and the upper bound of the 90% confidence interval based on the RMSEA is .204 (see the table).

Next, I specified the measurement part of Figure 10.6 as a standard four-factor CFA model. Values of selected fit statistics for this four-factor CFA model are listed in Table 14.2. The exact-fit hypothesis is not rejected,  $\chi^2_M(48) = 62.468, p = .078$ . The relative improvement in fit of the four-factor CFA model over that of the one-factor CFA model is statistically significant,  $\chi^2_D(6) = 504.329, p < .001$ , and values of approximate-fit indexes for the four-factor model are generally favorable (e.g., upper-bound RMSEA = .056; CFI = .986; see the table).

Inspection of the residuals for the four-factor CFA model indicated few apparent problems. For example, two absolute correlation residuals (computed in EQS) just exceed .10, which is not a bad result in a larger model. A total of three standardized residuals are significant at the .05 level. One of these results was for the “Happy” (percent time happy) and “Mood<sub>2</sub>” indicators of the subjective well-being factor. One of the largest modification indexes (5.380) was for an error covariance between the same pair of variables.

Because it seems reasonable that common item content across the two indicators just mentioned could explain shared error variance, I respecified the four-factor CFA model by allowing the error covariance between the “Happy” and “Mood<sub>2</sub>” indicators to be freely estimated in a third analysis. The results are summarized in Table 14.2. Its fit to the data is statistically better than that of the four-factor CFA model with no correlated errors,  $\chi^2_D(1) = 5.806, p = .016$ . The exact-fit hypothesis is not rejected for the respecified measurement model,  $\chi^2_M(47) = 56.662, p = .158$ . Values of other fit statistics are generally favorable (RMSEA = .028; CFI = .991). Finally, no absolute correlation residuals exceeded .10.

Based on the results just described, the four-factor CFA model in Figure 14.1 was retained but with an error correlation. In contrast, Houghton and Jinkerson’s (2007) final measurement model was a standard four-factor model, so my conclusion differs somewhat from theirs. Reported in Table 14.3 are estimates of pattern coefficients and error variances for the CFA model in Figure 14.1 but with an error correlation. Values of standardized pattern coefficients for indicators of some factors, such as job satisfaction (range = .749–.839), are uniformly high. A few other standardized coefficients are rather low, such as .433 for the self-talk indicator of constructive thinking, so evidence for convergent validity is mixed. Values of  $R^2$  for indicators range from .188 to .817 (see the table).

Estimates of factor variances and covariances and of the error covariance for the final CFA measurement model are listed in Table 14.4. Exercise 1 asks you to verify that a total of two factor covariances are not statistically significant at the .05 level. Of greater interest are the estimated factor correlations, which range from −.480 to .466. These moderate intercorrelations suggest discriminant validity. The error covariance is −.043, and the corresponding error correlation is −.243. This correlation does not seem

**TABLE 14.3. Maximum Likelihood Estimates of Pattern Coefficients and Residuals for a Measurement Model of Thought Strategies and Job Satisfaction**

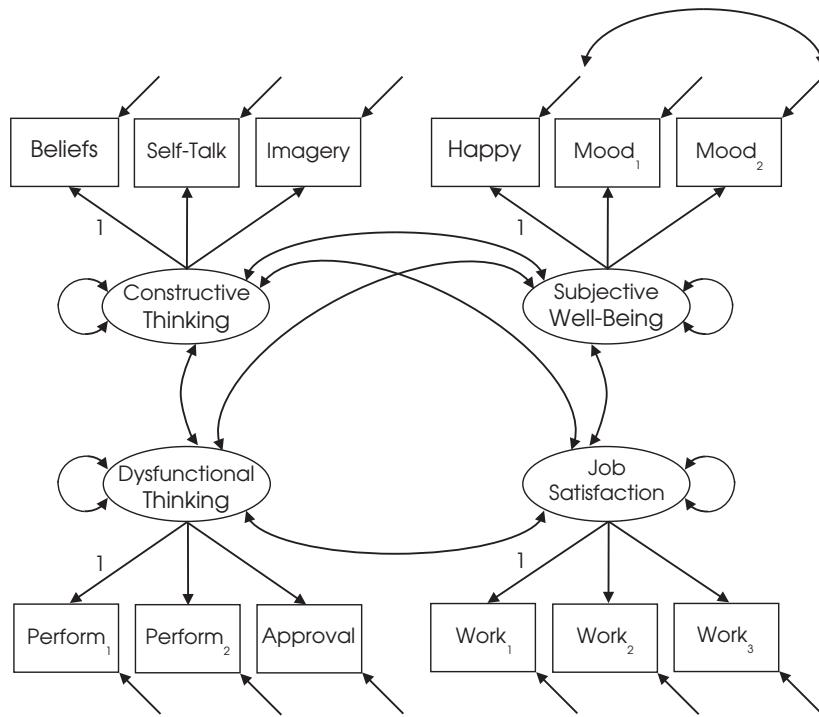
Indicator	Pattern coefficients				Error variances			
	Unstandardized		Standardized		Unstandardized		Standardized	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
<u>Job satisfaction</u>								
Work <sub>1</sub>	1.000	—	.839	.030	.261	.042	.297	.051
Work <sub>2</sub>	1.035	.082	.802	.032	.369	.051	.357	.052
Work <sub>3</sub>	.891	.073	.749	.035	.386	.044	.439	.052
<u>Subjective well-being</u>								
Happy	1.000	—	.671	.061	.174	.026	.550	.081
Mood <sub>1</sub>	1.490	.227	.739	.053	.262	.045	.453	.078
Mood <sub>2</sub>	.821	.126	.591	.062	.179	.022	.651	.073
<u>Dysfunctional thinking</u>								
Perform <sub>1</sub>	1.000	—	.830	.029	.106	.015	.311	.048
Perform <sub>2</sub>	1.133	.079	.904	.026	.068	.017	.183	.047
Approval	.993	.088	.660	.040	.302	.030	.564	.053
<u>Constructive thinking</u>								
Beliefs	1.000	—	.648	.065	.293	.044	.580	.084
Self-Talk	1.056	.179	.433	.062	1.026	.098	.812	.053
Imagery	1.890	.340	.870	.073	.243	.127	.242	.127

Note. Est., estimate. All results were computed by Stata. The standardized solution is completely standardized.

large, but its presence helps to “clean up” local fit problems in the standard four-factor CFA model without this parameter.

The analyses described next concern the second step of two-step modeling—the testing of SR models, with the measurement part established in the first step but with alternative versions of structural models. The first SR model analyzed is one with a just-identified structural component. Because this SR model and the CFA measurement model in Figure 14.1 have the same number of paths among the factors (6), they are equivalent models. This fact is verified by the observation of identical global fit statistics in Table 14.2 for the two models just mentioned. Equivalence also implies that estimates of pattern coefficients and error variances and covariance will be identical within rounding error for the two models. Accordingly, we consider just the parameter estimates for the structural part of the SR model.

Presented in Figure 14.2 are estimates for the just-identified structural model. The direct effects in the figure depicted with dashed lines were predicted by Houghton and Jinkerson (2007) to be zero. Neither the unstandardized coefficient for the direct effect



**FIGURE 14.1.** Measurement component in a structural regression model of thought strategies and job satisfaction with compact symbolism for indicator error terms.

of constructive thinking on dysfunctional thinking ( $-.131$ ) nor the standardized coefficient ( $-.124$ ) is significant at the .05 level. It is no surprise, then, that constructive thinking explains only about 1.5% of the variance in dysfunctional thinking ( $R^2 = .015$ ). The other two sets (unstandardized, standardized) of coefficients for direct effects of constructive thinking on subjective well-being and job satisfaction are also not statistically significant. This is consistent with predictions. Direct effects of dysfunctional thinking on subjective well-being and of subjective well-being on job satisfaction are both significant and appreciable in standardized magnitude (respectively,  $-.470$ ,  $.382$ ). These results support the hypothesis that effects of dysfunctional thinking on job satisfaction are largely indirect through subjective well-being. About 25% of the variance in both subjective well-being and job satisfaction is explained ( $R^2$ 's are, respectively,  $.237$  and  $.245$ ).

The final SR model retained by Houghton and Jinkerson (2007) had the four paths in the structural model represented with solid lines in Figure 14.2. Values of selected fit statistics for this restricted SR model are reported in Table 14.2. The exact-fit hypothesis is not rejected,  $\chi^2_M(49) = 60.010$ ,  $p = .135$ , and its fit is not statistically worse than that of the unrestricted SR model with six direct effects,  $\chi^2_D(2) = 3.348$ ,  $p = .188$ . Inspection of the correlation residuals for the restricted SR model indicates some problems. For example, the residual for the "Work<sub>2</sub>" indicator of job satisfaction and the "Imagery" indicator of constructive thinking is  $.142$ . Other absolute residuals  $> .10$  involved the

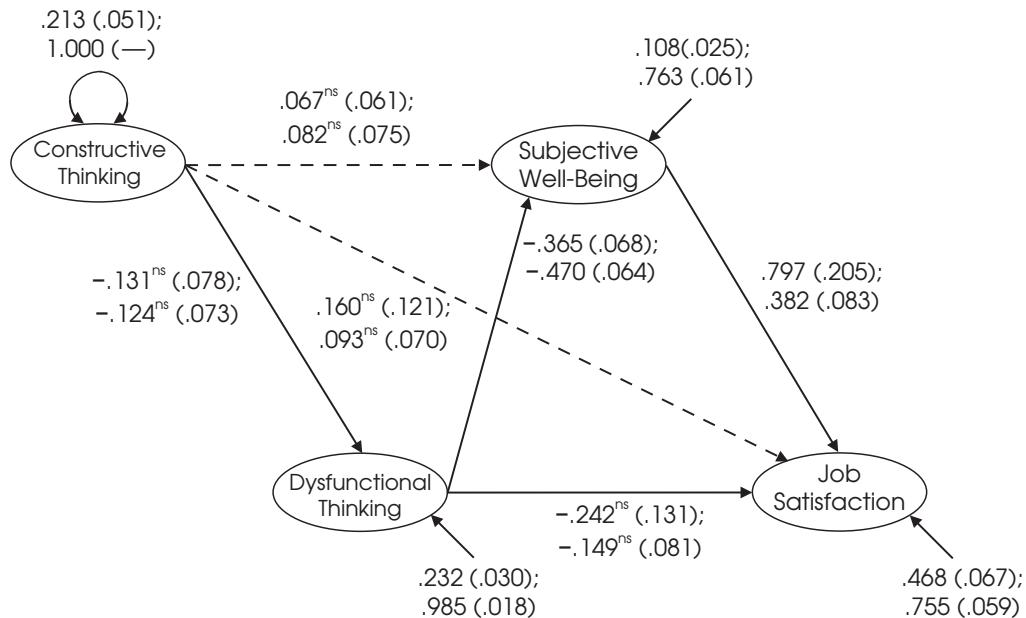
“Beliefs” indicator of constructive thinking and both mood indicators of subjective well-being. This example illustrates how deleting paths that are not significant can appreciably worsen fit elsewhere in the model. Thus, I would retain the SR model with the just-identified structural part (see Figure 14.2). Exercise 2 asks you to calculate a standardized effect decomposition for the just-identified structural model. You can download all computer files for this analysis in Amos, EQS, lavaan, LISREL, Mplus, and Stata from this book’s website.

I used the Power Analysis procedure in STATISTICA Advanced (StatSoft, 2013) to estimate power for the final SR model, given  $N = 263$ ,  $df_M = 47$ , and  $\alpha = .05$ . Assuming  $\varepsilon_1 = .08$  for the test of the close-fit hypothesis ( $\varepsilon_0 \leq .05$ ), power is .869. Now assuming  $\varepsilon_1 = .01$  for the test of the not-close-fit hypothesis ( $\varepsilon_0 \geq .05$ ), power is .767. These results say that the probability of rejecting a false model or detecting a correct model is reasonably good. Although the sample size for this analysis is not large, there are sufficient model degrees of freedom to offset the negative impact of a smaller sample on power.

**TABLE 14.4. Maximum Likelihood Estimates of Factor Variances and Covariances and Error Covariance for a Measurement Model of Thought Strategies and Job Satisfaction**

Parameter	Unstandardized		Standardized	
	Est.	SE	Est.	SE
Factor variances and covariances				
Job Satisfaction	.620	.082	1.000	—
Subjective Well-Being	.142	.031	1.000	—
Dysfunctional Thinking	.236	.031	1.000	—
Constructive Thinking	.213	.051	1.000	—
Constructive $\curvearrowleft$ Dysfunctional	-.028	.018	-.124	.073
Constructive $\curvearrowleft$ Subjective Well-Being	.024	.014	.140	.080
Constructive $\curvearrowleft$ Job Satisfaction	.060	.029	.165	.074
Dysfunctional $\curvearrowleft$ Subjective Well-Being	-.088	.018	-.480	.063
Dysfunctional $\curvearrowleft$ Job Satisfaction	-.132	.030	-.344	.064
Subjective Well-Being $\curvearrowleft$ Job Satisfaction	.139	.027	.466	.066
Error covariance				
Happy $\curvearrowleft$ Mood <sub>2</sub>	-.043	.018	-.243	.116

Note. All results were computed by Stata. The standardized solution is completely standardized.



**FIGURE 14.2.** Structural component in a structural regression model of thought strategies and job satisfaction with compact symbolism for disturbances. Estimates in the top row are unstandardized (standard error); estimates in the bottom row are standardized (standard error). Standardized estimates are from a completely standardized solution. All estimates are statistically significant at the .05 level except for those designated “ns,” which means not significant.

## EQUIVALENT SR MODELS

It is often possible to generate equivalent versions of SR models. An equivalent version of a fully latent SR model with a just-identified structural part was mentioned earlier: the model respecified as a CFA model, which assumes only covariances between factors. Regardless of whether or not the structural part of an SR model is just-identified, it may be possible to generate equivalent versions of it using the replacing rules (Table 12.6), but verify that the respecified and original structural models are d-separation equivalent. With the structural part of an SR model held constant, it may also be possible to generate equivalent versions of the measurement part using the reversed indicator rule (e.g., Figure 13.2). Given no change in the structural part, alternative SR models with equivalent measurement parts will fit the same data equally well; see Hershberger and Marcoulides (2013) for examples.

Equivalent versions of the just-identified structural model in Figure 14.2 for analysis of the Houghton and Jinkerson (2007) data include any other possible just-identified variation of this model. This includes structural models where causal effects “flow” in the opposite direction, such as from job satisfaction to subjective well-being to dysfunctional thinking to constructive thinking. Houghton and Jinkerson (2007) offered

a detailed rationale for their directionality specifications, but without such arguments, there is no way to prefer one just-identified structural model over an equivalent variation.

## SINGLE INDICATORS IN A NONRECURSIVE MODEL

This example was introduced in Chapter 10. Briefly, Chang et al. (2007) administered measures about occupational commitment, organizational commitment, and turnover intention to 177 nurses. Because these authors reported reliability coefficients (Table 10.1), we can use the specification that explicitly controls for measurement error in single indicators—see Figure 10.7. The structural model represents three hypotheses: (1) affective, continuance, and normative organizational commitment affect organizational turnover intention; (2) affective, continuance, and normative occupational commitment affect occupational turnover intention; and (3) organizational and occupational turnover intention mutually cause each other.

I fitted the model in Figure 10.7 to the covariance matrix based on the correlations and standard deviations in Table 14.5 using ML estimation in the lavaan package for R (Rosseel, 2012). You can download the lavaan script and output files for this analysis from this book’s website as well as the computer files for the same analysis in LISREL

**TABLE 14.5. Input Data (Correlations, Standard Deviations, Score Reliabilities) for Analysis of a Nonrecursive Model of Organizational and Occupational Commitment and Turnover Intention**

Variable	1	2	3	4	5	6	7	8
<b>Organizational commitment</b>								
1. Affective	.82							
2. Continuance	-.10	.70						
3. Normative	.66	.10	.74					
<b>Occupational commitment</b>								
4. Affective	.48	.06	.42	.86				
5. Continuance	.08	.58	.15	.22	.71			
6. Normative	.48	.12	.44	.69	.34	.84		
<b>Turnover intention</b>								
7. Organizational	-.53	-.04	-.58	-.34	-.13	-.34	.86	
8. Occupational	-.50	-.02	-.40	-.63	-.28	-.58	.56	.88
M	4.33	4.07	4.02	5.15	4.17	4.44	4.15	3.65
SD	1.04	.98	.97	1.07	.78	1.09	1.40	1.50

*Note.* These data are from Chang et al. (2007); N = 177. Values in the diagonal are internal consistency (Cronbach’s alpha) score reliability coefficients.

(Scientific Software International, 2013). The analysis in lavaan converged normally to an admissible solution. Values of selected fit statistics generated by lavaan are as follows:

$$\begin{aligned}\chi^2_M(4) &= 9.420, p = .051 \\ \text{RMSEA} &= .087, 90\% \text{ CI } [0, .161], p_{\epsilon_0 \leq .05} = .159 \\ \text{CFI} &= .991; \text{ SRMR} = .018\end{aligned}$$

The model just passes the exact-fit test at the .05 level; it also passes the close-fit test at the same level. Values of the CFI and SRMR are generally favorable, but results on the RMSEA are poor: the upper bound of its confidence interval (.161) suggests poor fit within the limits of sampling error (i.e., the poor-fit test is failed). Next we examine the residuals.

No absolute correlation residual exceeded .10. A few standardized residuals were statistically significant, of which the largest in absolute value ( $z = 2.665$ ) concerned the association between the indicator of continuance organizational commitment and the indicator of occupational turnover intention. A respecification consistent with this result is to add to the initial model in Figure 10.7 a direct effect from the continuance organizational commitment factor to the occupational turnover intention factor. The model so respecified was fitted to the same data. Reported next are values of selected fit statistics for the respecified model:

$$\begin{aligned}\chi^2_M(3) &= .809, p = .847 \\ \text{RMSEA} &= 0, 90\% \text{ CI } [0, .070], p_{\epsilon_0 \leq .05} = .913 \\ \text{CFI} &= 1.000; \text{ SRMR} = .005\end{aligned}$$

These results are all favorable. Also, the largest absolute correlation residual is .02, and there are no significant standardized residuals for the revised model. The improvement in fit of the revised model relative to that of the initial model is also significant, or

$$\chi^2_D(1) = 9.420 - .809 = 8.611, p = .003$$

Based on all these results, the respecified nonrecursive model with a direct effect from continuance organizational commitment to occupational turnover intention is retained.

There are no free parameters in the measurement part of the respecified SR model except for the variances of the six exogenous factors and their pairwise covariances. To save space, these estimates are not reported here. Presented in Table 14.6 are the ML estimates of the direct effects (including two reciprocal effects) and the disturbance variances and covariance in the structural part of the respecified model. The best predictor of organizational turnover intention is normative organizational commitment. The standardized path coefficient is  $-.665$ , so a stronger sense of obligation to stay within the organization predicts lower levels of the intention to leave that organization. The best

**TABLE 14.6. Maximum Likelihood Estimates for the Structural Component in a Nonrecursive Model of Organizational and Occupational Commitment and Turnover Intention**

Parameter	Unstandardized	SE	Standardized
<u>Direct effects</u>			
AOC → OrgTI	-.062	.413	-.045
COC → OrgTI	.030	.174	.019
NOC → OrgTI	-1.033	.408	-.665
APC → OccTI	-.754	.231	-.532
CPC → OccTI	-1.441	.647	-.672
NPC → OccTI	.094	.309	.067
COC → OccTI	.996	.454	.579
<u>Reciprocal effects</u>			
OccTI → OrgTI	.037	.147	.040
OrgTI → OccTI	.287	.145	.265
<u>Disturbance variances and covariance</u>			
OrgTI	.767	.192	.458
OccTI	.461	.157	.234
OrgTI ↗ OccTI	.226	.178	.380

Note. AOC, affective organizational commitment; COC, continuance organizational commitment; NOC, normative organizational commitment; APC, affective occupational commitment; CPC, continuance occupational commitment; NPC, normative occupational commitment; OrgTI, organizational turnover intention; OccTI, occupational turnover intention. Standardized estimates for disturbance variances are proportions of unexplained variance. All results were computed by lavaan. The standardized solution is completely standardized.

predictor of occupational turnover intention is continuance occupational commitment, for which the standardized path coefficient is  $-.672$ ; thus, higher estimation of the cost to leave a discipline predicts a lower level of the intention to do so.

The second strongest predictor of occupational turnover intention is continuance organizational commitment, and, surprisingly, the standardized coefficient for this predictor is positive, or  $.579$ . Higher perceived costs of leaving an organization therefore predicts a higher level of intention to leave the discipline. This result is an example of a suppression effect because, although the Pearson correlation between continuance organizational commitment and occupational turnover intention is about zero ( $-.02$ ; see Table 14.5), the standardized weight for the continuance organizational commitment is positive ( $.579$ ) once other predictors are held constant.

As expected, the direct effects of organizational turnover intention and occupational turnover intention on each other are positive. Of the two, the standardized mag-

nitude of the effect of organizational turnover intention on occupational turnover intention (.287) is stronger than the effect in the other direction (.037). If there is reciprocal causation between these two variables, then it is stronger in one direction than in the other. In other words, the impact of the intention to leave one's place of work on the intention to leave one's profession is greater than the magnitude of the influence in the other direction by a ratio of almost eight (.287/.037 = 7.76). Exercise 3 asks you to rerun this analysis but to constrain the direct effects of organizational and occupational turnover intention on each other to equality and then comment on the values of the corresponding path coefficients.

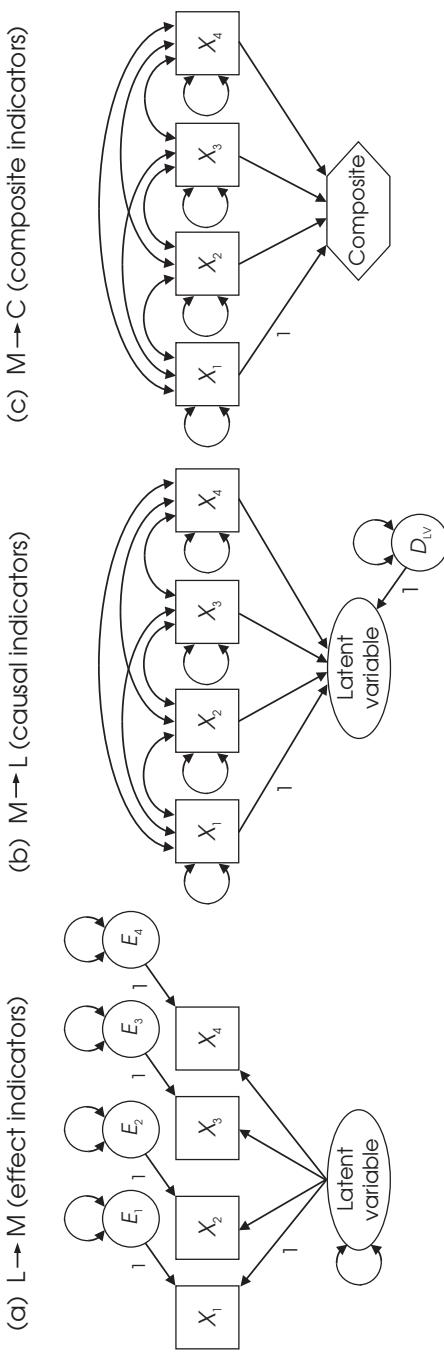
Two special topics in the analysis of nonrecursive structural models are described in the appendices of the chapter. These issues are relevant whether the structural model is a path model or part of an SR model. Appendix 14.B deals with effect decomposition in nonrecursive structural models and the assumption of equilibrium, and Appendix 14.C is about the estimation of corrected  $R^2$ -type proportions of explained variance for endogenous variables involved in feedback loops.

## **ANALYZING FORMATIVE MEASUREMENT MODELS IN SEM**

A reflective measurement model with four effect (reflective) indicators,  $X_1-X_4$ , is depicted in Figure 14.3(a). Grace and Bollen (2008) use the term **L → M block** (latent to manifest) to describe the “flow” of causation from the factor to the indicators. Reflective measurement assumes that indicators with equally reliable scores are interchangeable. It also requires positive intercorrelations among the indicators of a common factor. Measurement error in such models is represented at the indicator level by the terms  $E_1-E_4$ . The factor in Figure 14.3(a) is exogenous, but a factor can also be endogenous in a reflective measurement model (e.g., SR models). In two-step modeling, pattern coefficients for effect indicators should remain invariant over respecification of the structural model.

A formative measurement model with causal indicators is represented in Figure 14.3(b). Causal indicators have a conceptual unity in that they all measure the same latent continuum (Bollen & Bauldry, 2011). But that latent variable is caused by its indicators, which is described as an **M → L block** (manifest to latent) (Grace & Bollen, 2008). For example, income, education, and occupation, among other variables, determine a person's standing on a latent socioeconomic status (SES) dimension. A change in any of the variables just listed could affect the level of SES. In economics, costs of food, shelter, household furnishings, transportation, health care, and durable goods are examples of causal indicators of a formative cost of living factor. The disturbance in Figure 14.3(b) represents the possibility that causal indicators  $X_1-X_4$  do not explain all the variance in the corresponding latent variable. To scale the latent variable in Figure 14.3(b), the unstandardized direct effect of one of its causal indicators,  $X_1$ , is fixed to 1.0, which is a scaling constant.

Causal indicators are exogenous variables and thus have no error terms. Causal indicators are free to vary and covary, which explains the symbols in Figure 14.3(b) that



**FIGURE 14.3.** Directionalities of relations between indicators and (a) a latent variable with effect indicators, (b) a latent variable with causal indicators, and (c) a linear combination (composite) with composite indicators.  $M$ , manifest;  $L$ , latent;  $C$ , composite.

represent their variances and covariances. Measurement error in a formative measurement model like the one in Figure 14.3(b) is manifested in the disturbance term ( $D_{LV}$ )—that is, at the construct level, not at the indicator level as in reflective measurement (Figure 14.3(a)). Exercise 4 asks you to respecify an  $M \rightarrow L$  block like the one in Figure 14.3(b) but with three causal indicators so that measurement error is represented at the indicator level instead of at the construct level.

Causal indicators may have *any* pattern of intercorrelations, positive, negative, or even zero. What connects them is the researcher's theory that a set of causal indicators matches up with a target construct (Bollen & Bauldry, 2011). Direct effects of causal indicators should remain invariant over different outcomes of the latent variable measured by those indicators. Also, omitting a causal indicator that covaries with measured causal indicators may lead to bias. This can happen because the omission of such an indicator changes the empirical definition of the construct. So unlike effect indicators, causal indicators are not generally interchangeable.

The stumbling block in SEM to analyzing models where a factor has causal indicators only and its disturbance variance is freely estimated is identification. This is because it can be difficult to specify such a model that reflects the researcher's hypotheses and is identified. The need to scale a factor with causal indicators was mentioned earlier in this section, but meeting this requirement is not difficult (Bollen & Bauldry, 2011, describe other options). MacCallum and Browne (1993) noted that in order for the disturbance variance of a factor with causal indicators only to be identified, that factor must have direct effects on at least two other endogenous variables, such as factors with effect indicators. This requirement is the **2+ emitted paths rule**. If a factor measured with causal indicators only emits a single path, its disturbance variance will be underidentified. In models with multiple indirect pathways from effect-indicated factors to other such factors that pass through causal-indicated factors, some path coefficients may be underidentified.

The first two models in Figure 14.3 illustrate the identification issues just raised. Specifically, the model in Figure 14.3(a) with effect indicators is a standard CFA model. This model is identified (Rule 10.1); thus, it could be analyzed as a "stand-alone" model. But the model in Figure 14.3(b) with causal indicators is not identified. In order to estimate its parameters, it would be necessary to "embed" it within a larger SR model where the causal-indicated factor emits at least two direct effects, among other requirements for identification.

One way to deal with the identification problem is to add effect indicators to causal-indicated factors; that is, specify a MIMIC (multiple indicators and multiple causes) factor with both effect and causal indicators. Adding two effect indicators means that a factor previously with only causal indicators will now emit at least two directs (see Diamantopoulos, Riefler, & Roth, 2008). *Any such respecification requires a rationale*. For example, Hershberger (1994) described a MIMIC depression factor with indicators that represented various behaviors. Some of these indicators, such as "crying" and "feeling depressed," were specified as effect indicators because they are symptoms of depression. But another indicator, "feeling lonely," was specified as a causal indicator because this feeling may lead to depression, not vice versa.

If the disturbance variance in Figure 14.3(b) were fixed to zero (i.e., it is dropped from the model), the formerly latent variable is converted to a composite, or a linear combination of the indicators. The model just described is illustrated in Figure 14.3(c). It is called an **M → C block** (manifest to composite) (Grace & Bollen, 2008). These authors represent composites in diagrams with hexagons, which are also used in Figure 14.3(c). The hexagon is not a standard symbol, but it conveys the fact that a composite with no disturbance is not latent. Grace and Bollen also distinguish between a **fixed-weights composite** where coefficients for direct effects of composite indicators are specified a priori, such as unit weighting where all coefficients equal 1.0, and an **unknown weights composite** where the coefficients are estimated from the data. Figure 14.3(c) assumes an unknown weights composite.

**Composite indicators**, such as  $X_1-X_4$  in Figure 14.3(c), can be any arbitrary combination of variables; that is, they are not required to have a conceptual unity (Bollen & Bauldry, 2011). This means that (1) a set of composite indicators is not assumed to be unidimensional and (2) composite indicators can have any pattern of intercorrelations. Coefficients for composite indicators are not generally expected to be invariant over different outcomes predicted by the composite. This is because such coefficients could be optimized by the computer to predict a particular outcome. If the outcome changes, then the coefficients for the composite indicators could also change. This situation describes an unknown weights composite. An alternative is to assign a prior weights based on specific hypotheses regardless of the outcome variable.

A composite represents a convenient way to summarize the effects of several variables that may have little to do with each other. This is why the coefficients for composite indicators generally have little, if any, causal interpretation: The coefficients act simply to form the composite. For example, a composite of the variables age, gender, and ethnicity would have an imprecise unity in being demographic variables, and direct effects of such a composite would estimate the combined influence of these variables (Bollen & Bauldry, 2011). A drawback is that individual effects of age, gender, or ethnicity are lost when these variables are summed to form a composite. Instead, demographic variables are usually represented as individual covariates in structural equation models, a point elaborated next.

Covariates in SR models may be specified as correlated exogenous variables with direct effects on latent variables. A covariate is included in order to control for the effect of the corresponding variable. For example, including gender as a cause of a latent variable controls for differences between women and men when the computer calculates the path coefficients for other presumed causes of the same latent variable. Although covariates may have direct effects on factors, covariates do not measure those factors; that is, they are neither effect indicators nor causal indicators. Covariates can also be specified as causes of effect indicators or causal indicators, which may avoid bias in estimating the correspondence between latent variables and their indicators (Bollen & Bauldry, 2011).

Worland, Weeks, Janes, and Strock (1984) administered measures of verbal reasoning and scholastic achievement (reading, arithmetic, and spelling) within a sample of 158 adolescents. They also collected teacher reports about the classroom adjustment

(motivation, harmony, and emotional stability) and measured family SES and the degree of severe parental psychiatric disturbance. For pedagogical reasons, I generated the hypothetical correlations and standard deviations presented in Table 14.7 to match the basic pattern of associations reported by Worland et al. (1984) among these nine variables.

Suppose that *risk* is conceptualized as a latent variable with the causal indicators low family SES, parental psychopathology, and adolescent verbal IQ; that is, risk is affected by any combination of the variables just mentioned (plus error variance). Inter-correlations among these variables are not all positive (see Table 14.7), but this is irrelevant for causal indicators. Presented in Figure 14.4 is an SR model where a latent risk variable has causal indicators only. The risk factor emits two direct effects onto reflective factors (achievement, classroom adjustment), each measured with effect indicators only. These specifications meet the 2+ emitted paths rule and identify the disturbance variance for risk. It also reflects the hypothesis that the association between achievement and classroom adjustment is spurious due to risk. This assumption may not be plausible. For example, achievement may affect classroom adjustment (e.g., students with better scholastic skills may be better adjusted). But including the direct effect or, alternatively, a disturbance covariance between the two respective factors, would render Figure 14.4 not identified.

Exercise 5 asks to verify that  $df_M = 22$  for the model in Figure 14.4. I fitted this model to the covariance matrix based on the data in Table 14.7 using ML estimation in

**TABLE 14.7. Input Data (Hypothetical Correlations and Standard Deviations) for Analysis of a Model of Risk as a Latent Variable with Causal Indicators**

Variable	1	2	3	4	5	6	7	8	9
<u>Risk</u>									
1. Parental psychiatric		1.00							
2. Low family SES	.22		1.00						
3. Verbal IQ	-.43		-.49		1.00				
<u>Achievement</u>									
4. Reading	-.39		-.43	.58	1.00				
5. Arithmetic	-.24		-.37	.50	.73	1.00			
6. Spelling	-.31		-.33	.43	.78	.72	1.00		
<u>Classroom adjustment</u>									
7. Motivation	-.25		-.25	.39	.52	.53	.54	1.00	
8. Harmony	-.25		-.26	.41	.54	.43	.47	.77	1.00
9. Stability	-.16		-.18	.31	.50	.46	.47	.60	.62
SD	13.00	13.50	13.10	12.50	13.50	14.20	9.50	11.10	8.70

Note. N = 158.

LISREL (Scientific Software International, 2013) and Mplus (Müthen & Müthen, 1998–2014). All computer files for these analyses can be downloaded from this book’s website. It is somewhat tricky to specify a causal-indicated factor in LISREL SIMPLIS syntax or Mplus syntax. In LISREL, it is necessary to specify the “SO” option of the “LISREL output” command. This option suppresses LISREL’s automatic checking of the scale for each latent variable regardless of user specifications. A factor with causal indicators only is specified in Mplus using its syntax for regression (keyword “on”) instead of its syntax for measurement (keyword “by”), which is for factors with effect indicators only. But use of the keyword “by” in Mplus is required to specify the direct effects emitted by a causal-indicated factor. It is also necessary in Mplus syntax to explicitly specify that all such direct effects are free parameters.

Analyses in LISREL and Mplus both generated converged and admissible solutions. Values of fit statistics and parameter estimates are similar across both computer programs. Results from Mplus are as follows:

$$\begin{aligned}\chi^2_M(22) &= 35.308, p = .036 \\ \text{RMSEA} &= .062, 90\% \text{ CI } [.016, .098] \\ \text{CFI} &= .980; \text{SRMR} = .031\end{aligned}$$

The model fails the chi-square test at the .05 level. Values of the CFI and SRMR are not bad, but the RMSEA results are marginal. No absolute correlation residuals (calculated in EQS) exceeded .10, but among larger positive residuals were ones for pairwise associations between indicators of achievement and indicators of classroom adjustment. Thus, the model underpredicts these cross-factor associations. Although only three standardized residuals generated by Mplus were statistically significant, several others were almost so at the .05 level. Based on all these results, the model with the causal-indicated risk factor in Figure 14.4 is rejected.

The absence of a direct effect from achievement to classroom adjustment or a disturbance correlation between these two factors in Figure 14.4 may be a specification error, but adding either parameter to the model would make it not identified. Listed next are some options to respecify the model in order to estimate either parameter:

1. Respecify the causal-indicated risk factor in the original model as a MIMIC factor with at least one effect indicator, such as adolescent verbal IQ. This respecification makes sense because it assumes that family or parent variables (causal indicators) affect a characteristic of adolescents (effect indicator) through the latent risk variable. After respecification of risk as a MIMIC factor, it can be demonstrated using the replacing rules that model 1 with the direct effect

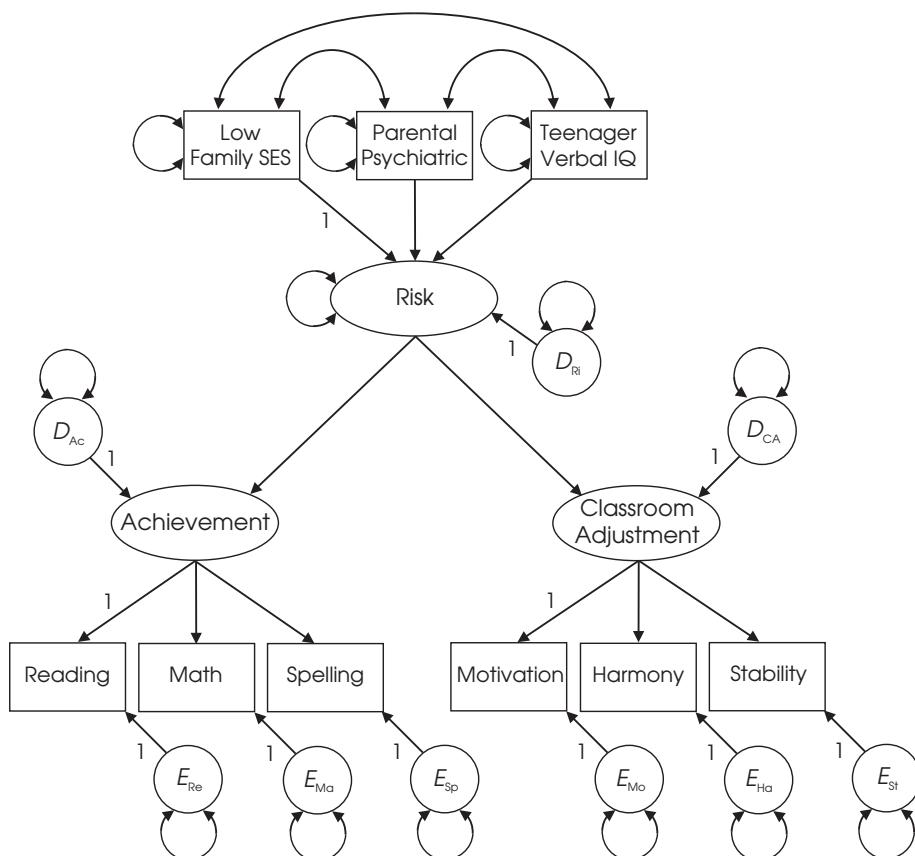
Achievement → Classroom Adjustment

but no disturbance correlation between these two factors and model 2 with the disturbance correlation but no direct effect between the same two factors are equivalent. Thus,

model 1 and model 2 would have exactly the same fit to the data, so in practice they are not empirically distinguishable.

2. Drop the disturbance  $D_{Ri}$  from the model in Figure 14.4, which would convert the latent risk variable into a weighted combination of its indicators (i.e., a composite). This option is not ideal. Dropping the disturbance by fixing its variance to zero would be akin to assuming that the indicators explain all the variance in risk, which is unlikely. MacCallum and Browne (1993) show that dropping a composite that emits a single path and converting the indirect effects of its indicators to direct effects on other variables result in an equivalent model.

3. Drop the latent risk variable altogether and replace it with direct effects from low family SES, parental psychopathology, and adolescent verbal IQ in the original model to each of two endogenous factors with effect indicators. Because the latent risk variable in the original model emits two direct effects (see Figure 14.4) instead of just one, this respecification would not generate an equivalent model.



**FIGURE 14.4.** An identified model of risk as a latent variable with causal indicators.

Analyzed next is a model with risk as a MIMIC factor with two causal indicators (low family SES, parental psychopathology) and one effect indicator (adolescent verbal IQ). Model 1 has a direct effect from the achievement factor to the classroom adjustment factor but no disturbance correlation. You should verify that  $df_M = 23$  for model 1. I fitted this model to the data in Table 14.7 using ML estimation in LISREL and Mplus. No special syntax is needed for this respecified model with a MIMIC factor and two reflective factors. Reported next are values of selected fit statistics calculated by Mplus:

$$\begin{aligned}\chi^2_{M1}(23) &= 35.308, p = .049 \\ \text{RMSEA} &= .058, 90\% \text{ CI } [.005, .094] \\ \text{CFI} &= .983; \text{SRMR} = .031\end{aligned}$$

The respecified model 1 just fails the chi-square test at the .05 level, and values of the RMSEA and CFI are marginally better than those for the original model in Figure 14.4.<sup>1</sup> There may be ways to improve the fit of model 1 that are not pursued here, but the estimated standardized direct effect of achievement on classroom adjustment calculated by Mplus is .627.

Analysis of respecified model 2 with a correlation between the disturbances of the achievement and classroom adjustment factors but with no direct effect between these factors generates exactly the same values of all fit statistics as model 1. In the Mplus output for model 2, the estimated disturbance correlation is .507. Computer files for the analyses in LISREL and Mplus of models with risk as a MIMIC factor can also be downloaded from this book's website.

Formative measurement and the analysis of composites are better known in economics, commerce, and biology than in the social sciences. For example, Grace (2006) and Grace and Bollen (2008) describe the analysis of composites in the environmental sciences. There is a special issue about formative measurement in the *Journal of Business Research* (Diamantopoulos, 2008). Jarvis, MacKenzie, and Podsakoff (2003) advised readers in the consumer research area—and the rest of us, too—not to automatically specify factors with effect indicators only because doing so may result in specification error, perhaps owing to lack of familiarity with formative measurement. This familiarity should make researchers aware that there are options for specifying the directionality of effects between latent variables and their indicators. This possibility should also prompt researchers to think very hard about measurement.

But formative measurement is no panacea. Because causal indicators are exogenous, their variances and covariances are not explained, which makes it more difficult to assess the construct validity of such indicators (Edwards, 2010), but Diamantopoulos and Winklhofer (2001) offer suggestions. Edwards (2010) notes that the lack of internal consistency expected for causal indicators can lead to misunderstanding. Suppose that

<sup>1</sup>Note that the model chi-squares for both the original model (Figure 14.4) and model 1 as just described are equal, 35.308, but the two models have different degrees of freedom, respectively, 22 versus 23. This explains the difference in the RMSEA and CFI results across the two models.

a researcher observes low intercorrelations among a set of measures and concludes that they must be formative indicators. This decision is not justified because low intercorrelations in this case could merely indicate poorly constructed measures. Howell, Breivik, and Wilcox (2007) conclude that (1) formative measurement is not an equally attractive alternative to reflective measurement, and (2) researchers should try to include effect indicators whenever other indicators are specified as causal indicators of the same construct, but see Bollen (2007) for other views.

An alternative to SEM for analyzing models with both measurement and structural parts is **partial least squares path modeling**. It is a two-stage, iterative, components-based method that estimates hypothetical constructs as linear combinations (composites). In the first stage of iterative estimation, weights that associate indicators with components are estimated. Observed variables specified as effect indicators are eventually regressed on their common component. In the second stage, weights that associate the components are estimated with an emphasis on maximizing predictive power. Although SEM is better for testing strong hypotheses about measurement, the partial least squares method is well suited for situations where (1) prediction is emphasized over theory testing and (2) it is difficult to meet the requirements for larger samples or identification of causal-indicated factors in SEM—see Topic Box 14.1 for more information.

#### TOPIC BOX 14.1

### Partial Least Squares Path Modeling

A good starting point for outlining the logic of partial least squares path modeling (PLS-PM) is the distinction between principal components analysis versus common factor analysis. Briefly, principal components analysis analyzes total variance and estimates factors as linear combinations of indicators, but common factor analysis analyzes shared (common) variance only and makes an explicit distinction between indicators, underlying factors, and error (unique) variance. Of these two EFA methods, it is principal components analysis that is directly analogous to PLS-PM.

The idea behind PLS-PM is based on **soft modeling**, an approach developed by H. Wold (1982) for situations where theory about measurement is not strong, but the goal is to estimate predictive relations between latent variables. Latent variables are initially estimated as simple composites based on their indicators in an iterative algorithm. Later in iterative estimation, indicators are regressed on the composites. The method is basically an extension of canonical correlation but one that (1) explicitly distinguishes indicators and components and (2) permits the estimation of direct or indirect effects among components. Similar to canonical correlation, indicators in PLS-PM are weighted in order to maximize prediction. In contrast, the goal of estimation in SEM is to minimize residual covariances, which may not maximize prediction.

Estimation methods in PLS-PM make fewer demands of the data. For example, they do not generally assume a particular distributional form for the original scores, and the process of iterative estimation is not as complex. Consequently, it may be possible to apply PLS-PM in smaller samples than SEM, and there are generally no problems concerning inadmissible solutions. There are also few issues about identification in PLS-PM. For example, there is no problem in PLS-PM with specifying that a component with causal indicators emits a single direct effect. These characteristics make the analysis of complex models with many indicators (effect or causal) easier in PLS-PM compared with SEM.

A drawback of PLS-PM is that its limited-information estimators are statistically inferior to those generated under full-information methods (e.g., ML in SEM) in terms of bias and consistency, but this is less so in very large samples. Standard errors in PLS-PM are estimated using adjunct methods, including bootstrapping. Researchers generally evaluate models in PLS-PM by inspecting values of pattern coefficients, path coefficients, and  $R^2$ -type statistics for outcome variables (i.e., local fit testing). There are statistics for global fit testing in PLS-PM, but it is harder to evaluate overall model fit in PLS-PM compared with SEM. One could argue that PLS-PM, which generally analyzes unknown weights composites, does not really estimate substantive latent variables compared with SEM. The advantages of PLS-PM over SEM include flexibility, robustness, and fewer demands concerning distributional assumptions and requirements for identification (see Rigdon (2013) and McIntosh, Edwards, & Antonakis (2014) for more information).

Several commercial and open-source computer tools are now available for PLS-PM. For example, SmartPLS (Ringle, Wende, & Becker, 2014) is a commercial program with a graphical interface where users generate models on the screen with a palette of drawing tools. There is a free student version. The PLS-Graph program (Chin, 2001) has similar features and is free for academic users. There are also PLS-PM packages for R, such as semPLS (Monecke, 2014).

## SUMMARY

The evaluation of an SR model is essentially a simultaneous path analysis and confirmatory factor analysis. Single- or multiple-indicator assessment of constructs is represented in the measurement part of an SR model, and presumed causal effects are represented in the structural part. In two-step modeling, hypotheses about measurement are evaluated in the first step. An acceptable measurement model is required before going to the second step, which involves testing hypotheses about the structural model. The specification of reflective measurement wherein effect indicators are specified as caused by latent variables is not appropriate in all research problems. An alternative is formative measurement where indicators are conceptualized as causes of latent variables.

The evaluation of SR models represents the apex in the SEM family for the analysis of covariances. Part IV deals with some advanced methods, starting with the analysis of means. Best practices in SEM are considered in the last chapter, which may be the most important one in the book.

### LEARN MORE

Bollen and Bauldry (2011) define effect, causal, and composite indicators, Edwards (2010) criticizes the concept of formative measurement, and Rigdon (2013) introduces the technique of PLS-PM from the perspective of SEM.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, 16, 265–284.

Edwards, J. R. (2010). The fallacy of formative measurement. *Organizational Research Methods*, 14, 370–388.

Rigdon, E. E. (2013). Partial least squares path modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 81–116). Charlotte, NC: IAP.

### EXERCISES

1. Evaluate the factor covariances in Table 14.4 for statistical significance at the .05 level.
2. Use the results in Figure 14.2 to conduct a standardized effect decomposition.
3. Reanalyze the final nonrecursive model of turnover intention (see Figure 14.2 for the initial model) by constraining to equality the direct effects of the feedback loop. Compare the unstandardized and standardized coefficients.
4. Respecify Figure 14.3(b) but with just three causal indicators in order to control for measurement error at the indicator level.
5. Verify that  $df_M = 22$  for Figure 14.4.

## **Appendix 14.A**

### Constraint Interaction in SR Models

Recall that constraint interaction for CFA models is indicated when the value of the chi-square difference statistic for the test of the equality of the pattern coefficients for indicators of different factors depends on how the factors are scaled (Appendix 13.B). Steiger (2002) shows that the same phenomenon can happen with SR models where some factors have only two indicators and when estimates of direct effects on two or more different endogenous factors are constrained to be equal. Constraint interaction can also result in an incorrect standardized solution for an SR model if it is calculated in the way described earlier (unstandardized solution with ULI constraints, then standardize that solution).

The presence of constraint interaction can be detected the same way for SR and CFA models: While imposing the equality constraint, change the value of each identification constraint for the factors from 1.0 to another positive scaling constant and then rerun the analysis. If the value of the model chi-square changes by an amount that exceeds what is expected by rounding error, there is constraint interaction. Steiger (2002) suggests a way to deal with constraint interaction in SR models: If the analysis of standardized factors can be justified, the method of constrained estimation can be used to test hypotheses of equal standardized path coefficients and to generate correct standard errors. Constrained estimation of an SR model standardizes all factors, exogenous and endogenous.

## Appendix 14.B

### Effect Decomposition in Nonrecursive Models and the Equilibrium Assumption

The tracing rule does not apply to nonrecursive structural models with feedback loops. Variables in feedback loops have indirect effects—and thus total effects—on *themselves*, which is apparent in effect decompositions calculated by SEM computer programs for nonrecursive models. Consider the reciprocal relation  $Y_1 \leftrightarrow Y_2$ . Suppose that the standardized direct effect of  $Y_1$  on  $Y_2$  is .40 and that the effect in the other direction is .20. An indirect effect of  $Y_1$  on itself would be the sequence

$$Y_1 \rightarrow Y_2 \rightarrow Y_1$$

which is estimated as  $.40 \times .20$ , or .08. There are additional indirect effects of  $Y_1$  on itself through  $Y_2$ , however, because cycles of mutual influence in feedback loops are theoretically infinite. The indirect effect

$$Y_1 \rightarrow Y_2 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_1$$

is one of these, and its estimate is  $.40 \times .20 \times .40 \times .20$ , or .0064. Mathematically, these terms head quickly to zero, but the total effect of  $Y_1$  on itself is an estimate of all possible cycles through  $Y_2$ . Indirect and total effects of  $Y_2$  on itself are similarly derived.

Calculation of indirect and total effects among variables in a feedback loop as just described assumes equilibrium. Recall that there is no statistical test of whether the equilibrium assumption is tenable when the data are cross sectional. In a computer simulation study, Kaplan et al. (2001) found that the **stability index**, which is printed in the output of some SEM computer programs, did not accurately measure the degree of bias due to lack of equilibrium. It is based on certain mathematical properties of the matrix of coefficients for direct effects among all the endogenous variables in a structural model, not just those involved in feedback loops. These properties concern whether estimates of the direct effects would get infinitely larger over time. If so, the system is said to “explode” because it may never reach equilibrium, given the observed direct effects among the endogenous variables. The mathematics of the stability index are complex (Kaplan et al., 2001, pp. 317–322). A standard interpretation of this index is that values less than 1.0 are taken as positive evidence for equilibrium, but values greater than 1.0 suggest the lack of equilibrium. But this interpretation is not generally supported by Kaplan and colleagues’ simulation results, which emphasizes the need to evaluate equilibrium on rational grounds.

## Appendix 14.C

### Corrected Proportions of Explained Variance for Nonrecursive Models

Several authors note that  $R^2$  calculated as one minus the ratio of the disturbance variance over the total variance may be inappropriate for endogenous variables involved in feedback loops. This is because the disturbances of such variables may be correlated with one of their presumed causes, which violates the least squares requirement that the residuals (disturbances) are uncorrelated with all predictors (causal variables). Some corrected  $R^2$  statistics for nonrecursive models are described next:

1. The **Bentler-Raykov corrected  $R^2$**  (Bentler & Raykov, 2000) is based on a respecification that repartitions the variance of endogenous variables controlling for correlations between disturbances and causal variables. This statistic is automatically printed by EQS, Stata, and some other SEM computer tools for nonrecursive models.
2. Version 8 of LISREL (Scientific Software International, 2006) printed a **reduced-form  $R^2$**  for each endogenous variable in a structural model. In **reduced form**, the endogenous variables are regressed on the exogenous variables only. This regression also has the consequence that all direct effects of disturbances on their respective endogenous variables are removed or blocked, which also removes any contribution from all other endogenous variables (Hayduk, 2006). For nonrecursive models, the value of the reduced-form  $R^2$  can be substantially less than that of  $R^2$  for the same variable.
3. Hayduk (2006) describes the **blocked-error  $R^2$**  for variables in feedback loops or with correlated errors. It is calculated by blocking the influence of the disturbance of just the variable in question (the focal endogenous variable). An advantage of this statistic is that it equals the value of  $R^2$  for each endogenous variable in a recursive model. The blocked-error  $R^2$  is automatically printed by Version 9 of LISREL (Scientific Software International, 2013) for nonrecursive models. Hayduk (2006) describes a method for calculating the blocked-error  $R^2$  using any SEM program that prints the predicted covariance matrix when all parameters are fixed to equal user-specified values.

Depending on the model and data, the corrected  $R^2$ 's just described can be either smaller or larger than that of  $R^2$  for endogenous variables in feedback loops. For example, reported next are values of  $R^2$  (calculated using the results in Table 14.6), the Bentler-Raykov  $R^2$  (computed in EQS), the reduced-form  $R^2$  (computed in LISREL 8), and the blocked-error  $R^2$  (computed in LISREL 9) for the two endogenous variables involved in a causal loop in Figure 10.7. These

results are from the analysis of the final model with a direct effect from continuance organizational commitment to occupational turnover intention:

Endogenous variable	$R^2$	BR $R^2$	RF $R^2$	BE $R^2$
OrgTI	.542	.541	.520	.521
OccTI	.766	.764	.658	.690

*Note.* BR, Bentler-Raykov; RF, reduced-form; BE, blocked-error; OrgTI, organizational turnover intention; OccTI, occupational turnover intention.

The three sets of corrected  $R^2$  values for the same model and data are equally correct because they represent somewhat different ways to control for predicted correlations between causal variables and disturbances. Indicate in written reports the particular  $R^2$  statistic used to estimate the proportions of explained variance for endogenous variables in nonrecursive models.

## **Part IV**

# Advanced Techniques and Best Practices



# 15

## Mean Structures and Latent Growth Models

---

The basic datum of SEM, the covariance, does not convey information about means. If only covariances are analyzed, then all variables are effectively mean-deviated (centered) so that all substantive latent variables must have means of zero. Sometimes this loss of information is too restrictive, such as when the means of repeated measures variables are expected to differ. Means are estimated in SEM by adding a mean structure to the model's basic covariance structure. The input data for the analysis of a model with a mean structure are covariances and means (or the raw scores). The SEM approach is distinguished by the capability to test hypotheses about the error covariances or about the means of substantive latent variables. The analysis of latent growth models with longitudinal data from a single sample is also considered in this chapter. Latent growth models are analyzed in several different research areas, so they have wide application.

---

### LOGIC OF MEAN STRUCTURES

Multiple regression provides the basic rationale for analyzing covariance structures of observed variables in SEM. It also provides the logic for analyzing means. Recall that unstandardized regression equations have both a covariance structure ( $B$  weights) and a mean structure in the form of the intercept ( $A$ ) (e.g., Equation 2.1). For example, consider the scores on variables  $X$  and  $Y$  in Table 15.1. The unstandardized regression equation for these data is

$$\hat{Y} = .455X + 20.000$$

The regression coefficient, .455, conveys no information about the means of either variable (see Equation 2.2). The intercept, 20.000, reflects the mean of both variables and the

**TABLE 15.1. Example Bivariate Data Set**

Case	Raw scores		Constant $\Delta$
	X	Y	
A	3	24	1
B	8	20	1
C	10	22	1
D	15	32	1
E	19	27	1
M	11.000	25.000	—
SD	6.205	4.690	—
$s^2$	38.500	22.000	—

Note.  $r_{XY} = .601$ .

regression coefficient, albeit with a single number. Given  $M_X = 11.000$  and  $M_Y = 25.000$  (Table 15.1), the intercept can be expressed according to Equation 2.3 as

$$A = 25.000 - .455 (11.000) = 20.000$$

Likewise, the mean of Y can be expressed as a function of the intercept, regression coefficient, and mean of X, as follows:

$$M_Y = 20.000 + .455 (11.000) = 25.000$$

How a computer calculates the intercept in regression analysis provides the key to understanding the analysis of means in SEM. Look again at Table 15.1 and in particular at the column labeled  $\Delta$ , which represents a constant that equals 1 for every case in this application of McArdle–McDonald RAM symbolism. Results of two regression analyses with the constant are summarized in Table 15.2. Both analyses were conducted by instructing the computer to omit from the analysis the intercept term it would otherwise automatically calculate. In the first analysis, Y is regressed on both X and the constant. Note that the coefficient for X is the same as before, .455, and for the constant it is 20.000, the intercept. The second analysis in Table 15.2 concerns the regression of X on the constant. The coefficient in this analysis is 11.000, or the mean of X. These results illustrate two basic principles about mean structures:

---

For a continuous criterion and predictor, (Rule 15.1)

1. if the criterion is regressed on both the predictor and constant, the unstandardized coefficient for the constant is the intercept; and
  2. if the predictor is regressed on the constant, the unstandardized coefficient for the constant is the mean of the predictor.
-

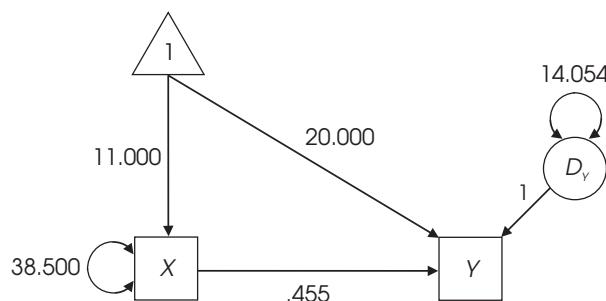
**TABLE 15.2. Results of Regression Analyses with a Constant for the Data in Table 15.1**

Regression	Predictor(s)	Unstandardized coefficient(s)
1. Y on X and $\Delta$	X	.455
	$\Delta$	20.000
2. X on $\Delta$	$\Delta$	11.000

A path-analytic representation of the analyses is presented in Figure 15.1, if we now assume that  $X$  causes  $Y$ . Unlike a standard path model, the one in the figure has both a covariance structure and a mean structure. The covariance structure includes the direct effect of  $X$  and the variances of both  $X$  and the disturbance  $D_Y$ . Analyzing this covariance structure with the data in Table 15.1 using OLS estimation yields an unstandardized path coefficient of .455—the same as the unstandardized regression coefficient—and a disturbance variance of 14.054.<sup>1</sup> No means are represented in this covariance structure.

The mean structure in Figure 15.1 consists of direct effects of the constant on both  $X$  and  $Y$ . Although the constant is depicted as exogenous in the figure, it is not a variable in the usual sense because it has no variance. The unstandardized coefficient for the direct effect of the constant on  $X$  is 11.000, or the mean of  $X$ , just as in the corresponding regression analysis (Table 15.2). The mean of  $X$  is thus represented in the mean structure of the path model in the form of an unstandardized path coefficient. Because the constant has no indirect effects on  $X$  through other variables, the unstandardized coefficient for the path  $\Delta \rightarrow X$  is also the total effect.

The unstandardized coefficient for the direct effect of the constant on endogenous variable  $Y$  is 20.000, or the intercept from regressing  $Y$  on  $X$ . The constant also has an indirect effect on  $Y$  through  $X$ . Using the tracing rule for this model, we obtain this result:

**FIGURE 15.1.** A path model with a mean structure.

<sup>1</sup>The error variance is calculated in bivariate OLS estimation as  $(1 - r_{XY}^2)s_Y^2 = (1 - .601^2) 22.000 = 14.054$ .

$$\begin{aligned}\text{Total effect of } \Delta \text{ on } Y &= \Delta \rightarrow Y + \Delta \rightarrow X \rightarrow Y \\ &= 20.000 + .455 (11.000) = 25.000\end{aligned}$$

which equals the mean of  $Y$ . Two additional principles about mean structures can thus be expressed in path analytic language:

---

For a continuous endogenous and exogenous variable, (Rule 15.2)

1. the mean of the endogenous variable is a function of three parameters—  
(a) the intercept, (b) the unstandardized path coefficient, and (c) the mean of the exogenous variable; and
  2. the predicted mean for either variable is the total effect of the constant.
- 

The mean structure in Figure 15.1 is just-identified (i.e., two means, two direct effects of  $\Delta$ ), so the predicted means for  $X$  and  $Y$  equal their observed counterparts. This fact is elaborated next.

When an SEM computer tool analyzes means, it automatically creates a constant on which variables in the model are regressed. A variable is included in a mean structure by specifying that the constant has a total effect on it. This leads to two more principles:

---

In structural equation models with continuous variables, (Rule 15.3)

1. for exogenous variables, the unstandardized path coefficient for the direct effect of the constant is a mean; and
  2. for endogenous variables, the direct effect of the constant is an intercept but the total effect is a mean.
- 

If a variable is excluded from the mean structure, its mean is assumed to be zero. Error terms are *never* included in mean structures because their means are always assumed to equal zero. The mean structure may not be identified if the mean of an error term is inadvertently specified as a free parameter. Three points warrant special mention:

1. There is no standard symbol in the SEM literature for mean structures. The symbol  $\Delta$  is used in diagrams here so that you can quickly recognize the presence of a mean structure and determine which variables it includes. But it is not absolutely necessary to explicitly represent mean structures in model diagrams. Some authors just present the covariance structure in a diagram and report estimates about means in accompanying tables.
2. When a mean structure is added to a path model (e.g., Figure 15.1), an intercept is included in the equation for each endogenous variable, and estimates of those intercepts are reported in computer output along with the estimates of other free parameters.

The presence of an intercept makes it easier to calculate a predicted score for each case in the metric of the corresponding endogenous variable. This can be done manually by creating in a data editor a new composite where the weights match those of the path coefficients and intercept from the equation for a particular outcome. The `sem` command in Stata can automatically save predicted scores to the raw data file for either observed or latent endogenous variables.

3. Special forms of the ML method for incomplete raw data files, including the expectation–maximization (EM) algorithm, estimate both covariances and means; that is, they add a mean structure to the model. Depending on how these special methods are implemented in a particular computer tool, it may or may not be necessary to explicitly specify a mean structure even if the original model has only a covariance structure.

## **IDENTIFICATION OF MEAN STRUCTURES**

The principles listed next concern identification:

---

The parameters of a model with a mean structure include the (Rule 15.4)

1. means of the exogenous variables (except errors);
  2. intercepts of the endogenous variables; and
  3. number of parameters in the covariance structure counted in the usual way for that type of model.
- 

A simple rule for counting the number of observations is stated next:

---

If  $v$  is the number of observed variables, then the number (Rule 15.5)  
of observations equals  $v(v + 3)/2$  when means are analyzed.

---

The value of the expression in Rule 15.5 gives the total number of variances, nonredundant covariances, and means of observed variables. If there are three continuous variables, for example, then there are  $3(6)/2$ , or nine observations, including three means, three variances, and three unique covariances (e.g., see the lower right side of Table 4.1).

In order for a mean structure to be identified, the number of its parameters cannot exceed the total number of observed means. The identification status of a mean structure must be considered separately from that of the covariance structure. For example, an overidentified covariance structure will not identify an underidentified mean structure, and vice versa. If the mean structure is just-identified, it has as many free parameters (direct effects of the constant) as observed means; therefore, (1) the predicted means (total effects of the constant) will exactly equal the corresponding observed means, and

(2) the fit of the model with just the covariance structure will be identical to that of the model with both the covariance and mean structures.

For example, the mean structure of the path model in Figure 15.1 has two parameters,

$$\Delta \rightarrow X \quad \text{and} \quad \Delta \rightarrow Y$$

respectively, the mean of  $X$  and the intercept when regressing  $Y$  on  $X$ . Because there are two observed means ( $M_X, M_Y$ ), the mean structure here is just-identified. It was demonstrated earlier for this model that the total effect of the constant on  $X$  is 11.000 and on  $Y$  it is 25.000. Each of these predicted means equals the corresponding observed mean (Table 15.1). It is only when the mean structure is overidentified that the predicted means could differ from the observed values; that is, some **mean residuals** may not equal zero. Mean residuals are differences between observed means and predicted means. A **standardized mean residual** is the ratio of a mean residual over the estimated standard error of that difference (i.e., a  $z$  test in large samples).

## ESTIMATION OF MEAN STRUCTURES

Most estimation methods described in earlier chapters for analyzing models with covariance structures only can be applied to models with both covariance and mean structures. Incremental fit indexes, such as the Bentler CFI, may not be calculated for models with mean structures, or they may be calculated for just the covariance part of the model. The independence model is more difficult to define for models with mean structures. For example, an independence model where both covariances and means are fixed to zero may be very unrealistic. An alternative for outcomes that are not repeated measures variables allows for means to equal their observed (sample) values. For repeated measures, though, the null model should set the means as equal (i.e., no change), such as when means from the same variable are fixed to equal the value at the first measurement occasion (Kenny, 2014a). Check the documentation of your SEM computer tool to determine how it defines the independence model when means are analyzed.

## LATENT GROWTH MODELS

The term **latent growth model** (LGM)—also known as a **latent curve model**—refers to a class of models for longitudinal data that can be analyzed in SEM or other statistical techniques, such as hierarchical linear modeling (HLM) (Garson, 2013). It may be the most common type of structural equation model with a mean structure evaluated in a single sample. The kinds of latent growth models outlined in this section have been described by several authors in SEM (Bollen & Curran, 2006; Preacher, Wichman, Mac-

Callum, & Briggs, 2008), are specified as SR models with a mean structure and can be analyzed with standard SEM software.

The analysis of an LGM in SEM typically requires scores that have the same units (metric) across time and can be said to measure the same construct at each assessment. It also generally requires that the data are **time structured**, which means that all cases are tested at the same intervals. These intervals need not be equal. Suppose that a group of children are observed at 3, 6, 12, and 24 months of age. If other children are tested at, say, 4, 10, 15, and 30 months, their data cannot be analyzed together with those tested at other intervals, if time-structured data are required. In contrast, HLM does not require time-structured data.<sup>2</sup> Another advantage of HLM is that it is more flexible than SEM concerning missing or unbalanced data (different numbers of cases are tested at different occasions). The SEM approach offers these relative advantages: the availability of global fit statistics and the capability to simultaneously analyze multiple growth curves or to model growth of latent variables, not just observed variables.

The raw scores are not generally needed to analyze an LGM. This is because such models can be analyzed with matrix summaries, if all endogenous variables are continuous. But these summaries must include the covariances (or correlations and standard deviations) and means of all variables, even those that are not repeated measures variables. Willett and Sayer (1994) note that inspection of the **empirical growth record** (raw scores for each case) can help to determine whether it may be necessary to model curvilinear change over time. Curvilinear growth means that the rate of change is conditional in that it depends on the particular measurement occasion. Simple linear growth is uniform over all occasions. It is also possible to generate predicted growth curves for individual cases, but only when a raw data file is analyzed.

As noted by Bauer (2003) and Curran (2003), latent growth models are actually multilevel (two-level) models that explicitly acknowledge the fact that scores are clustered under individuals (repeated measures). Scores from the same case are probably not independent, and this lack of independence must be controlled in the analysis. The parameterization of an LGM in HLM is different, but HLM and SEM computer tools generate the same basic parameter estimates for the same model and data. This is a point of isomorphism between HLM and SEM (Curran, 2003).

## DETAILED EXAMPLE

The data for this example are from Browne and Du Toit (1991), who analyzed scores from a six-trial computerized air traffic controller task where the score is the number of successful landings. This task was administered by Kanfer and Ackerman (1989) within a sample of 137 military pilots. This sample size is so small that technical problems may

<sup>2</sup>Some SEM computer tools, such as *Onyx* and *Mplus*, do not require time-structured data when analyzing latent growth models.

be encountered in the analysis. To avoid this problem, I specified a more realistic sample size of  $N = 250$ . This changes the standard errors and values of some fit statistics, but not the basic parameter estimates for this pedagogical example.

Presented in Table 15.3 are summary statistics for the six trials. The mean increases from 11.77 at trial 1 to 34.20 at trial 6. The standard deviation increases, too, from 7.60 to 9.62 from the first trial to the last trial; that is, the scores become more variable over trials. Correlations between adjacent trials are generally higher than those between nonadjacent trials (see the table), which is typical for learning trial data. The ability variable in the table refers to a measure of general cognitive aptitude where a higher score indicates a better result. This variable is analyzed later as a predictor of both initial performance and change in performance. Mean scores over trials exhibit both linear and curvilinear trends, which are apparent in Figure 15.2. The specific form of the curvilinear trend is a negative quadratic trend, which says that increases in performance decelerate over trials.

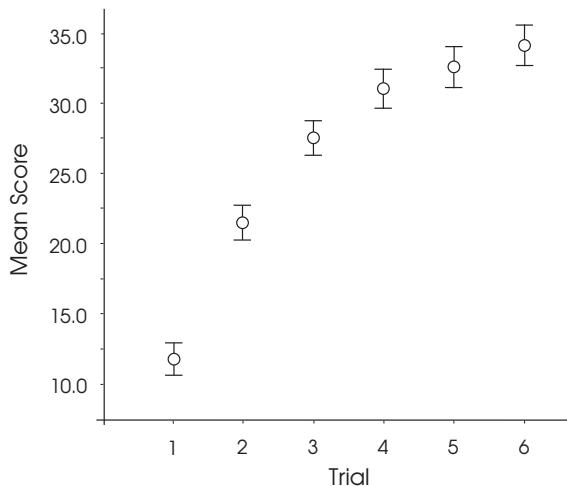
## Modeling Change

Latent growth models are often analyzed in two steps. The first concerns a basic change model of just the repeated measures variables. This model attempts to explain the covariances and means of these variables. Given an acceptable change model, the second step adds covariates to the model that may predict change. For example, does level of general cognitive ability predict initial performance in the air traffic controller task or the rate of improvement over subsequent trials? The two-step approach makes it easier to detect

**TABLE 15.3. Input Data (Correlations, Standard Deviations, Means) for Latent Growth Models of Performance on a Computerized Air Traffic Controller Task**

Variable	Trial						Ability
	1	2	3	4	5	6	
<b>Trial</b>							
1	1.00						
2	.77	1.00					
3	.59	.81	1.00				
4	.50	.72	.89	1.00			
5	.48	.69	.84	.91	1.00		
6	.46	.68	.80	.88	.93	1.00	
Ability	.50	.46	.36	.26	.28	.28	1.00
M	11.77	21.39	27.50	31.02	32.58	34.20	.70
SD	7.60	8.44	8.95	9.21	9.49	9.62	5.62

*Note.* These data are from Browne and Du Toit (1991);  $N = 250$ .



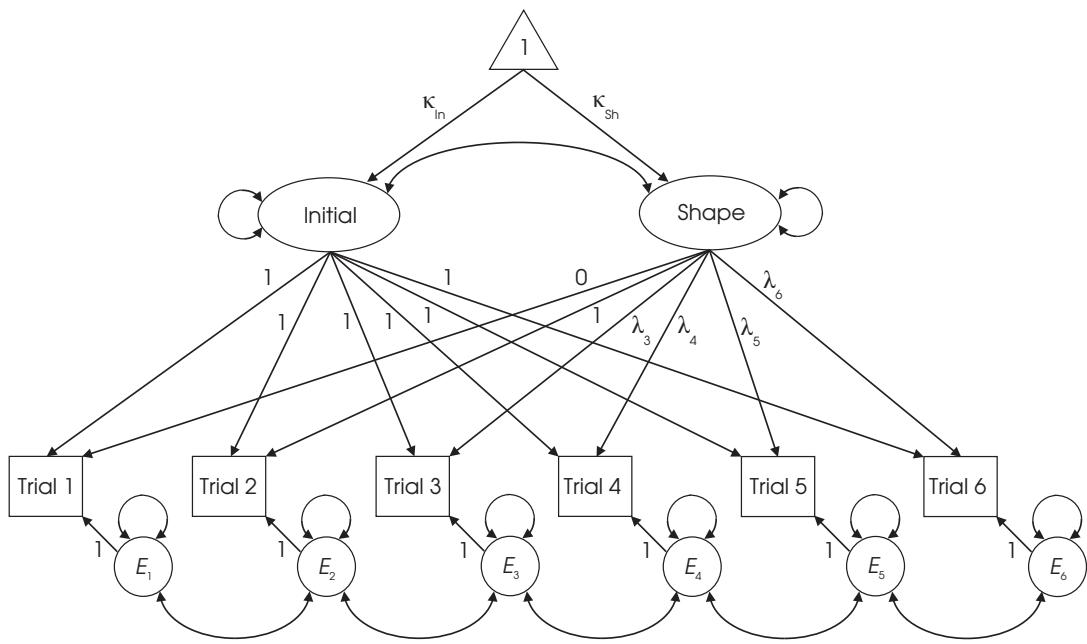
**FIGURE 15.2.** Means and 95% confidence intervals for the learning trial data in Table 15.3.

potential specification error compared with the analysis of a prediction model in a single step. There is a similar rationale for analyzing SR models in two steps.

A basic change model for the air traffic controller task is presented in Figure 15.3. It has the following features:

1. The model in the figure represents a strategy for analyzing latent growth models described by Meredith and Tisak (1990) and Kaplan (2009) and referred to as **nonlinear curve fitting**. Models analyzed in this approach have two latent growth factors, designated as Initial and Shape in the figure. The indicators of both factors are repeated measures variables, or the six trials of the learning task. The six trials are specified as endogenous variables with error terms. This implies that estimates for the latent growth factors are adjusted for unexplained variation (which includes measurement error) in their indicators (trials 1–6 on the learning task).
2. Because the Initial factor is analogous to the intercept in a latent variable regression equation, the unstandardized pattern coefficients for this factor are all fixed to 1.0 (see Figure 15.3). Two pattern coefficients for the Shape factor are fixed to constants that correspond to times of measurement, beginning with zero for the first trial and then 1.0 for the second trial. The former (the fixed coefficient of 0) specification sets the initial level (origin) at trial 1. Consequently, the Initial factor will be based on the trial 1 measurement.<sup>3</sup> The latter specification (i.e., the fixed coefficient of 1.0) scales the Shape factor, which is required for identification.

<sup>3</sup>The origin can be set to other times besides the first observation. For example, the weights  $(-1, 0)$  for, respectively, trials 1 and 2 would specify that the initial level is now based on the second measurement. Where the origin is set may affect estimates of factor variances and covariances; see Willett and Sayer (1994) for more information.



**FIGURE 15.3.** Latent growth model of change in performance on a computerized air traffic controller task.  $\kappa$ , means of exogenous latent growth factors;  $\lambda$ , freely estimated pattern coefficients for the Shape factor.

3. The remaining pattern coefficients for the Shape factor, designated as  $\lambda_3-\lambda_6$  for trials 3–6 in Figure 15.3, are free parameters. This specification results in what is basically an empirical trend made up of linear or curvilinear facets that optimally fit the Shape factor to the data in a particular sample. Relative values of these freely estimated coefficients together with the fixed coefficients for trials 1 and 2 describe the overall pattern of change. For example, if the relative increases in values of all pattern coefficients for the Shape factor are not constant over trials, then change is curvilinear.

4. The covariance between the Initial and Shape factors indicates the degree to which initial performance predicts the rate of subsequent change. A positive covariance would indicate that better performance on trial 1 predicts a higher rate of improvement over subsequent trials (higher initial level and then faster growth), and a negative covariance would indicate just the opposite (lower initial level but then faster growth). A covariance close to zero would say that initial level of performance has no bearing on the rate of subsequent improvement.

5. The constant in Figure 15.3 has direct effects on the Initial and Shape factors. This specification includes the means of the latent growth factors as free parameters, which are designated in the figure as  $\kappa_{In}$  and  $\kappa_{Sh}$  for, respectively, the Initial and Shape factors. The character  $\kappa$  (lowercase Greek letter kappa) designates means of exogenous

factors in LISREL symbolism. The mean of the Initial factor,  $\kappa_{In}$ , is the average initial level of performance at trial 1, controlling for error variance at the first trial. This average is a characteristic of the whole sample. In contrast, the variance of the Initial factor (see the figure) reflects variation around the average initial level. The mean for the Shape factor,  $\kappa_{Sh}$ , reflects the average amount of improvement in performance from trial 1 to trial 2, also adjusted for unexplained variation. The variance of the Shape factor provides information about the range of individual differences in the rate of improvement over subsequent trials.

6. Means over the six trials, which are endogenous, are not model parameters. But the unstandardized total effects of the constant on them are predicted means that can be compared with the observed means. We can apply the tracing rule to Figure 15.3 to see how the model generates predicted means, which are determined by the factor means and pattern coefficients. For example, the constant has the two indirect effects on trial 1 presented next:

$$\begin{aligned}\triangle \rightarrow \text{Initial} \rightarrow \text{Trial 1} &= \kappa_{In} \times 1.0 = \kappa_{In} \\ \triangle \rightarrow \text{Shape} \rightarrow \text{Trial 1} &= \kappa_{Sh} \times 0 = 0\end{aligned}$$

The total effect of the constant on trial 1 is the sum of the two indirect effects just listed, or  $\kappa_{In}$ . In words, the predicted mean for trial 1 is the mean of the Initial factor. The predicted mean for trial 2 is the sum of both indirect effects of constant through the latent growth factors, or

$$\begin{aligned}\triangle \rightarrow \text{Initial} \rightarrow \text{Trial 2} &= \kappa_{In} \times 1.0 = \kappa_{In} \\ \triangle \rightarrow \text{Shape} \rightarrow \text{Trial 2} &= \kappa_{Sh} \times 1.0 = \kappa_{Sh}\end{aligned}$$

The predicted mean for trial 2 thus equals  $\kappa_{In} + \kappa_{Sh}$ . The latter term,  $\kappa_{Sh}$ , indicates the average amount of improvement from trial 1 to trial 2.

Applying the tracing rule once again to Figure 15.3, we find that the total effect of the constant on the third trial is the sum of the indirect effects, listed next:

$$\begin{aligned}\triangle \rightarrow \text{Initial} \rightarrow \text{Trial 3} &= \kappa_{In} \times 1.0 = \kappa_{In} \\ \triangle \rightarrow \text{Shape} \rightarrow \text{Trial 3} &= \kappa_{Sh} \times \lambda_3\end{aligned}$$

So the predicted mean for the third trial equals  $\kappa_{In} + \lambda_3 (\kappa_{Sh})$ . The pattern coefficient in this expression,  $\lambda_3$ , indicates the proportion of the average improvement over the first two trials that must be added to the initial mean in order to generate the predicted mean for the third trial. If  $\lambda_3 = 1.50$ , for example, then the average for trial 3 is predicted to equal the initial mean plus 1.5 times the average improvement over the first two trials. Exercise 1 asks you to generate and interpret in symbolic form the predicted mean for trial 4.

7. The error terms of the adjacent trials are assumed to covary in Figure 15.3. Other patterns are possible, including no error covariances (the errors are independent) or the specification of additional error covariances, such as  $E_1 \curvearrowleft E_3$ . The capability to explicitly model the error covariance structure is an advantage of SEM over more traditional statistical techniques. For example, repeated measures ANOVA assumes that the error variances are equal and independent, which is unlikely for learning trial data.<sup>4</sup> The technique of MANOVA (multivariate ANOVA) makes less restrictive assumptions about error variances (e.g., they can covary), but both ANOVA and MANOVA treat individual differences in growth trajectories as error variance. In contrast, a goal of analyzing an LGM in SEM is to explicitly model these differences.

The change model in Figure 15.3 has 20 free parameters. These include (1) two factor means (of Initial and Shape); (2) eight variances (of two factors and six error terms); (3) six covariances (one between the factors and five error covariances); and (4) four pattern coefficients (trials 3–6 for the Shape factor). With six observed variables, there are  $6(9)/2$ , or 27 observations (6 variances, 15 unique covariances, and 6 means) available to estimate the model, so  $df_M = 7$ . I fitted the change model in Figure 15.3 to the data in Table 15.3 for the six trials with the ML method of Amos (Amos Development Corporation, 1983–2013). The Amos program calculates variances with  $N$  in the denominator, not  $N - 1$ . You can download the Amos input and output files for this analysis from this book’s website. Computer files for the same analysis in EQS, lavaan, LISREL, Mplus, and Stata are also available.

Analysis of the change model in Figure 15.3 converged to an admissible solution. Values of selected fit statistics are reported in the first row of Table 15.4. The Amos program does not automatically print the SRMR. The exact-fit hypothesis is rejected at the .05 level,  $\chi^2_M(7) = 16.991$ ,  $p = .017$ , and the upper bound of 90% confidence interval based on the RMSEA, .122, is a troubling result. The value of the Bentler CFI, .995, does not suggest an obvious problem.<sup>5</sup> Things are clearer in local fit testing: None of standardized residuals or standardized mean residuals are statistically significant; no absolute correlation residual exceeds .10; and the predicted means are close to the observed values. Given these results, the change model in Figure 15.3 is retained. Reported in the second row of Table 15.4 are values of selected fit statistics for a change model with no correlated errors. The fit of this constrained model is poor, so the hypothesis of independent errors for these data is rejected.

Estimates of free parameters for the change model in Figure 15.3 are reported in Table 15.5. Unstandardized direct effects of the constant on the latent growth factors are estimated means. The mean of the Initial factor is 11.763, which is close to the observed average score on the first trial (11.77; see Table 15.3). The two means just stated are not

<sup>4</sup>Such data are also likely to violate the sphericity assumption in repeated measures ANOVA that the variances of all population pairwise difference scores are equal.

<sup>5</sup>The default independence model in Amos requires the observed variables to be uncorrelated, but their means and variances are not constrained.

**TABLE 15.4. Values of Selected Fit Statistics for Latent Growth Models of Performance on a Computerized Air Traffic Controller Task**

Model	$\chi^2_M$	$df_M$	p	$\chi^2_D$	$df_D$	p	RMSEA (90% CI)	CFI
Change model	16.991	7	.017	—	—	—	.076 (.030–.122)	.995
Change model, uncorrelated errors	105.824	12	< .001	88.833	5	< .001	.177 (.147–.209)	.949
Prediction model	27.333	11	.004	—	—	—	.077 (.041–.114)	.991

Note. CI, confidence interval. All results were computed by Amos.

identical because one is for an observed variable and the other is for a latent variable. The estimated mean of the Shape factor is 9.597, which indicates that the average increase from trial 1 to trial 2 is about 9.60 points. The statistical significance of the variances of the latent growth factors may be of interest. For example, the estimated variances of the Initial and Shape factors are, respectively, 50.430 and 12.929, and each is significant at the .01 level (see Table 15.5). These results say that participants are not homogeneous in either their initial performance or their rate of improvement.

The estimated covariance between the latent growth factors is -7.221. Given its standard error of 3.705 (Table 15.5), this covariance falls just short of statistical significance at the .05 level ( $z = 1.95$ ,  $p = .051$ ). The corresponding estimated factor correlation is -.283. This result says that *higher* levels of performance on trial 1 predict *lower* rates of improvement after trial 1; that is, those who start off with better initial performance show relatively less improvement over later trials, and vice versa. Other results in the table concern error variances of trials 1–6 and their error covariances. The change model explains relatively high proportions of the total variance of the indicators. The  $R^2$  values range from .699 for trial 2 to .905 for trial 5, or a majority of the variance for each trial. The range of estimated error correlations is .057–.434. As expected in a learning study with multiple trials, the error correlations are positive.

Freely estimated pattern coefficients for indicators of the Shape factor (trials 3–6) are also reported in Table 15.5. The unstandardized coefficient for trial 3 is 1.639. This result indicates that the average score on trial 3 is predicted to equal the sum of the mean for the Initial factor and 1.639 times the improvement from trial 1 to trial 2 (i.e., the mean of the Shape factor). For trial 4, the unstandardized pattern coefficient is 2.015 (see the table). It has a similar interpretation except that the proportion of the improvement in performance over the first and second trials is 2.015. Exercise 2 asks you to interpret the unstandardized pattern coefficients reported in Table 15.5 for trials 5 and 6 as indicators of the Shape factor.

Listed next are the six unstandardized pattern coefficients for, respectively, trials 1–6 as indicators of the Shape factor (Figure 15.3, Table 15.5):

0            1.0            1.639            2.015            2.171            2.323

**TABLE 15.5. Maximum Likelihood Parameter Estimates for a Latent Growth Model of Change in Performance on a Computerized Air Traffic Controller Task**

Parameter	Unstandardized	SE	Standardized
<u>Mean structure</u>			
<u>Latent growth factor means</u>			
△ → Initial	11.763	.482	0
△ → Shape	9.597	.346	0
<u>Covariance structure</u>			
<u>Pattern coefficients</u>			
Shape → Trial 1	0	—	0
Shape → Trial 2	1.000	—	.430
Shape → Trial 3	1.639	.042	.672
Shape → Trial 4	2.015	.057	.798
Shape → Trial 5	2.171	.061	.830
Shape → Trial 6	2.323	.066	.840
<u>Variances and covariances</u>			
<u>Latent growth factors</u>			
Initial	50.430	8.276	1.000
Shape	12.929	2.175	1.000
Initial ↗ Shape	-7.221	3.705	-.283
<u>Indicator error terms</u>			
Trial 1	7.484	6.881	.129
Trial 2	21.096	2.825	.301
Trial 3	15.359	1.823	.200
Trial 4	8.553	1.707	.104
Trial 5	8.439	2.137	.095
Trial 6	12.260	2.337	.124
$E_1 \curvearrowright E_2$	5.457	3.504	.434
$E_2 \curvearrowright E_3$	6.259	1.185	.348
$E_3 \curvearrowright E_4$	3.939	1.282	.344
$E_4 \curvearrowright E_5$	.488	.906	.057
$E_5 \curvearrowright E_6$	3.881	1.899	.382

Note. All unstandardized pattern coefficients for the Initial factor are fixed to 1.0. Standardized estimates for error terms are proportions of unexplained variance calculated as the ratio of the error variance over the predicted variance for the corresponding trial. All results were computed by Amos. The standardized solution is completely standardized.

These coefficients are represented on the Y-axis in Figure 15.4 over the six trials represented on the X-axis. Note that the shape of the curve described by the unstandardized pattern coefficients in Figure 15.4 matches that described by the original means over the trials in Figure 15.2. Specifically, both curves have a positive linear trend and a negative quadratic trend. The unstandardized pattern coefficients thus optimize the fit of the Shape factor to the data.

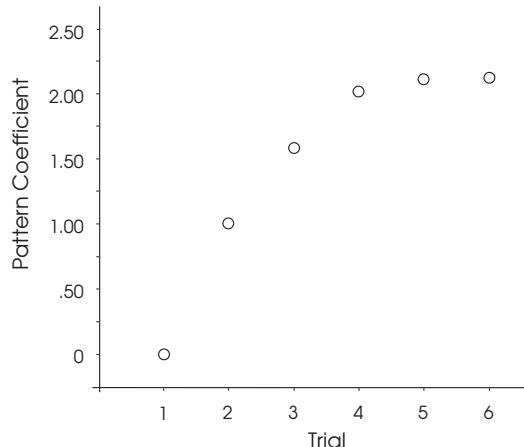
The observed means for trials 1 and 2 are, respectively, 11.77 and 21.39 (Table 15.3). Given the results in Table 15.5, the predicted means for these trials are calculated as total effects of the constant as follows:

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 1} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 1} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 1} \\ &= 11.763 (1.0) + 9.597 (0) = 11.763\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 2} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 2} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 2} \\ &= 11.763 (1.0) + 9.597 (1.0) = 21.360\end{aligned}$$

Each predicted mean just calculated is similar to its observed counterpart. For trial 3,  $\hat{\lambda}_3 = 1.639$ , so its implied mean is calculated as follows:

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 3} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 3} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 3} \\ &= 11.763 (1.0) + 9.597 (1.639) = 27.492\end{aligned}$$



**FIGURE 15.4.** Unstandardized pattern coefficients for indicators of the Shape factor in the change model. Coefficients for trials 1–2 are fixed parameters that equal, respectively, 0 and 1.0. Values for trials 3–6 are freely estimated.

which is close to the observed mean for this trial, 27.50 (Table 15.3). Exercise 3 asks you to calculate predicted means for trials 4–6. Completing this exercise should convince you that the change model closely reproduces the observed means. We are ready for the next step.

## Predicting Change

Predictors are added to a change model by regressing the latent growth factors on them, as illustrated in Figure 15.5. Because the Initial and Shape factors are endogenous in the prediction model, they have disturbances, and the disturbance covariance represents their association controlling for ability. Unstandardized coefficients for direct effects of ability are designated in the figure as  $\gamma_{In}$  and  $\gamma_{Sh}$ . Because ability is exogenous, the direct effect of the constant equals its mean, which is designated as  $\kappa_{Ab}$  in the figure even though this variable is not latent.<sup>6</sup>

Direct effects of the constant on latent growth factors are not means in the prediction model; instead, they are intercepts from their regressions on ability. These intercepts are designated in Figure 15.5 as  $\alpha_{In}$  and  $\alpha_{Sh}$ , where the lowercase Greek letter alpha in LISREL symbolism represents the intercept for a latent endogenous variable. Means of latent growth factors are still total effects of the constant, which are outlined next for the prediction model:

$$\begin{aligned} \text{Total effect of } \Delta \text{ on Initial} &= \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} + \\ &\quad \Delta \rightarrow \text{Initial} \\ &= \kappa_{Ab} (\gamma_{In}) + \alpha_{In} \end{aligned} \tag{15.1}$$

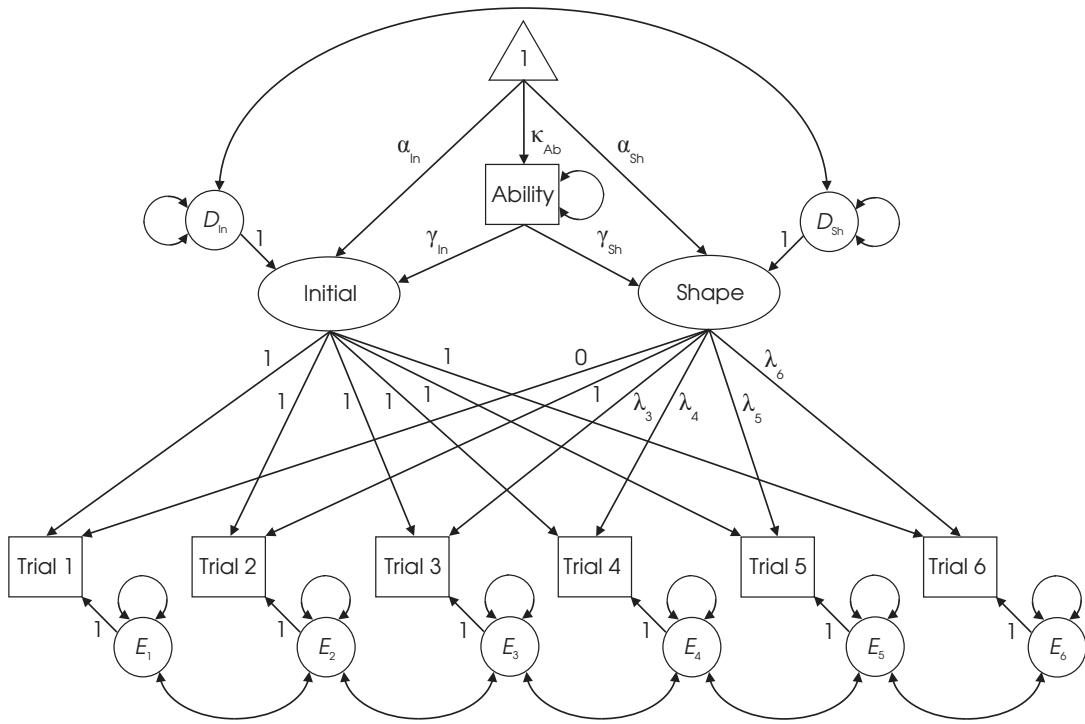
$$\begin{aligned} \text{Total effect of } \Delta \text{ on Shape} &= \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} + \\ &\quad \Delta \rightarrow \text{Shape} \\ &= \kappa_{Ab} (\gamma_{Sh}) + \alpha_{Sh} \end{aligned}$$

Exercise 4 asks you to interpret the expression just listed for the Initial factor. The measurement part of the prediction model in Figure 15.5 is identical to the measurement part of the change model in Figure 15.3, including error covariances between adjacent trials and freely estimated pattern coefficients for trials 3–6 as indicators of the empirical Shape factor.

With seven variables, a total of  $7(10)/2 = 35$  observations are available to estimate the 24 free parameters of the prediction model in Figure 15.5. These include (1) three direct effects of the constant (the mean of ability, two latent growth factor intercepts); (2) nine variances (of ability, two factor disturbances, and six error terms); (3) six covariances (one disturbance covariance and five error covariances); (4) two direct effects of

---

<sup>6</sup>It is not strictly required to include the predictor in the mean structure, but I do so here because the mean of the predictor is part of the input data for the analysis of the model in Figure 15.5.



**FIGURE 15.5.** Latent growth model for prediction of change in performance on a computerized air traffic controller task.  $\kappa$ , mean of the exogenous ability variable;  $\alpha$ , intercepts of endogenous latent growth factors;  $\gamma$ , path coefficients for direct effects of ability;  $\lambda$ , freely estimated pattern coefficients for the Shape factor.

ability on the latent growth factors; and (5) four pattern coefficients (trials 3–6 for the Shape factor), so  $df_M = 11$ .

I fitted the prediction model in Figure 15.5 to the data in Table 15.3 with Amos. Estimation with the ML method converged to an admissible solution, and values of selected fit statistics are reported in Table 15.4. The prediction model fails the chi-square test,  $\chi^2_M(11) = 27.333$ ,  $p = .004$ , and the upper bound of the 90% confidence interval based on the RMSEA, .114, exceeds .10, so the poor-fit hypothesis cannot be rejected. But at the level of local fit, the model seems fine (e.g., all absolute correlation residuals  $< .10$ ; predicted means are similar to the observed means), so the prediction model in Figure 15.5 is retained.

Reported in Table 15.6 are the ML parameter estimates for just the mean structure and the structural part of covariance structure for the prediction model in Figure 15.5. Estimates for the measurement part of the prediction model, including pattern coefficients and error variances and covariances, are similar to their counterparts in the change model (Table 15.5) and are not described next. The unstandardized coefficient for the direct effect of the constant on ability equals the observed mean of that variable, .70 (see Table 15.3). In contrast, the unstandardized direct effects of the constant on the

latent growth factors, Initial (11.287) and Shape (9.608), are intercepts. The means of the latent growth factors can be derived as total effects of the constant. Using results from Table 15.6, we can estimate these means as follows:

$$\text{Total effect of } \Delta \text{ on Initial} = .700 (.678) + 11.287 = 11.762$$

$$\text{Total effect of } \Delta \text{ on Shape} = .700 (-.096) + 9.608 = 9.541$$

where .678 and  $-.096$  are the unstandardized direct effects of ability on, respectively, the Initial factor and the Shape factor. These predicted means are similar to those estimated for the change model (Table 15.5), and they are interpreted the same way, too.

Both unstandardized coefficients (.678,  $-.096$ ) for direct effects of ability on the latent growth factors are statistically significant at the .05 level. More informative are values of the standardized coefficients. For every increase in ability of one standard deviation, the level of the Initial factor is expected to increase by .541 standard devia-

**TABLE 15.6. Selected Maximum Likelihood Parameter Estimates for a Latent Growth Model of Prediction of Change in Performance on a Computerized Air Traffic Controller Task**

Parameter	Unstandardized	SE	Standardized
<u>Mean structure</u>			
Predictor mean			
$\Delta \rightarrow$ Ability	.700	.355	0
<u>Latent growth factor intercepts</u>			
$\Delta \rightarrow$ Initial	11.287	.421	0
$\Delta \rightarrow$ Shape	9.608	.351	0
<u>Covariance structure</u>			
<u>Direct effects</u>			
Ability $\rightarrow$ Initial	.678	.074	.541
Ability $\rightarrow$ Shape	-.096	.043	-.153
<u>Disturbance variances and covariance</u>			
Initial	34.841	6.904	.707
Shape	12.145	1.996	.977
$D_{In} \curvearrowleft D_{Sh}$	-4.526	3.209	-.220

Note. Unstandardized pattern coefficient for the Shape factor over the six trials are, respectively, 0, 1.0, 1.647, 2.027, 2.185, and 2.338. Standardized estimates for disturbance terms are proportions of unexplained variance calculated as the ratio of the disturbance variance over the predicted variance for the corresponding latent growth factor. All results were computed by Amos. The standardized solution is completely standardized.

tions (Table 15.6), so higher general ability predicts better trial 1 performance. Ability explains about 29.3% of the total variation in initial performance ( $R^2 = 1 - .707 = .293$ ). But given a one standard deviation increase in ability, the level of the Shape factor is predicted to decrease by .153 standard deviations. This means that participants of higher ability show less improvement relatively over subsequent trials, and vice versa. The percentage of explained variation for the Shape factor is about 2.3% ( $R^2 = 1 - .977 = .023$ ). Thus, ability better predicts initial performance than the rate of subsequent improvement. The estimated disturbance correlation is  $-.220$ , which says that higher initial levels of performance are associated with lower rates of subsequent improvement after controlling for ability. This result parallels a similar one for the change model (Table 15.5).

The unstandardized total effect of the constant on each trial in Figure 15.5 is the predicted mean for that trial. Each of these total effects is comprised of four indirect pathways. Given the results from Table 15.6 listed next

$$\hat{\kappa}_{Ab} = .700, \hat{\alpha}_{In} = 11.287, \hat{\alpha}_{Sh} = 9.608 \\ \hat{\gamma}_{In} = .678, \hat{\gamma}_{Sh} = -.096, \text{ and } \hat{\lambda}_3 = 1.647$$

we can calculate the predicted mean for the third trial as follows:

$$\begin{aligned} \text{Total effect of } \Delta \text{ on Trial 3} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 3} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} \rightarrow \text{Trial 3} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 3} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} \rightarrow \text{Trial 3} \\ &= 11.287 (1.0) + .700 (.678) (1.0) + \\ &\quad 9.608 (1.647) + .700 (-.096) (1.647) \\ &= 27.475 \end{aligned}$$

The observed mean for trial 3 is 27.50, which is close to the predicted mean. Exercise 5 asks you to calculate the predicted means for trials 1–2 and 4–6 for the prediction model. Many SEM computer tools automatically generated predicted means, which is convenient in a larger model.

## COMPARISON WITH A POLYNOMIAL GROWTH MODEL

Presented in Figure 15.6 is a polynomial change model for the six trials of the air traffic controller task.<sup>7</sup> In contrast to the change model in Figure 15.3, which we analyzed in nonlinear curve fitting, the change model in Figure 15.6 has separate Linear and Qua-

<sup>7</sup>The ability predictor would be added to the polynomial change model in Figure 15.6 in the usual way (i.e., all latent growth factors are regressed on ability) in order to specify a polynomial prediction of change model, but this possibility is not elaborated further.

dratic latent growth factors. All unstandardized pattern coefficients are fixed to equal the constants displayed in the figure. For example, because the pattern coefficients for the Linear factor, or

$$(0, 1, 2, 3, 4, 5)$$

are evenly spaced over the trials, they specify a linear trend. Coefficients for the Quadratic factor, or

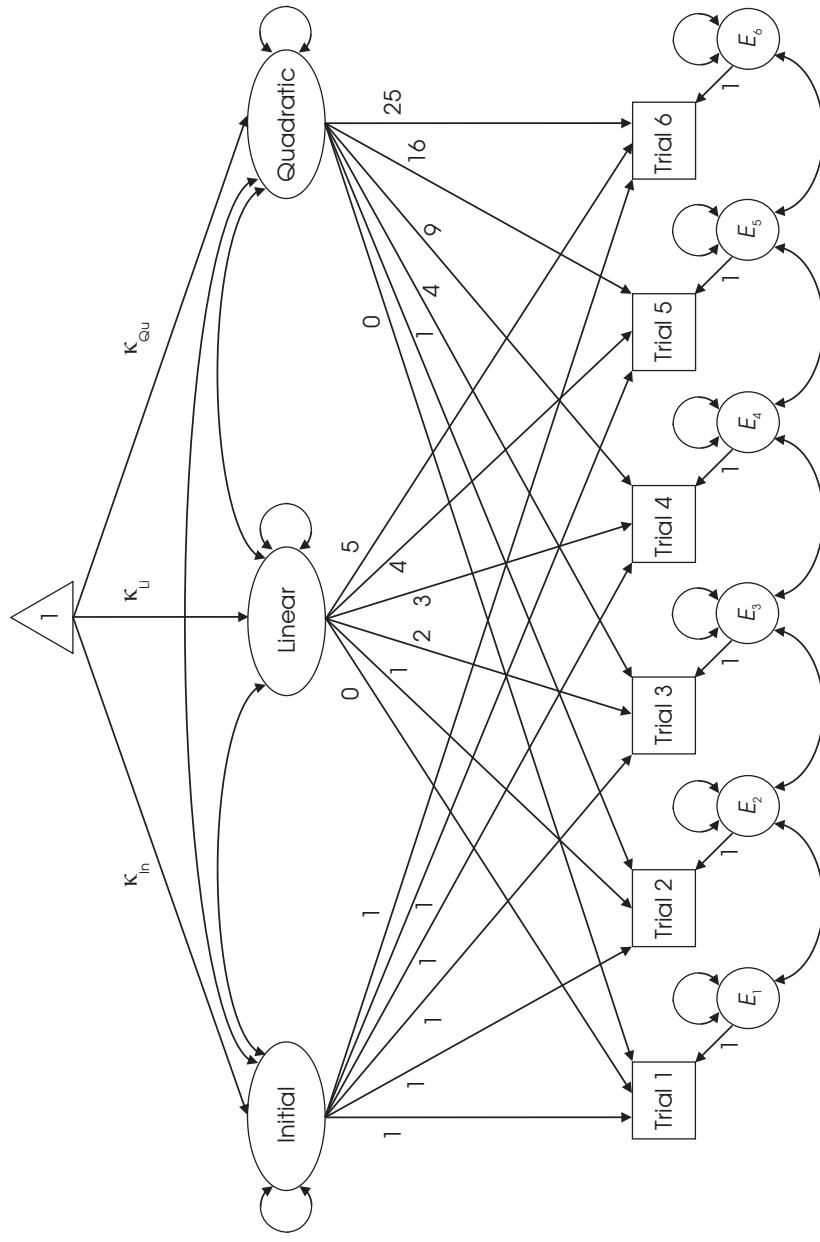
$$(0, 1, 4, 9, 16, 25)$$

are just the squares of the corresponding weights for the Linear factor. The Initial factor is defined in the usual way (all pattern coefficients are fixed to 1.0). The three latent growth factors are assumed to covary.

There are three direct effects of the constant in Figure 15.6, one on each of the exogenous latent growth factors. Each direct effect just mentioned is also a total effect that equals the mean of the corresponding factor, represented with the symbols  $\kappa_{In}$ ,  $\kappa_{Li}$ , and  $\kappa_{Qu}$  in the figure. The interpretation of  $\kappa_{In}$  is unchanged (the mean performance at trial 1). The value of  $\kappa_{Li}$  is the average linear slope over all six trials, and a positive versus negative value of  $\kappa_{Li}$  indicates, respectively, a positive versus negative trend. The average departure from linearity over trials (i.e., the curve is accelerating or decelerating) is indicated by the value of  $\kappa_{Qu}$ , and positive versus negative values of  $\kappa_{Qu}$  indicate, respectively, a positive or negative quadratic trend. Within the limits of identification, it may be possible to specify additional curvilinear trends, such as by adding a Cubic factor to Figure 15.6 that covaries with the Initial, Linear, and Quadratic factors and where the pattern coefficients for the Cubic factor are the cubed values of those for the Linear factor. However, it is rarely necessary to estimate curvilinear trends higher than a cubic one for most behavioral science data.

An advantage of the polynomial change model in Figure 15.6 is that it can be tested hierarchically, first with just the Initial factor (i.e., no change is predicted), then next with just the Initial and Linear factors (i.e., only linear change is expected), and finally with all three latent growth factors (i.e., linear and quadratic trends are estimated). At each step, the improvement in fit due to adding a latent growth factor to the model should be appreciable; otherwise, a more parsimonious model with fewer latent growth factors would be preferred. A drawback is increased complexity relative to a change model of the kind analyzed in nonlinear curve fitting, which has a latent shape factor that “captures” all linear or nonlinear trends in the data (see Figure 15.3). Another challenge is that at least four waves of data are needed in order to identify a polynomial model like the one in Figure 15.6. Just three waves of data are required in nonlinear curve fitting.

If we fit the polynomial change model in Figure 15.6 to the data in Table 15.3 for just the six trials, the computer will issue a warning or error message that the variance-covariance matrix of the three latent growth factors, Initial, Linear, and Quadratic, is



**FIGURE 15.6.** A polynomial change model with separate linear and quadratic latent growth factors. All unstandardized pattern coefficients are fixed parameters.  $\kappa_i$  means of exogenous latent growth factors.

nonpositive definite. Thus, the solution is inadmissible owing to an invalid parameter matrix. (Try this analysis on your own.) Perhaps the sample size ( $N = 250$ ) is just too small for the relatively complex model in Figure 15.6 or some sample correlations are so high (e.g.,  $r = .91$  for trials 4 and 5; Table 15.3) that extreme collinearity causes the analysis of the more complex polynomial growth model to fail. Polynomial factors can be very highly correlated, and the data for this example may not be precise enough to adequately distinguish between the Linear and Quadratic factors. Another potential difficulty is that the variances of polynomial factors beyond a linear slope factor are often relatively small, which can restrict the statistical power of tests for curvilinear trajectories. For this example, it is good to know about nonlinear curve fitting, which provides a simpler—and workable—alternative (Figure 15.3) to a polynomial growth model with separate Linear and Quadratic factors (Figure 15.6).

## EXTENSIONS OF LATENT GROWTH MODELS

The basic framework for univariate growth curve modeling in a single sample can be extended in many ways. For example, the ability variable in Figure 15.5 is a **time-invariant predictor** that was measured only once. It is also possible to include a **time-varying predictor** that is a repeated measures variable, typically assessed at the same intervals as the indicators of the latent growth factors. It is also no special problem to analyze latent growth models with binary or ordinal repeated measures. Such outcomes are simply specified as indicators of latent response variables, which are then regressed on the latent growth factors. Thresholds associate categorical outcomes with their latent response variable counterparts (e.g., Figure 13.6); see Masyn, Petras, and Liu (2014) for more details.

The ability predictor in Figure 15.5 is represented as an error-free exogenous variable. Given an a priori estimate of its error variance, one could use the method described earlier to control for measurement error in single indicators (e.g., Figure 10.7). The same method could be used to explicitly control for measurement error in the repeated measures variable (trials): Specify each trial as the single indicator of a latent variable that is regressed on the latent growth factor. Another way to control for measurement error on the predictor side is to use multiple indicators of an exogenous factor specified to predict the latent growth factors; that is, the prediction side of a growth model can be fully latent, too. It is straightforward in SEM to estimate indirect effects among predictors of latent growth factors. Even another variation is to analyze a model where the repeated measures variables are all latent and each is measured with multiple indicators. Such a model is called a **curve-of-factors model** because the latent repeated measures variables are regressed on second-order latent growth factors in order to explain or predict change in the former; see Park and Schutz (2005).

An alternative to a polynomial model like the one shown in Figure 15.6 is a **piecewise latent trajectory model** that captures nonlinear change at particular times or phases within a longer period of observation. Each phase or “piece” has its own latent

growth factor, and the end of one piece is the beginning of the next piece, if the phases are in sequential order. The specification of a piecewise growth model may be especially well suited for the study of phenomena that occur during certain critical periods or in treatment outcome studies where symptom reduction may be most apparent during phases of active treatment; see Flora (2008) for more information.

It is possible to analyze multivariate latent growth models of change across two or more domains. If these domains are measured at the same points in time, then the model reflects a **parallel growth process** (Kaplan, 2009). For example, George (2006) analyzed data from a longitudinal annual survey of attitudes about the utility of science in everyday life. The model analyzed was one of **cross-domain change** in which the within-domain latent growth factors were allowed to covary across the domains. The results indicated that while students' interest in science courses steadily declines during the middle school and high school years, their views of science utility generally increase over the same time. Higher initial interest in science classes predicted a more positive attitude about science utility, and changes in one domain covaried positively with changes in the other domain. Furthermore, initial levels of these domains were negatively associated with changes in the other domain. For example, students who in Grade 7 expressed more positive attitudes about science utility exhibited a more gradual decline in their interest in science classes from Grades 7 to 11.

A **factor-of-curves model** is a kind of parallel growth process model where separate first-order latent growth factors explain the trajectories within each domain (e.g., interest in science, attitudes about science utility), and second-order latent growth factors explain the associations among the first-order latent growth factors. This type of model can be more parsimonious compared with one where all associations among first-order latent growth factors are specified as unanalyzed (i.e., they covary), but it can be difficult to specify the model when two or more repeated measures outcomes show different patterns of change (Park & Schutz, 2005). See Bishop, Geiser, and Cole (2015) for more information.

Bollen and Curran (2004) describe **autoregressive latent trajectory** (ALT) models in which the indicators of latent growth factors are allowed to have direct and indirect effects on each other over time. An **autoregressive structure** is one in which past values of a variable are used to predict future values of that same variable (i.e., a Markov chain); that is, lagged (prior) variables are specified as predictors of later measurements on the same variable. For example, the specification for the empirical example presented next

Trial 1 → Trial 2 → Trial 3 → Trial 4 → Trial 5 → Trial 6

illustrates an autoregressive model of lag one where performance on the previous trial affects performance on the current trial. The structure just listed implies indirect effects, too, such as the effect of trial 1 on trial 3 through trial 2, the mediator. The autoregressive part of an ALT model is basically a Markov chain that controls for direct effects among the indicators not explained by the latent growth factors.

There are many statistical techniques for analyzing autoregressive structures, including the **autoregressive integrative moving average** (ARIMA) model, which uses shifts and lags in a time series to uncover patterns, such as seasonal trends or various kinds of intervention effects. In contrast, a standard growth model does not incorporate lagged effects among the indicators. Instead, indicators are assumed to be spuriously associated (locally independent) because of common causes, in this case the latent growth factors and correlated errors (e.g., Figure 15.5). Bollen and Curran (2004) argue that this assumption is unrealistic in certain cases. The basic logic of an ALT model can be extended to analysis of panel data from one or more repeated measures variables. A drawback is that integration of the autoregressive and latent growth curve parts of an ALT model rests on the strong assumption that neither part is misspecified, but non-linearity in the latent growth component can violate this requirement (Voelkle, 2008). Little (2013, pp. 271–273) notes that the simultaneous estimation of the autoregressive and latent growth curve parts of an ALT model with the same covariance information in the data conflates two questions that cannot be answered in the same model.

## SUMMARY

Means are estimated in SEM when the computer regresses exogenous or endogenous variables on a constant that equals 1.0. The parameters of a mean structure include the means of the exogenous variables and the intercepts of the endogenous variables. Means of endogenous variables are not model parameters, but predicted means of these variables, calculated as total effects of the constant, can be compared with the observed means. In order to be identified, the number of parameters in a mean structure cannot exceed the number of observed means. A latent growth model for longitudinal data is basically an SR model with a mean structure. Each repeated measures variable is specified as an indicator of at least two different latent growth factors. One of these factors represents initial status (origin), and the pattern coefficients for all indicators are fixed to 1.0. In nonlinear curve fitting, the only other latent growth factor estimates the shape of the growth trajectory. In a polynomial model, there is a separate slope factor for each individual trend, beginning with a linear change factor, followed next by a quadratic change factor, and so on. Latent growth factors are usually assumed to covary, which allows for the possibility that initial status is related to the rate of subsequent change. The next chapter introduces multiple-samples analysis in SEM. It also deals with the topic of measurement invariance in CFA.

## LEARN MORE

Books by Bollen and Curran (2006) and Preacher et al. (2008) are excellent resources for latent growth modeling. Park and Schutz (2005) describe latent growth models in exercise and sport, but the examples are meaningful for readers in other areas.

- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.
- Park, I., & Schutz, R. W. (2005). An introduction to latent growth models: Analysis of repeated measures physical performance data. *Research Quarterly for Exercise and Sport*, 76, 176–192.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

### EXERCISES

1. Apply the tracing rule to Figure 15.3 and interpret the symbolic expression for the predicted mean of trial 4.
2. Interpret the unstandardized pattern coefficients in Table 15.5 for trials 5 and 6 as indicators of the Shape factor in Figure 15.3.
3. Given the information in Tables 15.3 and 15.5, calculate the predicted means for trials 4–6.
4. Interpret Equation 15.1 for the Initial factor in Figure 15.5.
5. Calculate predicted means for trials 1–2 and 4–6 in Figure 15.5 (see Table 15.6).

## 16

# Multiple-Samples Analysis and Measurement Invariance

---

This chapter concerns SEM analyses where a model is simultaneously fitted to data matrices from two or more samples. The hypothesis that certain unstandardized parameters differ over populations is tested by imposing cross-group equality constraints on estimates of those parameters. If model fit while imposing the equality constraints is poor, there is evidence for group differences on those parameters. Multiple-samples CFA is concerned with whether a set of indicators measures the same constructs with equal precision over different samples, a property that is called measurement invariance. Testing for measurement invariance in CFA is explained and illustrated with two examples, first for a model with continuous indicators and then for a model with categorical indicators (Likert-scale items).

---

## RATIONALE OF MULTIPLE-SAMPLES SEM

The main question in a **multiple-samples SEM analysis** is whether values of model parameters of substantive interest vary appreciably across different samples.<sup>1</sup> Another way to express this question is in terms of interaction: Does sample membership moderate the relations specified in the model? If so, there is evidence for a group  $\times$  model interaction, where the computer must be allowed to derive separate estimates of some parameters in each sample in order for the model to have acceptable fit over all samples.

Perhaps the simplest way to address these questions is to separately estimate the same model within each of two or more different samples and then compare the unstan-

---

<sup>1</sup>Statistical significance of differences over samples in estimates for the same parameter are irrelevant, if those differences are of trivial magnitude.

dardized solutions. Recall that unstandardized estimates instead of standardized estimates should generally be compared across different samples. For the same reason, covariance matrices with means (or the raw scores) should be analyzed when the model has both covariance and mean structures. If the unstandardized estimates for the same parameter are substantially different, then the populations from which the samples were selected may not be equal on that parameter.

More sophisticated comparisons are available by using an SEM computer tool to simultaneously fit a model to data from multiple samples. Through the specification of cross-group equality constraints, group differences on any individual parameter or set of parameters can be directly tested. An equality constraint forces the computer to derive equal unstandardized estimates of that parameter within all samples. The fit of the constrained model can then be compared with that of the model without equality constraints with the chi-square difference test. If the fit of the constrained model is much worse, we can conclude that the parameters may not be equal in the populations from which those samples were drawn. This assumes that the unconstrained model fits the data from all samples reasonably well. Remember that estimates constrained to be equal in the unstandardized solution will typically be unequal in the standardized solution.

Note that LISREL can optionally print up to *four* different standardized solutions in multiple-samples analyses. The *within-groups standardized solution* and the *within-groups completely standardized solution* are both derived by standardizing the within-groups covariance matrices except that only the factors are standardized in the former versus all variables in the latter. In the LISREL *common metric standardized solution*, the weighted average factor covariance matrix over samples is a correlation matrix, but all variables are so scaled in the *common metric completely standardized solution*. Common metric standardized results may be more directly comparable over samples than within-groups standardized results, but the unstandardized estimates are still preferred for this purpose. Check the documentation of your SEM computer tool to see how it calculates the standardized solution in a multiple-samples analysis.

Any type of structural equation model—path, SR, or CFA model—can be tested across multiple samples. For example, Molina, Alegría, and Mahalingam (2013) analyzed a path model of self-rated health among Latin American adults. They found that the magnitude of indirect effects of psychological distress on self-rated health depended on both gender and ethnicity. Specifically, Cuban men's self-rated health was relatively more affected by psychological distress compared with other ethnic or gender groups. Benyamin, Ein-Dor, Ginzburg, and Solomon (2009) analyzed a latent growth model of self-reported health among Israeli veterans of the 1982 Lebanon War who were tested at 1, 2, 3, and 20 years after the conflict. The veterans were divided into two groups based on whether a combat stress reaction (CSR) was exhibited during the war. Benyamin et al. (2009) found that although the health trajectory of CSR veterans was positive over time (they improved), their levels of reported health remained lower than that of control veterans.

The goal of a multiple-samples CFA is to establish whether a set of indicators has the property of measurement invariance (defined in the next section) over two or more samples. Invariance testing in CFA involves (1) specifying a measurement model with a mean structure; (2) constraining to equality over samples the unstandardized parameters for pattern coefficients, intercepts, thresholds, or error variances and covariances; and (3) simultaneously fitting a series of models so constrained to the data in each multiple group.

## MEASUREMENT INVARIANCE

**Measurement invariance** concerns whether scores from the operationalization of a construct have the same meaning under different conditions (Meade & Lautenschlager, 2004). These different conditions could involve time of measurement, test administration methods, or populations. The *absence* of this property says that findings of differences between persons cannot be unambiguously isolated from differences owing to time, methods, or group membership (Horn & McArdle, 1992); that is, there is no clear basis for drawing inferences from the scores.

Stability in measurement parameters over time for the same population corresponds to **longitudinal measurement invariance**. It is well known in longitudinal research that administering the same test over multiple occasions does not guarantee that the same construct is assessed at each occasion. But if the results of a longitudinal factor analysis indicate that the same factor structure fits the data from multiple occasions equally well, then measurement over time may be invariant (Little, 2013, chap. 5). Similarly, finding that the same factor structure holds over different ways of giving the same test, such as Internet versus paper-and-pencil format, suggests that measurement is invariant over administration methods (Whitaker & McKinney, 2007). Because the logic of invariance testing over populations is basically the same as for invariance testing over time or methods, only invariance testing over populations is described next.

In this discussion we assume continuous indicators, but later we will deal with special considerations for categorical indicators. There are four basic kinds of measurement invariance. Unfortunately, different authors do not always use the same label to refer to the same invariance type. To avoid confusion, next I use the simplest versions of these labels, but I give alternative names in footnotes. Types of measurement invariance include (1) configural invariance, (2) weak invariance, (3) strong invariance, and (4) strict invariance (Wu, Li, & Zumbo, 2007). These types represent increasingly restrictive hypotheses about invariance, and each successive hypothesis requires more evidence than the preceding hypothesis. This means that measurement invariance is not an all-or-none property, and the task of the researcher is to decide about the particular level of measurement invariance (including none) supported by the data.

**Configural invariance** is the least restrictive level. It is tested by specifying the same CFA model in each group. In this model, both the number of factors and the cor-

respondence between factors and indicators are the same, but all parameters are freely estimated in each group. If this model is not consistent with the data, then measurement invariance does not hold at any basic level.<sup>2</sup> Otherwise, the hypothesis of configural invariance is retained, which says that the same factors are manifested in somewhat different ways in each group. These different ways refer to unequal pattern coefficients, intercepts, or error variances for some indicators over the groups. If factor scores were calculated assuming just configural invariance, a different weighting scheme would be needed for each group.

The hypothesis of **weak invariance** assumes configural invariance. It also requires equality of the unstandardized pattern coefficients.<sup>3</sup> This hypothesis is tested by (1) imposing an equality constraint over groups on the unstandardized coefficient of each indicator, and (2) comparing with the chi-square difference test the configural invariance model and the weak invariance model. If the fit of the weak invariance model is not appreciably worse than that of the invariance model, the more restrictive weak invariance hypothesis is retained. This outcome says that the constructs are manifested in the same way in each group; specifically, the slopes from regressing the indicators on their respective factors are equal across groups. Factor scores would be calculated using the same weighting scheme in all groups.

Gregorich (2006) offered two alternative accounts for rejecting the weak variance hypothesis. One possibility is that the factors—or at least a subset of items that correspond to those factors—have different meanings over groups. For example, psychological symptoms of depression are emphasized in Western samples, but somatic symptoms are more apparent in Asian samples (Ryder et al., 2008). Another possibility is an **extreme response style** (ERS) that affects the response variability. Low ERS is the tendency to avoid endorsing the most extreme response options (e.g., *never, always*) in favor of middling options (e.g., *sometimes*), and may be found in cultural groups that emphasize modesty or humility. High ERS is just the opposite: the most extreme response options are favored, which is a pattern that may be found in populations where decisiveness or firmness is encouraged. Cheung and Rensvold (2000) offer suggestions for detecting ERS and other response styles in multiple-samples CFA.

Retaining the weak invariance hypothesis justifies the formal comparison (e.g., with a significance test) of estimated factor variances or covariances over groups. This is because common variance from each indicator is allocated to the corresponding factor in the same way over groups. Any group differences in residual variation cannot contaminate group differences in common factor variation. But because the indicators are affected both by factors and sources of unique (residual) variation, weak invariance by itself does not support the formal comparison of observed variances or covariances over groups (Gregorich, 2006).

<sup>2</sup>**Dimensional invariance** requires only that indicators depend on the same number of factors across groups, but does not also require the same pattern of factor-indicator correspondence. The latter aspect of dimensional invariance is incompatible with the concept of measurement invariance.

<sup>3</sup>Weak invariance is also known as pattern invariance or metric invariance.

**Strong invariance** assumes weak invariance.<sup>4</sup> It also requires equal unstandardized intercepts over the groups. The intercept estimates the score on an indicator, given a true score of zero on the corresponding factor. Equality of intercepts says that different groups use the response scale of that indicator in the same way; that is, a person from one group and a person from a different group with the same level on the factor should obtain the same score on the indicator. The strong invariance hypothesis is supported if the fit of the model with equality-constrained unstandardized pattern coefficients *and* intercepts is not appreciably worse than that of the model with equality-constrained pattern coefficients only (i.e., weak invariance).

Rejection of strong invariance suggests the presence of a **differential additive (acquiescence) response style**, which refers to systematic influences unrelated to the factors that decrease or increase the overall level of responding on an indicator in a particular population (Cheung & Rensvold, 2000). Suppose that men and women differ in their willingness to acknowledge certain health problems described in a questionnaire. This difference could affect the observed means but not response variation, that is, the effect is additive. Other sources of differential additive response styles include cultural differences, cohort effects, or procedural differences in data collection. An example of these procedural differences is when patients are weighed in their street clothes in one clinic but wearing examination gowns in a different clinic (Gregorich, 2006). In this case, the constant that is added to true body weight depends on where patients were tested. It also contaminates estimates of mean weight differences over the two clinics. Note that if a response style affects all indicators, then invariance testing will not detect this pattern. Instead, the estimates of the construct will be influenced by response styles that are uniform over all indicators.

The pattern that an indicator has appreciably unequal pattern coefficients or intercepts over groups is **differential item functioning** (DIF). In this case, a person's score on the indicator, given his or her true score on the corresponding factor, will depend on his or her membership in a particular group, which violates the idea of measurement invariance. A goal in multiple-samples CFA is thus to locate the indicator(s) responsible for rejection of the hypothesis of weak invariance or strong invariance. In test development, items so flagged are candidates for deletion or alteration (e.g., change wording so it does not disadvantage members of a particular group), if the source of DIF is not part of the construct being tested (Karami, 2012).

Strong invariance guarantees that (1) group differences in estimated factor means will be unbiased and (2) group differences in indicator means or estimated factor scores will be directly related to the factor means and will not be distorted by differential additive response bias (Gregorich, 2006). The factors have a common meaning over groups and any constant effects on the indicators are canceled out when observed means are compared over groups. Thus, strong invariance is the minimal level required for meaningful interpretation of group mean contrasts. Some significance tests for mean contrasts, such as the standard *t* test, assume equal population variances. This assump-

---

<sup>4</sup>Strong invariance is also known as scalar invariance.

tion can be tested in CFA by imposing equality constraints on factor variances and then comparing the relative fits of the model so constrained with the unrestricted model. Rejection of the hypothesis of homoscedasticity in CFA would rule out the use of the standard  $t$  test, but other versions, such as the Welch–James test described in Appendix 16.A, do not rely on this assumption. Note that measurement invariance of the indicators requires neither equality of factor variances nor equality of factor means.

**Strict invariance** is the highest level of measurement variance. It assumes strong invariance *and* equality in error variances and covariances across the groups. This means that the indicators measure the same factors in each group with the same degrees of precision. Deshon (2004) and Wu et al. (2007) argue that residual invariance is required in order to claim that the factors are measured *identically* across groups. This is because unmodeled systematic effects on observed scores can be confounded with differences in pattern coefficients or intercepts over groups, which can result in specification error. Little (2013, p. 143) argued that because unique (residual) indicator variance reflects both random measurement error and specific variance that is systematic (e.g., Figure 9.2), the hypothesis of strict invariance is actually that the *sum* of these two components is equal for each indicator over the groups. Although it may be reasonable to assume that specific variance is invariant over groups, the expectation that the random error component would also be invariant may be less reasonable. Some researchers consider strict invariance as required for formal comparisons of observed differences in variances or covariances over groups. In this case, such differences are not confounded with group differences in error variances or covariances. Little (2013) cautions against enforcing this requirement because if the sum of the random and systematic parts of unique variance is not exactly equal, the amount of misfit due to equality-constrained residual variances must contaminate estimates elsewhere in the model (propagation of specification error).

Too many researchers fail to evaluate all aspects of measurement invariance. Vandenberg and Lance (2000) reviewed a total of 67 published measurement invariance studies and found that weak invariance (equal pattern coefficients) was evaluated in virtually all studies (99%). But residual variance and covariance equality was evaluated in 49% of the studies, and intercept equality was investigated in only 12% of reviewed works. Although researchers seem to appreciate the potential impact of pattern coefficient inequality on measurement invariance, many are not as aware of the implications of intercept inequality or residual inequality.

## TESTING STRATEGY AND RELATED ISSUES

The hierarchy of measurement invariance hypotheses corresponds to a model trimming strategy in which an initial unconstrained model (configural invariance) is gradually restricted by adding cross-group equality constraints in a sequence that corresponds to weak invariance, strong invariance, and then strict invariance. Stark, Chernyshenko, and Drasgow (2006) refer to this strategy as the **free baseline approach**. Failure to

retain the invariance hypothesis at a particular step means that even more restricted models are not considered.

It is also possible to test for measurement invariance through model building where constraints on an initially restricted model, such as one represented by the strict invariance hypothesis (equal pattern coefficients, intercepts, and residuals), are gradually released (e.g., next test strong invariance by allowing error variances and covariances to be freely estimated in each group). This method is the **constrained baseline approach** (Stark et al., 2006). The problem with this method is that it may not be clear which particular set of cross-group equality constraints—those for pattern coefficients, intercepts, or residuals—should be released, if the initial fully-constrained model is rejected. If theory is not specific, the choice may be arbitrary. Ideally, model trimming versus model building for the same data would each select the same final measurement model, but this is not guaranteed.

Because invariance testing involves the evaluation of a series of hierarchically related models, decisions made about models tested earlier affect results found for models tested later. Sometimes these decisions can have unintended consequences. For example, what seems like a small respecification in an earlier model can affect the choice of the final model at the end of the analysis. This is why Millsap and Olivera-Aguilar (2012) noted that effective use of computer tools for invariance testing relies heavily on the experience and judgment of the researcher.

Cheung and Rensvold (2002) remind us that the chi-square difference test in very large samples could be statistically significant even though the absolute differences in parameter estimates are of trivial magnitude. Thus, the outcome of the chi-square difference test could indicate lack of measurement invariance when the imposition of cross-group equality constraints makes relatively little difference in fit. One way to detect this outcome is to compare the unstandardized solutions across the groups. Another is to inspect changes in approximate fit indexes, but there are few guidelines for doing so in invariance testing. Cheung and Rensvold (2002) found in computer simulation studies that values of the Bentler CFI were relatively unaffected by model characteristics such as the number of indicators per factor. They suggested that changes in CFI values less than or equal to .01, or  $\Delta\text{CFI} \leq .01$ , indicate that the stricter invariance hypothesis should *not* be rejected. A second approximate-fit index that performed relatively well in the same simulation study is McDonald's (1989) noncentrality index (NCI).<sup>5</sup>

Meade, Johnson, and Braddy (2008) studied the performance of several approximate-fit indexes in generated data with different levels of measurement invariance. This included different factor structures (no invariance), different pattern coefficients, or different intercepts across two groups. In very large samples, such as 6,000 cases per group, the  $\chi^2_D$  statistic indicated lack of measurement invariance most of the time when there were just slight differences in model parameters over groups. In contrast, values of approximate fit indexes were generally less affected by group size and also by the

---

<sup>5</sup>NCI =  $\exp[-.5(\chi^2_M - df_M)/N]$  where exp is the exponential function  $e^x$  and e is the natural base, approximately 2.7183. The range of the NCI is 0–1.0 where 1.0 indicates the best fit. Mulaik (2009b) notes that values of the NCI tend to drop off quickly with small decreases in fit.

number of factors or indicators than the chi-square difference test in large samples. The CFI was among the best performing approximate fit indexes along with the McDonald NCI. Based on their results, Meade et al. (2008) suggested that changes in CFI values less than or equal to .002, or  $\Delta\text{CFI} \leq .002$ , may indicate deviations from measurement invariance that are functionally trivial when group sizes are very large.

Simulation results by Chen (2007) raise doubts about the accuracy of thresholds for invariance testing with approximate fit indexes. For example, in cases where the group sizes are both small ( $n < 300$ ) and unequal, the decision rules  $\Delta\text{CFI} \leq .005$  and  $\Delta\text{RMSEA} \leq .010$  were reasonably accurate in detecting the lack of invariance. But more stringent rules were needed when the group size was larger ( $n > 300$ ) and equal and the pattern of invariance was mixed. The latter means that there are at least two invariant parameters, each of which is from a different category, pattern coefficients, intercepts, or residual variances. In this case, the thresholds  $\Delta\text{CFI} \leq .010$  and  $\Delta\text{RMSEA} \leq .015$  worked reasonably well. Of the CFI and RMSEA, the latter tended to over-reject true invariant models when the sample size is smaller. Results by Cheung and Rensvold (2002), Meade et al. (2008), and Chen (2007) indicate that researchers working with very large samples should look more to approximate fit indexes than significance tests to establish measurement invariance, but no single rule works in all situations. Sass, Schmitt, and Marsh (2014) caution that thresholds based for approximate fit indexes may not perform especially well with misspecified models when the data are ordinal.

Power of significance tests in multiple-samples CFA are often low unless the group sizes are reasonably large. For example, Meade and Bauer (2007) found in simulation studies that power to detect population differences in pattern coefficients was only about .40 when the group size was  $n = 100$ . In contrast, power was generally high when the group size was  $n = 400$ , but power estimates for an intermediate group size of  $n = 200$  were quite variable. This is because power in invariance testing is affected not just by group size but also by model and data characteristics, such as magnitude of factor intercorrelations. Meade and Bauer's (2007) results did not indicate a general rule about a ratio of group size to the number of indicators that could ensure adequate power to detect the absence of measurement in variance when the group sizes are not large.

Byrne, Shavelson, and Muthén (1989) described **partial measurement invariance** as an intermediate state of invariance. For example, weak invariance assumes cross-group equality of each unstandardized pattern coefficient. If some, but not all, pattern coefficients are invariant, then only partial weak invariance holds. In this case, testing for intercept equality could still be performed because noninvariant pattern coefficients are freely estimated in each group, which controls for these differences. It is difficult that there are no clear guidelines for determining the degree of partial invariance in all situations that would be acceptable for concluding that the indicators measure approximately the same things over groups (Steenkamp & Baumgartner, 1998). Suppose that a single pattern coefficient out of 20 altogether is not invariant. There may be little harm in testing for the invariance of other parameters, such as intercepts or residuals, given that 19/20, or 95% of the pattern coefficients are equal across groups. But as more and more pattern coefficients are found to be unequal across the groups, such as  $> 10$  (i.e.,

the majority out of 20), there should be less and less confidence that the indicators define the same factors in each group.

In a simulation study, Steinmetz (2011) found negligible effects of a minority of one or two indicators with unequal pattern coefficients on the accuracy of group differences in average factor scores (composites) as estimators of true differences in factor means. But inequality of even one intercept had a substantial impact on the composites; specifically, an unequal intercept can lead to spurious composite differences between groups with equal factor means. The same inequality can also result in attenuated differences on composites for groups with unequal factor means. These results suggest that full invariance of the intercepts may be required for correct interpretation of group differences on observed variables; otherwise, group mean differences on the indicators may be confounded with differences in intercepts and factor means.

Asparouhov and Muthén (2014) describe the **alignment method** for estimating group-specific factor means and variances while assuming only approximate measurement invariance. It is intended to be more convenient to apply when analyzing data from many groups than traditional multiple-samples CFA as described to this point. The final (aligned) model in this method has the same fit to the data as the configural invariance model. In the alignment method, the degree of noninvariance in pattern coefficients or intercepts between each pair of groups is estimated with a loss function. The computer then uses Bayesian estimation to re-weight the estimates in the configural invariance model in order to minimize the degree of noninvariance in the aligned model. Measures of the amount of nonvariance are calculated for each measurement parameter. The alignment method relies heavily on significance testing in order to “discover” the optimal aligned model that minimizes noninvariance. Also, it does not yet work for models with complex indicators that depend on two or more factors. The alignment method is implemented in Mplus (Muthén & Muthén, 1998–2014).

Analyses for measurement invariance are often complex because multiple CFA models each of which represents different hypotheses about invariance, ranging from configural invariance at the lowest level through strict invariance at the highest level, are typically analyzed. Thus, you should thoroughly annotate your syntax files so that later on you can more easily remember exactly what you did at each step; that is, document your decisions at every point.

Presented next are two examples of testing for measurement invariance. Both examples feature the simultaneous analysis of a single-factor model with five indicators over groups from two different populations. The indicators are treated as continuous variables in the first example, and as ordinal in the second example (Likert-scale items). Special issues in each type of analysis, such as identification requirements, are emphasized. A limitation is that it is not possible in these secondary analyses to specify *a priori* hypotheses about which particular parameters may be invariant or not invariant over the two different populations in these analyses. Thus, there is an unavoidable exploratory bent to the analyses described next, as models are respecified to better fit the data. In actual primary analyses, the better practice is to respecify models according to theoretical predictions or results of previous empirical studies.

## EXAMPLE WITH CONTINUOUS INDICATORS

Dillman Carpentier et al. (2008) administered measures of parent–child conflict within a sample of 450 Hispanic adolescents. The total sample was divided into groups based on the primary language spoken at home, English ( $n_1 = 193$ ) or Spanish ( $n_2 = 257$ ). Correlations, standard deviations, and means reported by Millsap and Olivera-Aguilar (2012) for the five indicators within each group are presented in Table 16.1.

The initial CFA model with a mean structure is presented in Figure 16.1. The direct effect of the constant on the exogenous conflict factor estimates its mean, which is designated as  $\kappa$  in the figure. Direct effects of the constant on the indicators are intercepts in their regressions on the common factor. Intercepts are designated in the figure with the symbol  $\nu$  (lowercase Greek letter nu).<sup>6</sup> Total effects of the constant generate predicted indicator means. For example, the total effect of the constant on indicator  $X_1$  is

$$\begin{aligned} \text{Total effect of } \Delta \text{ on } X_1 &= \Delta \rightarrow \text{Conflict} \rightarrow X_1 + \\ &\quad \Delta \rightarrow X_1 \\ &= \kappa(\lambda_1) + \nu_1 \end{aligned} \tag{16.1}$$

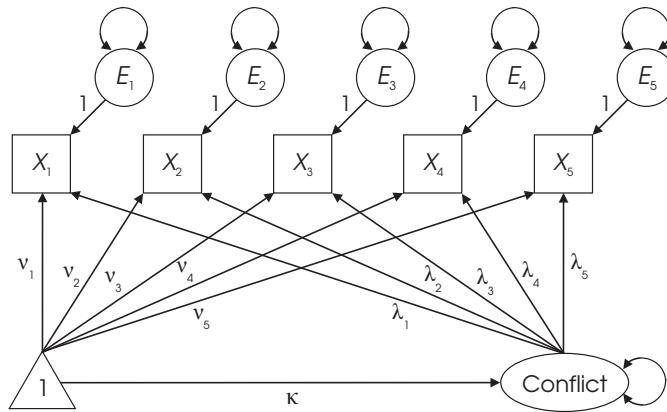
In words, the predicted mean of  $X_1$  in a particular group is a function of the group factor mean ( $\kappa$ ), the unstandardized coefficient when  $X_1$  is regressed on the common factor ( $\lambda_1$ ), and the intercept in this analysis ( $\nu_1$ ). Exercise 1 asks you to explain why Equation 16.1 implies that strict invariance is required in order to meaningfully compare the mean of  $X_1$  (or that of any other indicator) over groups.

**TABLE 16.1. Input Data (Correlations, Standard Deviations, Means) for a Single-Factor Model of Parent-Child Conflict Analyzed across English and Spanish Samples**

Indicator	English						
	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	M	SD
$X_1$	—	.381	.599	.416	.601	2.280	.887
$X_2$	.227	—	.393	.445	.404	1.518	.925
$X_3$	.400	.322	—	.476	.661	2.052	.972
$X_4$	.324	.330	.354	—	.519	1.689	1.014
$X_5$	.473	.370	.486	.540	—	1.684	.901
Spanish	M	2.113	1.175	1.708	1.366	1.319	
	SD	1.034	.597	.836	.785	.701	

Note. These data are from Millsap and Olivera-Aguilar (2012). English speaking (above diagonal),  $n_1 = 193$ ; Spanish speaking (below diagonal),  $n_2 = 257$ .

<sup>6</sup>In Mplus, the parameter matrix for intercepts of continuous indicators is labeled Nu. The same parameter matrix in LISREL is referred to as Tau-X.



**FIGURE 16.1.** Initial single-factor model of parent–child conflict analyzed over English and Spanish samples.  $\kappa$ , mean of the exogenous conflict factor;  $v$ , indicator intercepts;  $\lambda$ , indicator pattern coefficients.

For two reasons, the model in Figure 16.1 is not identified if it were estimated in a single sample. First, the mean structure would be underidentified: There are five observations (means of indicators  $X_1$ – $X_5$ ), but the mean structure has six parameters, including the factor mean ( $\kappa$ ) and five intercepts ( $v_1$ – $v_5$ ). Second, the latent conflict variable requires a scale. There are three basic options to scale the factor and identify the mean structure in Figure 16.1 in a multiple-samples CFA (Little, Slegers, & Card, 2006). These methods, outlined next, usually generate the same value of the model chi-square and all other fit statistics for the same model and data:

1. In the **reference group method**, one group is selected as the reference group, and the factor mean and variance are fixed to equal, respectively, 0 and 1.0 in this group. While imposing cross-group equality constraints on the unstandardized pattern coefficients and intercepts, factor means and variances are freely estimated in the other groups, each scaled relative to the values in the reference group. For example, the factor mean in other groups is estimated as the average difference across the set of indicators weighted by the pattern coefficients of those indicators. The factor variance in other groups is estimated as the proportional differences in the common variance of the indicators explained by the factor (Little et al., 2006).

2. A second option is the **marker variable method**, where the unstandardized pattern coefficient of one indicator per factor is fixed to 1.0 and its intercept is fixed to 0. The same indicator should be selected as the marker (reference) variable in each group. This method scales each factor in a metric related to that of the explained variances of the corresponding marker variable. The remaining pattern coefficients and intercepts are estimated but equated across the groups. Means and variances of the factors are freely estimated in all groups in this method.

A drawback of both methods is that the selection of a reference group or marker variable may be arbitrary. The choice should not affect the overall fit of the model most of the time, but Millsap (2001) described some exceptions for very unrestricted models. Both the reference group and marker variable methods require an invariance assumption but of a different type for each method. Because factor variances are fixed to 1.0 in the reference group, it must be assumed that the true variances in this group are equal across all factors. Because the pattern coefficient and intercept of the marker variable are fixed to, respectively, 1.0 and 0 in all groups, it is assumed that these coefficients are invariant over groups. This is because these fixed coefficients are excluded from tests of measurement invariance, so it must be assumed *a priori* that the unstandardized regression of the marker variable on its factor is identical in each group. If the researcher inadvertently selects a marker variable that is not invariant over groups, the results may be distorted. The third option, described next, avoids both of these problems.

3. The effects coding method is an option when all indicators of the same factor have the same metric (Chapter 9). In multiple-samples CFA, this method involves (1) fixing the average unstandardized pattern coefficient to 1.0 and (2) fixing the average intercept of the same indicators to 0. For the model in Figure 16.1, the average pattern coefficient is fixed to 1.0, or

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}{5} = 1.0 \quad (16.2)$$

in each group. There are a total of five algebraically equivalent reexpressions of Equation 16.2, of which just one is listed next:

$$\lambda_1 = 5 - \lambda_2 - \lambda_3 - \lambda_4 - \lambda_5 \quad (16.3)$$

The average intercept for the indicators is fixed to equal zero in each group, or

$$\frac{v_1 + v_2 + v_3 + v_4 + v_5}{5} = 0 \quad (16.4)$$

and presented next is just one of five algebraically equivalent versions of Equation 16.4:

$$v_1 = 0 - v_2 - v_3 - v_4 - v_5 \quad (16.5)$$

Equations 16.3 and 16.5 represent linear constraints where the value of one parameter, such as  $\lambda_1$  in Equation 16.3, is specified as a linear combination of the other parameters, or  $\lambda_2 - \lambda_5$ .

No individual pattern coefficient is fixed to 1.0, and no individual intercept is fixed to 0 in the effects coding method. Instead, pattern coefficients in each group are freely

estimated as an optimal balance around 1.0. This says that factors in each group are assigned a metric related to that of the average explained variance in their indicators. The intercepts are estimated as optimal averages of zero in each group. Factor means are estimated in all groups as optimal weighted averages of their indicators, and factor variances are estimated as the average amount of variation explained in the set of indicators for each factor (Little et al., 2006).

I fitted the single-factor model with a mean structure in Figure 16.1 to the data in Table 16.1 in Mplus (Muthén & Muthén, 1998–2014) using ML estimation. The effects coding method was used to scale the common factor and identify the mean structure. Reported in Table 16.2 are values of selected fit statistics for a total of seven invariance models from converged and admissible solutions. Model 1 represents the configural invariance hypothesis. With five indicators, there are  $5(8)/2 = 20$  observations in each of two groups for a total of 40 observations altogether. For each group, the free parameters include a total of four pattern coefficients and four intercepts. The only other free parameters in each group are the mean and variance of the common factor and the error variances of the five indicators. This makes for 15 free parameters in each of the two groups for a grand total of 30, so  $df_M = 40 - 30 = 10$  for the configural invariance model.

Model 1 is not rejected by the chi-square test,  $\chi^2_M(10) = 15.363, p = .119$  (see Table 16.2). The contribution to the overall model chi-square is 8.991 (58.5%) from the English group and 6.372 (41.5%) from the Spanish group. The upper bound of the 90% confidence interval based on the RMSEA, .095, is close to indicating poor fit. Inspection of the residuals indicated that the configural invariance model closely reproduces the means in both groups and the covariances in just the Spanish group. There is one statistically significant standardized residual in the English group,  $z = 2.466, p = .014$ , for indicators  $X_2$  and  $X_4$ . The correlation residual for this pair of indicators is .119, so the model underpredicts the observed correlation by this amount. Given all these results, model 1 is rejected.

Model 2 includes an error covariance between indicators  $X_2$  and  $X_4$  in the English group only. This modified configural invariance model is not rejected by the chi-square test,  $\chi^2_M(9) = 7.850, p = .549$ , and contributions to the overall chi-square from the English and Spanish groups are, respectively, 1.478 (18.8%) and 6.372 (81.2%). Values of other fit statistics for model 2 do not look problematic (Table 16.2). No severe problems were apparent in the residuals for both groups (e.g., no standardized mean residuals are statistically significant, no absolute correlation residuals exceed .10). Also, the relative fit of model 2 is statistically better than that of model 1,  $\chi^2_D(1) = 7.513, p = .006$ . The modified configural invariance model with a single error covariance for the English group only is thus retained.

Note that the claim for configural invariance (model 2) is possibly confounded with unmodeled systematic effects that underlie the error covariance in the model for the English group only. Here, I remind you of the argument that error covariance homogeneity is required in order to claim that the five indicators really measure the same factor with the same degree of precision, so there is also doubt for these data about the hypothesis of weak invariance.

**TABLE 16.2. Values of Selected Fit Statistics for Measurement Invariance Hypotheses for a Single-Factor Model of Parent-Child Conflict Analyzed across English and Spanish Samples**

Invariance model	Retained?	$\chi^2_M$	$df_M$	Model comparison	$\chi^2_D$	$df_D$	RMSEA (90% CI)	CFI	SRMR
1. Configural	N	15.363	10	—	—	—	.049 [0, .095]	.991	.026
2. Configural <sup>a</sup>	Y	7.850	9	2 vs. 1	7.513**	1	0 [0, .068]	1.000	.018
3. Weak <sup>a</sup>	Y	13.361	13	3 vs. 2	5.511	4	.011 [0, .068]	.999	.036
4. Strong <sup>a</sup>	N	22.717	17	4 vs. 3	9.356	4	.039 [0, .076]	.991	.046
5. Partial strong <sup>a, b</sup>	Y	13.462	15	5 vs. 3	.101	2	0 [0, .057]	1.000	.036
6. Partial strong <sup>a, b, c</sup>	N	87.658**	20	6 vs. 5	74.196**	5	.123 [.097, .149]	.891	.122
7. Partial strong <sup>a, b, d</sup>	N	24.931	16	7 vs. 5	11.469**	1	.050 [0, .086]	.986	.111

Note. CI, confidence interval. All results were computed by Mplus.

<sup>a</sup> $E_2 \rightarrow E_4$ , English group only; <sup>b</sup>Intercepts for  $X_3-X_5$  are invariant; <sup>c</sup>Equal error variances; <sup>d</sup>Equal factor variances.

\* $p < .05$ ; \*\* $p < .01$ .

Each unstandardized pattern coefficient is constrained to equality across the groups in model 3, which tests the weak invariance hypothesis. Note that the linear constraint among the pattern coefficients defined by Equation 16.3 is no longer needed in both groups, so it was dropped from the Mplus syntax for the Spanish group. This model passes the chi-square test,  $\chi^2_M(13) = 13.361, p = .423$ , and the relative contributions to the overall chi-square in the English and Spanish groups are, respectively, 3.617 (27.0%) and 9.744 (73.0%). Based on the chi-square difference test, the relative fit of model 3 is not statistically worse than that of the less restricted model 2, and values of approximate fit indexes do not indicate an obvious problem with model 3 (Table 16.2). Two standardized residuals are statistically significant, one in each group, but both of the corresponding absolute correlation residuals are  $< .10$ ; therefore, model 3 is retained.

In model 4, the intercept of each indicator is constrained to equality over groups, which tests the strong invariance hypothesis. This model is not rejected by the chi-square test— $\chi^2_M(17) = 22.717, p = .159$ , English group contribution (7.780; 34.2%), Spanish group contribution (14.937; 65.8%)—and values of other fit statistics look fine (Table 16.2). In each group, the standardized mean residuals for indicators  $X_1$  and  $X_2$  are both significant. This pattern suggests that model 4 does not predict the group means for this pair of indicators as well as it does the group means on indicators  $X_3$ – $X_5$ . For these reasons, the weak invariance hypothesis is rejected.

In model 5, the equality constraints on the intercepts of indicators  $X_1$  and  $X_2$  are released (they are freely estimated in both groups). The exact-fit hypothesis is not rejected for this model— $\chi^2_M(15) = 13.462, p = .567$ , English group contribution (3.809; 28.3%), Spanish group contribution (9.653; 71.7%)—and its relative fit is not statistically worse than that of model 3 (weak invariance) in which all intercepts are freely estimated in each group,  $\chi^2_D(2) = .101, p = .951$ . Values of all other fit statistics are acceptable (Table 16.2), and the residuals indicate no obvious local fit problems. Model 6 with equality-constrained error variances has poor fit (e.g.,  $\chi^2_M(20) = 87.658, p < .001$ ; RMSEA = .123). Further testing for error variance homogeneity was not pursued. The relative fit of model 7 where just the factor variances are constrained to equality is also statistically worse than that of model 5 where the factor variances are freely estimated in each group ( $\chi^2_D(1) = 11.469, p < .001$ ), and there are many significant standardized residuals in both groups for model 7. These results suggest that model 5 is the final invariance model.

Overall, these findings support partial strong invariance: There is evidence for pattern coefficient homogeneity, but there is only partial intercept homogeneity (3 of 5 indicators) across the English and Spanish samples. There is evidence against error variance homogeneity and factor variance homogeneity. These five indicators seem to measure roughly the same factor but with unequal precision over the groups. These interpretations are clouded by the need to include an error covariance in the model for just the English sample. Thus, an unmeasured common cause of two indicators ( $X_2, X_4$ ) for this group may confound the hypothesis of partial strong invariance. My conclusions about these models and data are somewhat different from those of Millsap and Olivera-Anguilar (2012, p. 387), but major areas of agreement include full pattern coefficient homogeneity and partial intercept homogeneity.

Reported in Table 16.3 are estimates for the parameters of the covariance structure for model 5 in both groups. Note in the table that the unstandardized pattern coefficient of each indicator is equal across the groups (weak invariance). As expected, standardized pattern coefficients for the same indicator are not equal over the groups. Error variances and factor variances are freely estimated in both groups except for the error covariance between indicators  $X_2$  and  $X_4$ , which is estimated in just the English group. Listed in Table 16.4 are unstandardized estimates for the mean structure in both groups. A total of three of five indicator intercepts are equal across the groups ( $X_3-X_5$ ), but different estimates in each group are required for indicators  $X_1-X_2$ . This pattern suggests that mean comparisons of the groups on the latter pair of indicators may confound group differences in factor means with differences in intercepts. Exercise 2 asks you to calculate predicted means for each group, given the results in Tables 16.3 and 16.4.

**TABLE 16.3. Maximum Likelihood Parameter Estimates for the Covariance Structure of a Single-Factor Model of Parent-Child Conflict Analyzed across English and Spanish Samples**

Parameter	English			Spanish		
	Unst.	SE	St.	Unst.	SE	St.
<u>Unconstrained estimates</u>						
<u>Factor variance</u>						
Conflict	.412	.051	1.000	.235	.027	1.000
<u>Error variances and covariance</u>						
$X_1$	.356	.047	.433	.741	.071	.737
$X_2$	.651	.069	.797	.273	.026	.742
$X_3$	.354	.049	.396	.427	.045	.581
$X_4$	.648	.073	.617	.370	.038	.617
$X_5$	.249	.041	.306	.165	.027	.339
$E_2 \curvearrowright E_4$	.139	.052	.213	—	—	—
<u>Equality-constrained estimates</u>						
<u>Pattern coefficients</u>						
Conflict $\rightarrow X_1$	1.062	.059	.753	1.062	.059	.513
Conflict $\rightarrow X_2$	.635	.055	.451	.635	.055	.507
Conflict $\rightarrow X_3$	1.144	.053	.777	1.144	.053	.647
Conflict $\rightarrow X_4$	.988	.056	.619	.988	.056	.619
Conflict $\rightarrow X_5$	1.171	.049	.833	1.171	.049	.813

Note. These estimates are for model 5 (see Table 16.2). Unst., unstandardized; St., standardized. Standardized estimates for error variances are proportions of unexplained variance. All results were computed by Mplus. The standardized solution is STDYX.

**TABLE 16.4. Unstandardized Maximum Likelihood Parameter Estimates for the Mean Structure of a Single-Factor Model of Parent–Child Conflict Analyzed across English and Spanish Samples**

Parameter	English		Spanish	
	Estimate	SE	Estimate	SE
<u>Unconstrained estimates</u>				
<u>Factor mean</u>				
△ → Conflict	1.843	.051	1.532	.039
<u>Indicator intercepts</u>				
△ → $X_1$	.323	.115	.486	.105
△ → $X_2$	.346	.113	.202	.089
<u>Equality-constrained estimates</u>				
<u>Indicator intercepts</u>				
△ → $X_3$	−.051	.095	−.051	.095
△ → $X_4$	−.144	.097	−.144	.097
△ → $X_5$	−.474	.087	−.474	.087

*Note.* All results were computed by Mplus. The standardized solution is STDYX.

The estimated means on the conflict factor for the English and Spanish samples are, respectively, 1.843 and 1.532 (Table 16.4). Because factor variances are not homogeneous across the groups (Table 16.3), it is not appropriate to pool the within-groups variances. Instead, I used a latent variable form of the Welch–James test (Appendix 16.A), which does not assume homoscedasticity. Given estimates from Tables 16.3 and 16.4 summarized next,

$$\begin{aligned}\hat{\kappa}_{\text{Eng}} &= 1.843, \hat{\sigma}_{\text{Eng}}^2 = .412, n_1 = 193 \\ \hat{\kappa}_{\text{Spa}} &= 1.532, \hat{\sigma}_{\text{Spa}}^2 = .235, n_2 = 257\end{aligned}$$

the results of the Welch–James test are

$$t(344.33) = \frac{1.843 - 1.532}{.0552} = 5.63, p < .001$$

which say that the group factor means are statistically different. Exercise 3 asks you to reproduce these results for the Welch–James test.

An effect size estimate of the magnitude of the group factor mean difference between the English and Spanish samples is more informative than results of a significance test. Because the within-groups variances are heteroscedastic, two different standardized mean differences can be calculated, each based on the standard deviation in just one group (Kline, 2013a, chap. 5), as follows:

$$d_{\text{Eng}} = \frac{1.843 - 1.532}{\sqrt{.412}} = \frac{.311}{.642} = .48$$

$$d_{\text{Spa}} = \frac{1.843 - 1.532}{\sqrt{.235}} = \frac{.311}{.485} = .64$$

Thus, the group contrast on the factor (.311) is approximately .50–.60 standard deviations in magnitude. This is the amount by which the English group reports higher average levels on the latent conflict factor compared with the Spanish group relative to the variation on the factor in both groups. Choi, Fan, and Hancock (2009) describe additional two-group latent contrast effect size measures.

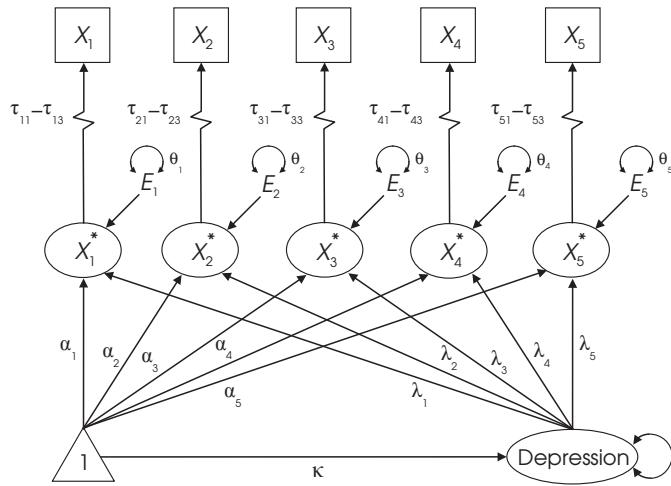
You can download from this book's website all Mplus data, syntax, and output files for this example. Computer files for analyses of the same models and data in LISREL (Scientific Software International, 2013) are also available. Because the SIMPLIS programming language in LISREL 9 does not support linear constraints among two or more parameters, I used LISREL's original, matrix-based programming language for this analysis.

## EXAMPLE WITH ORDINAL INDICATORS

The general definition of measurement invariance for ordinal indicators, such as Likert-scale items, can be stated as follows: For an individual item, it means that the probability of selecting a particular response option is the same across groups, given the same standing on the common factor that corresponds to that item. Measurement invariance is established if this property holds for all items (Millsap, 2011). But observed responses on ordinal indicators are only indirectly related to the common factors. This is because (1) each item is associated with a continuous latent response variable through a set of  $c - 1$  thresholds, where  $c$  is the number of item response categories; and (2) the latent response variables are the indicators of the common factors, not the observed variables (items) (e.g., Figure 13.6). Thus, there are special considerations for invariance testing when the data are ordinal.

Consider the single-factor CFA model in Figure 16.2. Its threshold structure associates each latent response variable  $X^*$  with an item  $X$ , each of which has a Likert scale with four response options (coded as 0, 1, 2, 3). The threshold parameters (three per item) are designated with the symbol  $\tau$ . The covariance structure in the figure specifies that the  $X^*$  variables are indicators of a common factor labeled depression. Parameters for the pattern coefficients, intercepts, and error variances of the  $X^*$  variables are designated, respectively, with the symbols  $\lambda$ ,  $\alpha$ , and  $\theta$ , and the mean of the exogenous depression factor is labeled  $\kappa$  in the figure.

Next we consider the minimum conditions required in order to identify the model of Figure 16.2 in a multiple-samples CFA with ordinal data (Millsap & Yun-Tein, 2004). These requirements assume that each item is polytomous with three or more response



**FIGURE 16.2.** Initial single-factor model of depression analyzed over white and African American samples shown with compact symbolism for error terms.  $\kappa$ , mean of the exogenous depression factor;  $\alpha$ , latent response variable intercepts;  $\lambda$ , latent response variable pattern coefficients;  $\theta$ , residual variances;  $\tau$ , item thresholds.

options and that each latent response variable is a simple indicator with a single pattern coefficient. These requirements assume theta parameterization, where the residual variance of each  $X^*$  variable is fixed to 1.0 in a group to be designated as the reference group. Error variances in other groups may be freely estimated. Theta parameterization in multiple-samples CFA with ordinal data is generally preferred over delta parameterization, where the error variances are not free parameters. This means that it is not possible to directly evaluate error homogeneity over groups in delta parameterization (see Newsom, 2015, chap. 2).

The model in Figure 16.2 is identified by imposing the constraints outlined next:

1. In the reference group, specify that the mean of the common factor is zero and standardize the variance of every residual. These specifications correspond in the figure to

$$\kappa = 0 \text{ and } \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 1.0$$

in just the reference group.

2. In every group, fix the direct effect of the constant on every  $X^*$  (i.e., the intercept,  $\alpha$ ) to zero. Next, select the same  $X^*$  across the groups as the marker (reference) variable and fix its unstandardized pattern coefficient to 1.0. The latter specification scales the common factor in the same way across the groups. It also assumes pattern coefficient homogeneity for the corresponding reference variable. Assuming that  $X_1^*$  is the reference variable, these requirements in Figure 16.2 correspond to the specifications that

$$\lambda_1 = 1.0 \text{ and } \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$$

in every group.

3. For every  $X^*$ , constrain one threshold parameter to equality across the groups. In addition, for each  $X^*$  that is a reference variable, constrain a second threshold parameter to equality over the groups. These constraints in Figure 16.2 correspond to

$$\begin{aligned}\tau_{11A} &= \tau_{11B}, \tau_{12A} = \tau_{12B} \\ \tau_{21A} &= \tau_{21B}, \tau_{31A} = \tau_{31B}, \tau_{41A} = \tau_{41B}, \tau_{51A} = \tau_{51B}\end{aligned}$$

where the last subscript for the parameters just listed indicates group A or group B in a two-samples analysis. Millsap and Yun-Tein (2004) describe identification requirements for models with binary items (e.g., true–false) or where some latent response variables are complex indicators with two or more pattern coefficients.

Parameters not mentioned in any of the three sets of identifying constraints are free parameters. This includes the variance of the common factor in Figure 16.2, which is freely estimated in each group. All pattern coefficients except those for reference variables ( $X_1^*$  in the figure) are also freely estimated in each group. In all groups except the reference group, the factor mean and residual variances are free parameters. The combination of identifying constraints and free parameters specifies the configural invariance model in the multiple-samples CFA with ordinal indicators for this example. If this configural invariance model is rejected as inconsistent with the data, then the measurement properties of the indicators are not invariant over groups at any level.

But if the model that specifies configural invariance is retained, then the weak invariance hypothesis can be tested. This is done by (1) constraining the unstandardized pattern coefficient for each latent response variable to equality over groups (coefficients for the reference variables are already equality-constrained to 1.0) and then (2) comparing the relative fit of the model so constrained with that of the configural invariance model. If the fit of the weak invariance model is not substantially worse than that of the configural invariance model, the weak invariance hypothesis is retained.

The strong invariance hypothesis assumes weak invariance. It is tested by imposing equality constraints on thresholds not already constrained for the sake of model identification (i.e., in the configural invariance and weak invariance models). This model is retained if its correspondence to the data is not appreciably worse than that of the weak invariance model. Equality of both pattern coefficients and thresholds are required in order to claim that ordinal indicators measure the same common factors, but perhaps with different degrees of precision. Assuming retention of the strong invariance model, the final model to be tested corresponds to the strict invariance hypothesis, which assumes equal error variances and covariances for the latent response variables over groups. Retention of the strict invariance hypothesis supports the claim that the indicators measure the same common factors in identical ways across all groups.

The hypothesis of measurement invariance at any level does not require homogeneity in factor variances or means over groups, so group differences in these parameters do not affect the measurement properties of ordinal indicators. It is also less relevant to formally compare common factor variances or means over groups in these types of analyses. This is because ordinal indicators are indirectly affected by the factor structure for the underlying latent response variables (see Figure 16.2), unlike continuous indicators (see Figure 16.1). Besides, values of means and variances for Likert-scale items are arbitrary because they depend on the particular numerical coding scheme for the response categories.

The data for the first of two groups compared in this empirical example were described in Chapter 13. Briefly, a sample of 2,004 white men completed five items about symptoms of depression from the Center for Epidemiologic Studies Depression (CES-D) scale (Radloff, 1977). The response categories are coded 0, 1, 2, or 3, where a higher score indicates greater frequency of the rated symptom over the prior week. This first group is designated as the reference group in this analysis. The second sample consists of 248 African American men who completed the same five items. Reported in Table 16.5 are the sample polychoric correlations and item thresholds for each group estimated in Mplus. All polychoric correlations are positive in both samples, and item thresholds do not appear to differ greatly over groups (see the table). Exercise 4 asks you to convert the thresholds in Table 16.5 to proportions of responses in each of the four response categories of item  $X_1$  for both groups.

**TABLE 16.5. Polychoric Correlations and Item Thresholds in White and African American Samples for Ratings of Depression Symptoms**

Indicator	$X_1^*$	$X_2^*$	$X_3^*$	$X_4^*$	$X_5^*$
<u>White (<math>n_1 = 2,004</math>)</u>					
$X_1^*$	—				
$X_2^*$	.437	—			
$X_3^*$	.471	.480	—		
$X_4^*$	.401	.418	.454	—	
$X_5^*$	.423	.489	.627	.465	—
<u>African American (<math>n_2 = 248</math>)</u>					
$X_1^*$	—				
$X_2^*$	.508	—			
$X_3^*$	.351	.373	—		
$X_4^*$	.305	.336	.398	—	
$X_5^*$	.464	.371	.531	.483	—

*Note.* Sample thresholds: white,  $X_1$ , .772, 1.420, 1.874;  $X_2$ , 1.044, 1.543, 1.874;  $X_3$ , .541, 1.152, 1.503;  $X_4$ , .288, 1.000, 1.500;  $X_5$ , .558, 1.252, 1.712; African American,  $X_1$ , .674, 1.487, 1.849;  $X_2$ , .753, 1.487, 1.622;  $X_3$ , .235, .726, .973;  $X_4$ , .361, 1.057, 1.374;  $X_5$ , .529, 1.233, 1.661. All results were computed by Mplus.

I fitted the single-factor CFA model with both mean and threshold structures in Figure 16.2 to the raw data file for this analysis using the WLSMV estimator and theta parameterization in Mplus. The initial model tested is that of configural invariance, which is specified by imposing the three sets of identifying constraints described earlier. Given  $p = 5$  items each with  $c = 4$  response categories and  $q = c - 1 = 3$  thresholds, there are a total of

$$2pq = 2(5)(3) = 30$$

independent response proportions over the two groups. In each group, there are  $p(p - 1)/2$  unique polychoric correlations (Table 16.5), so the total number across two groups is

$$p(p - 1) = 5(4) = 20$$

The total number of observations for this analysis over two groups is thus

$$2pq + p(p - 1) = 30 + 20 = 50$$

Freely estimated parameters include

1. one factor mean and  $p = 5$  residual variances in the African American group; and
2. two factor variances, a total of  $2(p - 1) = 2(4) = 8$  pattern coefficients, and a total of

$$2pq - p - 1 = 2(5)(3) - 5 - 1 = 30 - 6 = 24$$

thresholds in both groups.

The grand total for free parameters is  $1 + 5 + 2 + 8 + 24$ , or 40. This grand total also equals

$$2p(q + 1) = 2(5)(4) = 40$$

(Millsap & Yun-Tein, 2004). The degrees of freedom for the configural invariance model are thus  $df_M = 50 - 40 = 10$ .

Reported in Table 16.6 are values of fit statistics for the configural invariance model and a total of five additional invariance models. The model chi-squares for WLSMV estimation, designated as  $\chi^2_{SB}$  in the table, cannot be used for difference testing in the usual way, but specification of the “difftest” option in Mplus automatically generates correct values of scaled chi-square difference statistics,  $\hat{\chi}^2_D$ . Model 1 is rejected by the exact-fit test,  $\chi^2_{SB}(10) = 25.210$ ,  $p = .005$ , and the contribution to the overall model chi-square is 15.552 (61.7%) from the white group and 9.657 (38.3%) from the African American

**TABLE 16.6. Values of Selected Fit Statistics for Measurement Invariance Hypotheses for a Single-Factor Model of Depression Analyzed across White and African American Samples**

Invariance model	Retained?	$\chi^2_{\text{SB}}$	$df_M$	Model comparison	$\hat{\chi}^2_D$	$df_D$	RMSEA (90% CI)	CFI
1. Configural	N	25.210**	10	—	—	—	.037 [.019, .055]	.994
2. Configural <sup>a</sup>	Y	18.404*	9	2 vs. 1	5.066*	1	.030 [.009, .050]	.996
3. Weak <sup>a</sup>	Y	26.638*	13	3 vs. 2	8.656	4	.031 [.013, .047]	.995
4. Strong <sup>a</sup>	Y	35.633*	22	4 vs. 3	10.788	9	.023 [.007, .037]	.995
5. Strict <sup>a</sup>	N	71.607**	27	5 vs. 4	28.029**	5	.038 [.028, .049]	.983
6. Partial strict <sup>a,b</sup>	Y	39.555*	26	6 vs. 4	5.266	4	.022 [.004, .034]	.995

Note. CI, confidence interval. All results were computed by Mplus for theta parameterization.

<sup>a</sup> $E_1 \curvearrowright E_2$ , African American group only; <sup>b</sup>Error variance for  $X_3^*$  freely estimated, African American group only.

\* $p < .05$ ; \*\* $p < .01$ .

group. The correlation residual in the African American group for the pair of indicators  $X_1^*$  and  $X_2^*$  is .128, which is the amount by which model 1 underpredicts the sample polychoric correlation of .508 (Table 16.5) for these variables. This pair of indicators may share an omitted cause in the African American group only.

Given these results, an error covariance for the pair  $X_1^*$  and  $X_2^*$  was added to the original model in just the African American sample. The configural invariance model so respecified is model 2, which is rejected by the chi-square test,  $\chi^2_{\text{SB}}(9) = 18.404$ ,  $p = .031$  (Table 16.6). It is not surprising that most of the contribution to the overall model chi-square now comes from the white group (17.064, 92.7%) instead of the African American group (1.340, 7.3%). Values of absolute correlation residuals in both groups are  $< .10$ , and values of threshold residuals are also relatively small in both groups. Model 2 closely predicts the univariate proportions of responses in the four categories of items  $X_1$ – $X_5$ , and its fit is statistically better than that of model 1,  $\hat{\chi}^2_D(1) = 5.066$ ,  $p = .025$ . Values of other fit statistics are also reasonable, so model 2, the revised configural invariance model, is retained.

In model 3, the unstandardized pattern coefficients for indicators  $X_2^*$ – $X_4^*$  are constrained to equality across groups ( $X_1^*$  is the reference variable in both groups and its coefficient is already constrained to 1.0). This weak invariance model is rejected by the chi-square test,  $\chi^2_{\text{SB}}(13) = 26.638$ ,  $p = .014$ , and the contribution to the overall model chi-square is 15.666 (58.8%) from the white group and 10.972 (41.2%) from the African American group (Table 16.6). But no indications of severe misspecification are apparent in the residuals, and the fit of the weak invariance model is not statistically worse than that of the less restrictive configural invariance model,  $\hat{\chi}^2_D(4) = 8.656$ ,  $p = .070$ , so model 3 is retained.

In model 4, the hypothesis of strong invariance is specified by constraining to equality over groups all unstandardized thresholds not already so restricted in the original model. Model 4 is rejected by the chi-square test,  $\chi^2_{\text{SB}}(22) = 35.633$ ,  $p = .033$ , and the

contribution to the overall model chi-square is 15.468 (43.4%) from the white group and 20.165 (56.6%) from the African American group (Table 16.6). But its global fit is not statistically worse than that of model 3,  $\hat{\chi}_D^2(9) = 10.788$ ,  $p = .291$ , and inspection of the residuals suggests no severe local fit problems, so model 4 is retained.

Model 5 represents the strict invariance hypothesis, where the error variance for each of the five indicators is constrained to equal 1.0 across the groups. This model is rejected by the exact-fit test,  $\chi_{SB}^2(27) = 71.607$ ,  $p < .001$ , with contributions to the overall model chi-square of 19.026 (26.6%) in the white group and 52.582 (73.4%) in the African American group (Table 16.6). The fit of model 5 with equality-constrained residual variances is also statistically worse than that of model 4 without these constraints,  $\hat{\chi}_D^2(5) = 28.029$ ,  $p < .001$ . In both samples, the modification index for the error variance of the  $X_3^*$  indicator is significant (white,  $z = 8.284$ ,  $p < .001$ ; African American,  $z = 8.154$ ,  $p < .001$ ). Given all these results, the strict invariance hypothesis is rejected.

In model 6, the cross-group equality constraint for the error variance of the  $X_3^*$  indicator is released. This model of partial strict invariance is rejected by the chi-square test,  $\chi_{SB}^2(26) = 39.555$ ,  $p = .043$ , with contributions to the model chi-square of 14.542 (36.8%) in the white sample and 25.013 (63.2%) in the African American sample (Table 16.6). But because the fit of model 6 is not statistically inferior to that of model 4,  $\hat{\chi}_D^2(4) = 5.266$ ,  $p = .261$ , and no problems are apparent in the residuals, model 6 is retained.

To summarize, the five items just analyzed appear to measure a depression factor in roughly the same way for both white men and African American men (strong invariance, or equal pattern coefficients and thresholds), but with different degrees of precision for at least one indicator (partial strict invariance). The need to specify an error covariance for a pair of indicators in the African American sample only is additional evidence for error heterogeneity over the groups. The possibility that there are unmodeled systematic effects in the African American sample should temper enthusiasm for the interpretation of measurement invariance for these five items.

Parameter estimates for the final model (partial strict invariance, model 6) are reported in Tables 16.7–16.9. Results freely estimated in each group are presented in Table 16.7. The estimated factor mean of .086 in the African American sample is actually a contrast; that is, the mean for the African American men is higher by this amount than the mean for the white men. Factor variances and sizes for the white and African American samples are

$$\hat{\sigma}_W^2 = .577 \text{ and } \hat{\sigma}_{AA}^2 = .585 \\ n_W = 2,004 \text{ and } n_{AA} = 248$$

The variances are similar, so they are pooled in order to calculate a standardized mean difference:

$$d = \frac{.086}{\sqrt{\frac{2,003 (.577) + 247 (.585)}{2,250}}} = .11$$

**TABLE 16.7. Robust Weighted Least Squares Estimates of Unconstrained Parameters for a Single-Factor Model of Depression Analyzed across White and African American Samples**

Parameter	White			African American		
	Unst.	SE	St.	Unst.	SE	St.
<b>Depression factor</b>						
Variance	.577	.081	1.000	.585	.128	1.000
Mean	0	—	0	.086	.077	0
<b>Error variance and covariance</b>						
$X_3^*$	1.000	—	.372	3.498	.883	.672
$E_1 \curvearrowright E_2$	—	—	—	.192	.130	.192

Note. Unst., unstandardized; St., standardized. These estimates are for model 6 (Table 16.6). Standardized estimates for error variances are proportions of unexplained variance. All results were computed by Mplus for theta parameterization. The standardized solution is STDYX.

**TABLE 16.8. Robust Weighted Least Squares Estimates of Equality-Constrained Pattern Coefficients and Residual Variances for a Single-Factor Model of Depression Analyzed across White and African American Samples**

Parameter	White			African American		
	Unst.	SE	St.	Unst.	SE	St.
<b>Pattern coefficients</b>						
Depression → $X_1^*$	1.000	—	.605	1.000	—	.607
Depression → $X_2^*$	1.104	.104	.643	1.104	.104	.645
Depression → $X_3^*$	1.710	.159	.792	1.710	.159	.573
Depression → $X_4^*$	1.008	.086	.608	1.008	.086	.610
Depression → $X_5^*$	1.579	.144	.768	1.579	.144	.770
<b>Error variances</b>						
$X_1^*$	1.000	—	.634	1.000	—	.631
$X_2^*$	1.000	—	.587	1.000	—	.584
$X_4^*$	1.000	—	.630	1.000	—	.627
$X_5^*$	1.000	—	.410	1.000	—	.407

Note. Unst., unstandardized; St., standardized. These estimates are for model 6 (Table 16.6). Standardized estimates for error variances are proportions of unexplained variance. All results were computed by Mplus for theta parameterization. The standardized solution is STDYX.

**TABLE 16.9. Robust Weighted Least Squares Estimates  
of Equality-Constrained Thresholds for a Single-Factor Model  
of Depression Analyzed across White and African American Samples**

Item	Threshold	White			African American		
		Unst.	SE	St.	Unst.	SE	St.
$X_1$	1	.966	.045	.769	.966	.045	.767
	2	1.801	.062	1.434	1.801	.062	1.431
	3	2.360	.080	1.880	2.360	.080	1.875
$X_2$	1	1.326	.058	1.016	1.326	.058	1.013
	2	2.017	.075	1.546	2.017	.075	1.542
	3	2.412	.087	1.848	2.412	.087	1.843
$X_3$	1	.873	.062	.532	.873	.062	.383
	2	1.881	.093	1.148	1.881	.093	.824
	3	2.456	.113	1.498	2.456	.113	1.076
$X_4$	1	.382	.036	.303	.382	.036	.303
	2	1.276	.046	1.013	1.276	.046	1.011
	3	1.880	.058	1.493	1.880	.058	1.489
$X_5$	1	.882	.059	.565	.882	.059	.562
	2	1.967	.088	1.260	1.967	.088	1.255
	3	2.681	.113	1.717	2.681	.113	1.711

Note. Unst., unstandardized; St., standardized. These estimates are for model 6 (Table 16.6). All results were computed by Mplus for theta parameterization. The standardized solution is STDYX.

Thus, the mean of African American men on the depression factor is about .10 standard deviations higher than the mean of the white men. Other results in Table 16.7 are for parameters estimated in the African American sample only. These include the error variance for  $X_3^*$  (3.498) and the error covariance for the pair  $X_1^*$  and  $X_2^*$  (.192). Because the residual variances for this indicator pair are both 1.0 in the unstandardized solution, the error covariance just mentioned equals the error correlation.

Reported in Table 16.8 are equality-constrained estimates of pattern coefficients for all five indicators and residual variances for all indicators except  $X_3^*$ . Note in the table that the unstandardized estimates only are equal across the groups, which is expected. Given the results in Tables 16.7 and 16.8, Exercise 5 asks you to calculate  $R^2$  for each indicator in both groups. Estimates of equality-constrained thresholds for both samples are reported in Table 16.9. In the standardized solution of theta parameterization, thresholds are interpreted in the usual way; that is, as normal deviates each of which estimates the point on the corresponding latent response variable with a mean of 0 and a variance of 1.0 where the observed responses “cross over” from one response category to the next. In general, estimated item thresholds are similar across the groups, and the

standardized thresholds in Table 16.9 are similar to their observed counterparts in Table 16.5 for both white and African American men.

You can download the Mplus computer files for the analyses just described from this book's website. Also available are computer files for the same analysis in the lavaan package. Millsap and Yun-Tein (2004) describe differences between Mplus and LISREL in the analysis of CFA models for ordinal data over multiple samples when testing for measurement invariance. Hirschfeld and von Brachel (2014) demonstrate use of lavaan, semPlot, and semTools for measurement invariance testing in R.

## STRUCTURAL INVARIANCE

The analysis of an SR model with a mean structure over multiple samples follows the same basic rationale as for a CFA model with a mean structure. A notable difference is that direct effects of the constant on endogenous factors in SR models are intercepts for the regressions of those factors on their presumed causes, such as exogenous factors. It is total effects of the constant on the endogenous factors that estimate their means. Otherwise, testing for invariance in the measurement part of an SR model follows the same basic rationale as in CFA. It is also possible to test for **structural invariance**, or to determine whether the unstandardized coefficients for direct effects or disturbance variances and covariances are equal across groups. Such tests are conducted by imposing cross-group equality constraints on the corresponding parameter estimates for the structural model and then comparing the relative fits of the model so constrained with that of the model without equality constraints. If the fit of the constrained model is not much worse than that of the unrestricted model, there is evidence for structural invariance. Invariance over groups of the measurement part of an SR model should be conducted before testing for structural invariance. The path coefficients or disturbance terms may not be directly comparable if the same factors are not measured in the same way in each group.

## ALTERNATIVE STATISTICAL TECHNIQUES

Item response theory (IRT) offers a powerful set of techniques for detecting DIF within a set of items presumed to measure a common factor. Through the analysis of item characteristic curves, the presence of DIF is indicated when item difficulty, discrimination, or lower asymptote (e.g., due to a guessing effect) parameter estimates differ appreciably across groups with the same underlying true ability (Zumbo, 2007). There are also IRT-based methods that estimate the amount of **differential test functioning** (DTF), or whether total scores based on cumulative responses over subsets of items relate to the underlying dimension in the same way over different populations. Oliveri, Olson, Ercikan, and Zumbo (2012) describe the application of parametric and nonparametric forms of IRT and the technique of ordinal logistic regression to simultaneously estimate DIF and DTF over samples of English- versus French-speaking students who completed

an objective measure of problem solving. They found that differential functioning at the item level was not generally detected by analysis at the test level. The direction of DIF was such that some items favored English-speaking students but others favored French-speaking students, so the overall effect of bias at the item level canceled out when item responses were summed to form total scores. Karami (2012) and Millsap (2011) describe additional statistical techniques for detecting DIF.

The technique of EFA is another alternative for invariance testing. Just as in CFA, there are special methods in EFA for analyzing ordinal indicators that generate estimates of item thresholds and pattern coefficients for latent response variables, in addition to estimates of factor means, variances, and covariances (Wirth & Edwards, 2007). There are various statistical measures of the similarity or convergence of factor solutions for the same indicators over different groups in EFA (Nimon & Reio, 2011). Millsap (2011) describes invariance testing that begins with EFA for unrestricted measurement models and then progresses to CFA for restricted measurement models.

## SUMMARY

In multiple-samples SEM analyses, it is common to impose cross-group equality constraints on certain unstandardized estimates. This is done in order to test the hypothesis that the constrained parameters are equal in the populations from which two or more samples were selected. If the fit of the model with equality constraints is appreciably worse than that of the unconstrained model, then the hypothesis of equality is rejected. Multiple-samples CFA tests hypotheses about measurement invariance, or whether a set of indicators measures the same constructs in different groups. The most basic form is configural invariance, where the same CFA model is fitted to data from two or more samples but with no constraints other than those needed to identify the model. The next level is weak invariance, which assumes that the unstandardized pattern coefficient for each indicator is equal across the groups. The hypothesis of strong invariance requires equality over samples of intercepts for continuous indicators or thresholds for ordinal indicators. Strict invariance requires all the equality constraints as well as error variance and covariance homogeneity. Evidence for strict invariance is required in order to conclude that the indicators measure the factors identically in each sample.

## LEARN MORE

The authoritative book by Millsap (2011) deals with CFA, IRT, and other statistical methods for invariance testing with continuous or ordinal data. An additional example of testing for invariance is available in Wu, Li, and Zumbo (2007).

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

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research & Evaluation*, 12(3). Retrieved from <http://pareonline.net/pdf/v12n3.pdf>

### EXERCISES

1. Explain why Equation 16.1 says that strict invariance is required in order to meaningfully compare the mean of a continuous indicator over different groups.
2. Calculate the predicted means for all indicators in each group based on the results in Tables 16.3 and 16.4.
3. Calculate the Welch–James test for the group contrast on the conflict factor, given the information in Tables 16.3 and 16.4.
4. Convert the thresholds in Table 16.5 for item to proportions of responses in each of the four categories (0, 1, 2, 3) for item  $X_1$  in both groups.
5. Calculate  $R^2$  for each indicator in both groups using the results in Tables 16.7 and 16.8.

## Appendix 16.A

### Welch–James Test

The Welch–James (WJ) test for independent samples (James, 1951) assumes normality, but not homoscedasticity. The equation is

$$t(df_{\text{WJ}}) = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_{\text{WJ}}} \quad (16.6)$$

The degrees of freedom for the WJ test are estimated from the group variances and sizes as

$$df_{\text{WJ}} = \frac{\left( \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2} \right)^2}{\frac{(\hat{\sigma}_1^2)^2}{n_1^2(n_1-1)} + \frac{(\hat{\sigma}_2^2)^2}{n_2^2(n_2-1)}} \quad (16.7)$$

which is referred to as the **Welch–Satterthwaite equation**. Non-integer values of  $df_{\text{WJ}}$  are expected. The standard error of the WJ test statistic is

$$\hat{\sigma}_{\text{WJ}} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad (16.8)$$

A central  $t$  distributional calculator that accepts that either integer or non-integer  $df$  values and generates critical values for the WJ test is freely available over the Internet.<sup>1</sup>

<sup>1</sup>[www.usablestats.com/calcs/tinv](http://www.usablestats.com/calcs/tinv)

# Interaction Effects and Multilevel Structural Equation Modeling

---

This chapter introduces two classes of advanced techniques: estimation of interactive effects of observed and latent variables and multilevel SEM. Also described are techniques for estimating mediation when interaction is assumed, including conditional process analysis and causal mediation analysis. Entire books are written about some of these topics, so they cannot be reviewed here in great detail. Instead, I want to make you aware that these possibilities exist in SEM and cite enough advanced works so that you can learn more through additional study. As you read this chapter, keep in mind this advice credited to the French chemist and microbiologist Louis Pasteur: Chance only favors invention for minds that are prepared for discoveries by patient study and persevering efforts.

---

## INTERACTIVE EFFECTS OF OBSERVED VARIABLES

Estimation of interactive effects of continuous observed variables in SEM uses the same method as in **moderated multiple regression**. It involves **product terms** that represent interaction effects. A product term is literally the product of the scores from two different variables, such as  $XW = X \times W$ . The same basic method is used to estimate curvilinear effects (trends) except that **power terms** (also known as polynomials) are created by exponentiation where the scores (base numbers) are raised to a power, such as  $X^2 = X \times X$ , which represents a quadratic trend. Estimation of the curvilinear effects is not described here, but the principles are the same as for estimating interactive effects—see Cohen et al. (2003, chap. 6).

Consider the data in Table 17.1. Regressing Y on both X and W yields  $R^2 = .033$ , and the unstandardized prediction equation is

$$\hat{Y} = .112X - .064W + 8.873 \quad (17.1)$$

**TABLE 17.1. Data Set for Moderated Multiple Regression**

Predictors			Criterion
X	W	XW	Y
2	10	20	5
6	12	72	9
8	13	104	11
11	10	110	11
4	24	96	11
7	19	133	10
8	18	144	7
11	25	275	5

Note.  $M_X = 7.125$ ,  $SD_X = 3.137$ ,  $M_W = 16.375$ , and  $SD_W = 6.022$ .

where .112 is the slope of the regression plane along the X-axis across all levels of W and  $-.064$  is the slope of the regression plane along the W-axis across all levels of X (e.g., Figure 2.2). The intercept, 8.873, is the predicted score on Y, given  $X = W = 0$ . But scores of zero are not within the range of observed scores for X and W (see Table 17.1). Exercise 1 asks you to center the scores on X and W, or create the variables

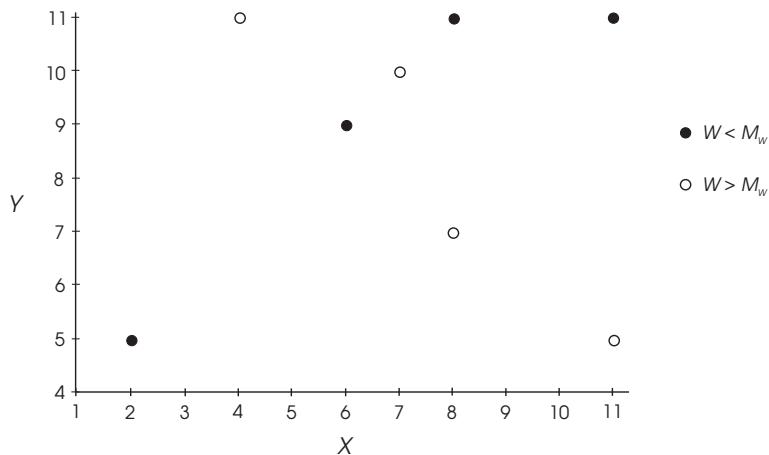
$$x = X - M_X \quad \text{and} \quad w = W - M_W$$

where  $x$  and  $w$  are mean-deviated scores, and then regress Y on  $x$  and  $w$ . You should verify for this second analysis that  $R^2 = .033$  and

$$\hat{Y} = .112x - .064w + 8.625 \quad (17.2)$$

The new intercept, or 8.625, equals the predicted score on Y, given  $X = M_X$  and  $W = M_W$ . Centering the predictors changed the intercept but neither  $R^2$  (.033) nor the partial regression coefficients (compare Equations 17.1 and 17.2).

Both analyses without a product term just described concern unconditional linear effects, but inspection of the scores in Table 17.1 indicates that effects are actually conditional: The linear relation between X and Y is *positive* for cases with lower scores on W, but it is *negative* for cases with higher scores on W. This pattern is depicted in Figure 17.1 where cases with scores  $< M_W$  are represented with closed circles and cases with scores  $> M_W$  are represented with open circles. There is a similar change in the direction of the relation between W and Y: It is positive at higher levels of X, negative at lower levels. Here, W moderates the relation of X to Y just as X moderates the relation of W to Y. This describes interaction, which is symmetrical. Exercise 2 asks you to explain the difference between moderation and mediation.

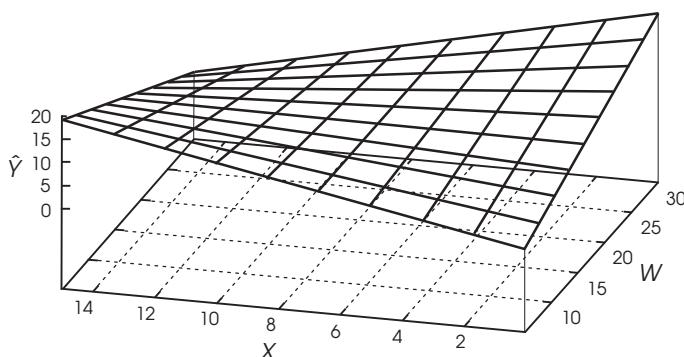


**FIGURE 17.1.** Scatterplot for the data set in Table 17.1 for variables  $X$  and  $Y$ . Closed circles indicate scores on  $W$  below the mean, and open circles indicate scores on  $W$  above the mean.

The product term  $XW$  in Table 17.1 represents the interactive effect when  $Y$  is regressed on  $X$ ,  $W$ , and  $XW$ . In the analysis just described,  $R^2 = .829$  and

$$\hat{Y} = 1.768X + .734W - .108XW - 3.118 \quad (17.3)$$

The intercept in Equation 17.3 has the usual interpretation:  $\hat{Y} = -3.118$ , given  $X = W = 0$ . The coefficient for  $XW$ , or  $-.108$ , says that the slope for the linear regression of  $Y$  on  $X$  decreases by  $.108$  for every increase in  $W$  of one point. Similarly, the slope of the regression line for predicting  $Y$  from  $W$  decreases by the same amount,  $.108$ , given a 1-point increase in  $X$ . Presented in Figure 17.2 is the regression surface defined by Equation 17.3. Note in the figure that slopes for regressing  $Y$  on  $X$  change across the levels of  $W$ ,



**FIGURE 17.2.** Unstandardized regression surface for predicting  $Y$  from  $X$ ,  $W$ , and  $XW$  for the data in Table 17.1.

and vice versa. Compare Figure 17.2 for a prediction equation with a product term with Figure 2.2 for an equation with no product term.

Coefficients for  $X$  and  $W$  in Equation 17.3 estimate *conditional* linear effects. For example, the coefficient 1.768 is the slope of the line for regressing  $Y$  on  $X$ , but only if  $W = 0$ . Similarly, the coefficient .734 is the slope of the  $Y$ -on- $W$  regression line, but only if  $X = 0$ . Note that scores of zero are not within the range of observed scores for either predictor (Table 17.1). Perhaps zero is not even a possible score on either predictor. If so, then interpretation of the regression coefficients for  $X$  or  $W$  in Equation 17.3 may have little practical value.

Centering the predictors (i.e., calculate  $x$  and  $w$ ) will deal with the problem that zero is not among their original scores. Next, take the product of the centered scores, or  $xw$ , and then regress  $Y$  on  $x$ ,  $w$ , and  $xw$ . Doing so for the data in Table 17.1 generates the results  $R^2 = .829$  and

$$\hat{Y} = -.001x - .035w - .108xw + 8.903 \quad (17.4)$$

Note that centering changed neither the value of  $R^2$  (.829) nor the coefficient for the product term, or  $-.108$ . But centering alters the partial regression coefficients for both predictors and the intercept, too (compare Equations 17.3 and 17.4). The latter, or 8.903, is the predicted score on  $Y$ , if  $x = w = 0$ . Scores of zero on the centered variables correspond to scores that equal the mean of each predictor in their original units. This means that  $\hat{Y} = 8.903$ , if  $X = M_X$  and  $W = M_W$ .

The coefficients for  $x$  and  $w$  in Equation 17.4 are interpreted as follows: The slope of the regression line for predicting  $Y$  from  $X$  is  $-.001$ , if  $w = 0$ ; that is, the case has an average score on  $W$ . That is, variables  $X$  and  $Y$  are basically unrelated, given  $W = M_W$ . Similarly, the slope for predicting  $Y$  from  $W$  is  $-.035$  for cases with average scores on  $X$ . In analyses with a product term, centering shifts the interpretation of the intercept and partial regression coefficients from scores of zero in the original units ( $X = W = 0$ ) (e.g., Equation 17.3) to scores of zero in mean-deviated units ( $x = w = 0$ ) (e.g., Equation 17.4), which corresponds to the means in the original units ( $M_X, M_W$ ). *This is the only effect of centering in moderated multiple regression.*

Some researchers believe that centering is *required* in moderated multiple regression, but this widespread conviction is false (Edwards, 2009). Centering is *optional*, but it may be useful when scores of zero are not possible on a particular predictor. There is also no problem with centering the scores on some predictors (e.g., those for which zero is not a valid score), while at the same time analyzing original scores on other predictors (e.g., those for which zero is a valid score). Centering does not alter **essential multicollinearity** in the data, or the overlap of effects of individual predictors with that of their interactive effect. Centering affects only **nonessential multicollinearity**, which refers to intercorrelations among original variables and product terms due to the scales of the variables and not to underlying effects in the data (Cohen et al., 2003). Exercise 3 asks you to calculate the intercorrelations among  $X$ ,  $W$ , and  $XW$  for the original scores in Table 17.1 and also among  $x$ ,  $w$ , and  $xw$  based on the centered scores. For these data, you

will find that centering results in lower intercorrelations among centered versus original scores, but centering has no other effect.

Interpretation of the interactive effect of continuous variables is described next. Begin with Equation 17.3 for regressing  $Y$  on  $X$ ,  $W$ , and  $XW$  for the data in Table 17.1. Next, rearrange the terms in this equation so that there is no product term. For example, the expression

$$\hat{Y} = (1.768 - .108W)X + (.734W - 3.118) \quad (17.5)$$

is algebraically equivalent to Equation 17.3 but has no product term. The regression coefficient for  $X$  in Equation 17.5, or

$$(1.768 - .108W) X$$

is the **simple slope** that depends on  $W$ . This means that as  $W$  changes, so does the slope of the regression line for predicting  $Y$  from  $X$ . The **simple intercept** in Equation 17.5, or

$$(.734W - 3.118)$$

is the value of the intercept for predicting  $Y$  from  $X$  that depends on  $W$  (see Figure 17.2).

Now substitute in Equation 17.5 meaningful values of  $W$  and inspect the resulting **simple regressions** for predicting  $Y$  from  $X$  as a function of  $W$ . For the data in Table 17.1,  $M_W = 16.375$  and  $SD_W = 6.022$ . Scores on  $W$  that fall  $-2$ ,  $-1$ ,  $0$ ,  $+1$ , and  $+2$  standard deviations from the mean are, respectively,

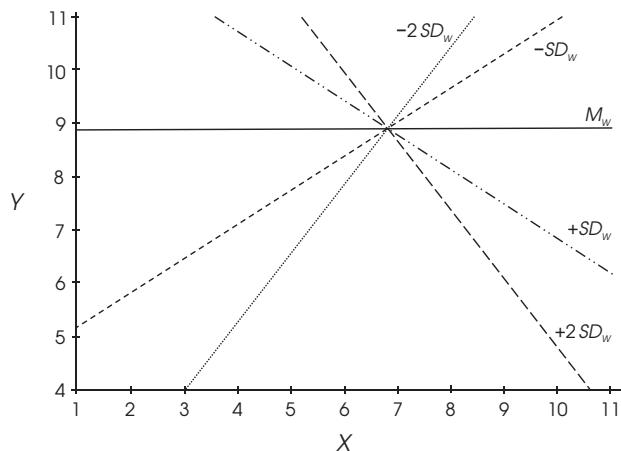
$$4.331, 10.353, 16.375, 22.397, \text{ and } 28.419$$

Suppose that  $W = 22.397$ , or  $M_W + SD_W$ . Plugging this constant into Equation 17.5 generates the prediction equation listed next:

$$\hat{Y}_{W = 22.397} = -.651X + 13.321$$

Presented in Table 17.2 are all simple regression equations for predicting  $Y$  from  $X$  at each of the five levels of  $W$  just defined. The same equations are plotted in Figure 17.3. The simple slopes progressively change from positive values for  $W < M_W$  to negative values for  $W > M_W$ . For cases with average scores of  $W$ , the simple slope is basically zero ( $-.001$ ), so  $X$  and  $Y$  are unrelated at this level of  $W$ . Note that the values of the simple intercepts in Table 17.2 and Figure 17.3 also vary over the levels of  $W$ .

In large samples assuming normality, the ratio of a simple slope over its standard error is a  $z$  test. A related concept is that of **regions of significance**, or a range of values on  $W$  for which the simple slope for predicting  $Y$  from  $X$  is statistically significant. A more useful explanatory device involves confidence intervals based on simple slopes called **confidence bands**, which can be plotted in order to interpret interaction. In this



**FIGURE 17.3.** Simple regression lines for predicting  $Y$  from  $X$  as a function of  $W$  for the data in Table 17.1.

case, a third variable  $W$  moderates the relation between  $X$  and  $Y$  where the confidence bands do not include slopes of zero. Preacher, Rucker, and Hayes (2007) describe computer utilities for analyzing simple slopes that are freely available over the Internet.<sup>1</sup> Scripts for visualizing interactions with spin plots in SAS/STAT are also freely available.<sup>2</sup>

Standardized regression coefficients (beta weights) do not have the normal interpretation for product terms. This is because, in most cases, the product of normal deviates from two original variables, such as  $z_X \times z_W$ , does not equal the  $z$  score of their unstandardized product, or  $z_{XW}$ . Computer procedures for multiple regression typically report beta weights for the data in Table 17.1 that correspond to the model

$$\hat{z}_Y = b_X z_X + b_W z_W + b_{XW} z_{XW} \quad (17.6)$$

**TABLE 17.2. Simple Regression Equations for Predicting  $Y$  from  $X$  Conditional on the Level of  $W$  for the Data in Table 17.1**

W		
Level	Score	Regression equation
$-2 SD_W$	4.331	$\hat{Y} = 1.300X + .061$
$-SD_W$	10.353	$\hat{Y} = .650X + 4.481$
$M_W$	16.375	$\hat{Y} = -.001X + 8.901$
$+SD_W$	22.397	$\hat{Y} = -.651X + 13.321$
$+2 SD_W$	28.419	$\hat{Y} = -1.301X + 17.742$

<sup>1</sup>[www.quantpsy.org/medn.htm](http://www.quantpsy.org/medn.htm)

<sup>2</sup>[www.ats.ucla.edu/stat/sas/faq/spplot/reg\\_int\\_cont.htm](http://www.ats.ucla.edu/stat/sas/faq/spplot/reg_int_cont.htm)

but the coefficient  $b_{XW}$  does not correctly estimate the standardized interaction effect. The correct standardized model would include a coefficient for the term “ $z_X \times z_W$ ,” not  $z_{XW}$ . Cohen et al. (2003, pp. 282–284) describe how to obtain correct standardized estimates of interaction.

## Extensions and Challenges

A **residualized product term** is created using the technique of **residual centering** that controls for effects of lower-order variables  $X$  and  $W$  and consequently is uncorrelated with them (Lance, 1988; Little, Bovaird, & Widaman, 2006). A residualized product is created in two steps by first regressing  $XW$  on  $X$  and  $W$ . The residuals from the regression analysis just described are uncorrelated with both  $X$  and  $W$  but still convey information about the interaction. In the second step,  $Y$  is regressed on  $X$ ,  $W$ , and  $XW_{\text{res}}$ , the residualized  $XW$  term created in the first analysis. Exercise 4 asks you to perform the two analyses just described for the data in Table 17.1.

The term  $XW$  represents a linear  $\times$  linear interaction (e.g., Figure 17.3). The term  $XW^2$  represents a linear  $\times$  quadratic interaction, which means that the linear relation of  $X$  to  $Y$  changes faster at higher (or lower) levels of  $W$ . It also means that the quadratic relation between  $W$  and  $Y$  changes at a constant (linear) rate across the levels of  $X$ . Estimation of the interactive effect just described requires that  $Y$  is regressed on  $X$ ,  $W$ ,  $W^2$ ,  $XW$ , and  $XW^2$  (e.g., Cohen et al., 2003, pp. 292–295). The product term  $XWZ$  represents the three-way linear interaction among these variables when all lower-order terms are also included in the equation. A three-way linear interaction means that the linear  $\times$  linear interaction between any two predictors, such as  $X$  and  $W$ , changes in a linear across the levels of the other predictor, or  $Z$  in this case—see Dawson and Richter (2006).

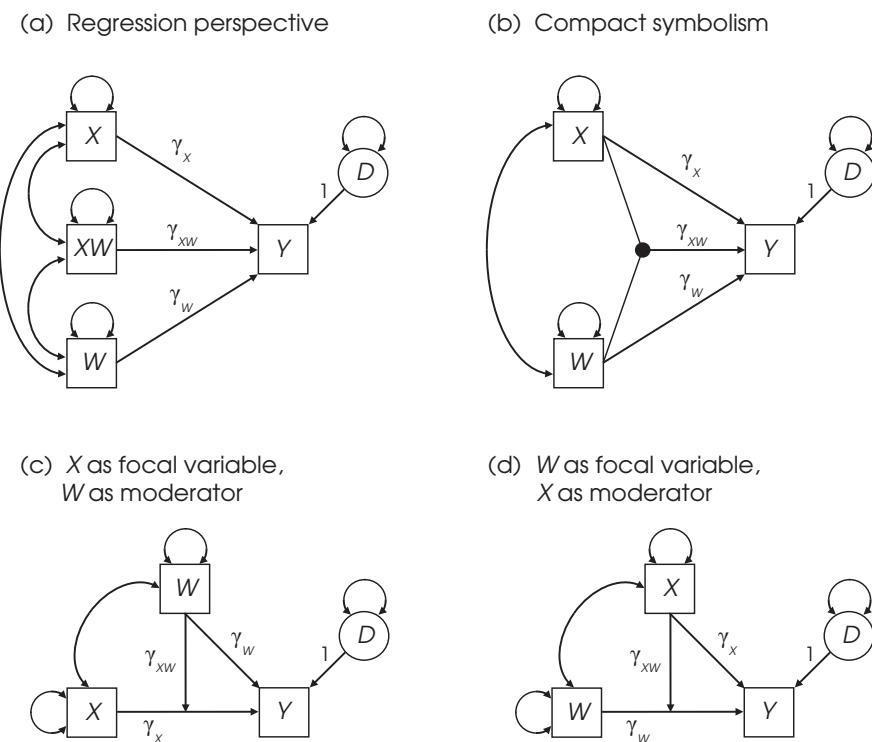
Judd, Kenny, and McClelland (2001) describe analyses for repeated measures data where difference scores are regressed on a presumed moderator. The difference score measures the effect of  $X$  on  $Y$  for each case, and the hypothesis of interaction is supported if the magnitudes of those effects vary over levels of the moderator. Edwards (1995) describes the analysis of interaction when congruence, or the degree of similarity between two constructs, such as supervisor–subordinate agreement, is the outcome variable. Cohen et al. (2003, chap. 9) and Hayes and Matthes (2009) describe the estimation of interactive effects of categorical variables. Interaction can also be analyzed in multilevel modeling, a point considered later in this chapter.

Score reliability is a critical problem. This is because measurement error can be much greater in product terms than in either constituent variable. This in turn reduces both the absolute coefficient for the product term and the power of its significance test (Edwards, 2009). Measurement error in the criterion that varies over the levels of a predictor can also bias the regression coefficient for product terms that involve that predictor (Baron & Kenny, 1986). One way to address these problems is to use predictor variables with excellent scores reliabilities (e.g.,  $> .90$ ). Another is to estimate the interactive effects of latent variables in the multiple-indicator (i.e., SEM) approach described in a later section. Both very large samples and very precise scores are needed for adequate power when testing for interaction (Aguinis, 1995).

## INTERACTIVE EFFECTS IN PATH ANALYSIS

Presented in Figure 17.4 are four different ways to represent the interactive effect of two continuous variables in **moderated path analysis**. Figure 17.4(a) depicts the regression of  $Y$  on  $X$ ,  $W$ , and  $XW$ , just as in moderated multiple regression. The symbols  $\gamma_X$ ,  $\gamma_W$ , and  $\gamma_{XW}$  designate, respectively, the direct effects of these variables, where  $\gamma_{XW}$  in particular estimates the interactive effect. Compact symbolism associated with Mplus is used in Figure 17.4(b) to represent the product term with a closed circle. The path emitted from the closed circle represents the interaction. The product term is not explicitly shown in Figures 17.4(c) and 17.4(d), but its analysis along with  $X$  and  $W$  is assumed. The hypothesis that the effect of focal variable  $X$  changes over the levels of the moderator  $W$  is represented in Figure 17.4(c) by the path from  $W$  that bisects the  $X$ -to- $Y$  path. The parameter for this effect is  $\gamma_{XW}$ , just as in Figures 17.4(a) and 17.4(b). The roles of focal variable and moderator are reversed in Figure 17.4(d), but this switch is allowed because interaction is symmetrical. All four models in Figure 17.4 are equivalent for the same data.

The models in Figure 17.4 do not have mean structures, so only slopes are analyzed, not intercepts. But it is no special problem in moderated path analysis to analyze means along with covariances, which adds a mean structure to the basic covariance structure. In this case, both simple slopes and simple intercepts (and regions of significance and



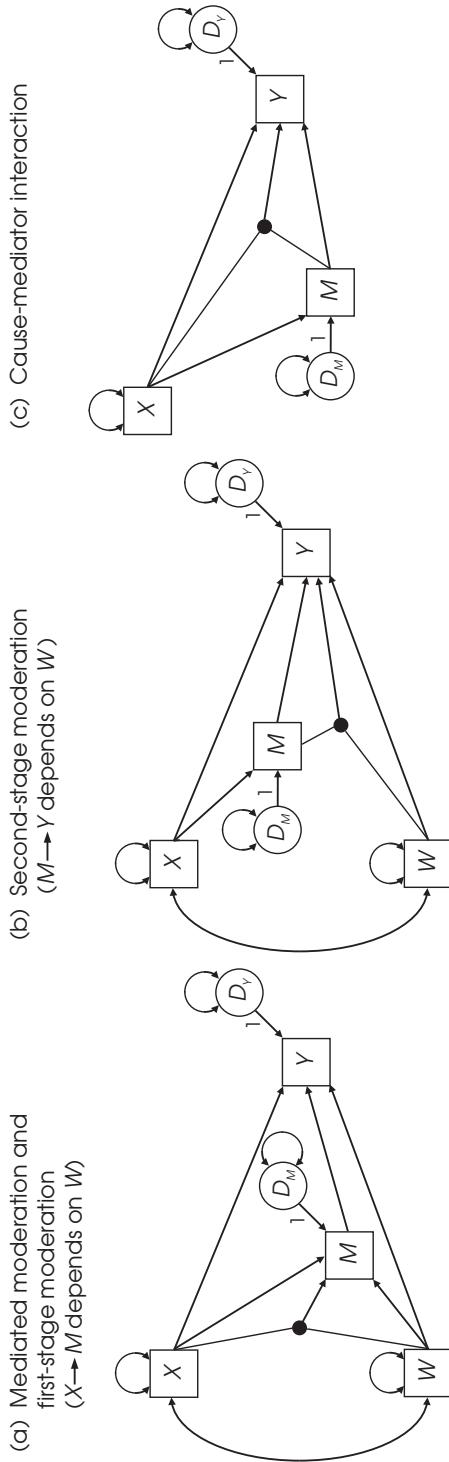
**FIGURE 17.4.** Path-analytic representations of an interactive effect of continuous observed variables.

confidence bands, too) can be estimated in moderated path analysis (Preacher et al., 2007). Keep in mind the following points:

1. Kenny (2013) reminds us that just as a mediational model is a causal model, so too is a model of moderation; thus, if the basic directionality assumptions are incorrect, the results may have little value. For example, an interactive effect between  $X$  and  $W$  can be reversed if the direct effect between  $X$  and  $Y$  is reversed.
2. Kenny (2013) gives this example of how curvilinear and interactive effects can be confounded. Suppose that  $X$  is income and  $Y$  is work motivation. Their relation is curvilinear such that their association is stronger at lower income levels. If variable  $W$  is age, then because younger workers earn less money, the “interaction” between age and income could be found, such that the relation between income and motivation is stronger for younger workers. In order to avoid confusing curvilinear and interactive effects, Edwards (2009) recommends the routine inclusion of the power terms  $X^2$  and  $W^2$  whenever estimating coefficients for the product term  $XW$ .
3. Edwards (2009) reminds us that although a product term can be represented as a causal variable in a path diagram (e.g., Figure 17.4(a)), it actually has no causal potency. This is because a product term does not represent a distinct entity apart from its components variables. It is merely a mathematical construction that represents joint effects when analyzed together with its components. This idea is consistent with Figures 17.4(c) and 17.4(d), where the product term is assumed but not depicted as causal.

## CONDITIONAL PROCESS MODELING

The researcher in **conditional process modeling** is concerned with analyzing the boundary conditions of direct or indirect effects, or the circumstances under which causal effects occur (Hayes, 2013a). In the discussion of mediation that follows, research designs with time precedence in measurement of the presumed causes, mediators, and outcomes are assumed; otherwise, the term *indirect effect* is preferred for cross-sectional designs. A key concept in conditional process modeling is that of **mediated moderation**, which describes when an interactive effect, or moderation, is transmitted at least in part through at least one intervening variable, or mediation (Baron & Kenny, 1986). Consider Figure 17.5(a), which features compact symbolism to represent the hypothesis that the interactive effect of  $X$  and  $W$  on  $Y$  is entirely indirect through  $M$ , the mediator. Lance (1988) tested basically the same path model where  $Y$  represents recall accuracy for the script of a lecture,  $X$  stands for memory demand,  $W$  symbolizes complexity of social perception, and  $M$  corresponds to specific behaviors mentioned in the script. The results indicated that the interactive effects of memory demand and perception complexity on recall accuracy were mainly indirect through the mention of specific behaviors.



**FIGURE 17.5.** Path models with compact symbolism for (a) mediated moderation and first-stage moderation of the path  $X \rightarrow M$ . (b) Second-stage moderation of the path  $M \rightarrow Y$ . (c) Interaction of the causal variable  $X$  with the mediator  $M$ .

**Moderated mediation**, also known as a **conditional indirect effect** (James & Brett, 1984; Preacher et al., 2007) is another key concept. It is indicated when the strength of at least one direct effect in an indirect pathway depends on the level of an external variable. The external variable in multiple-samples path analysis is group membership: if the size of an indirect effect differs over samples, then that effect is conditional. It is also possible to estimate conditional indirect effects in a single-sample analysis where the external variable is just another variable in the same model, a point that is elaborated next.

There is more than one kind of moderated mediation. Look back at Figure 17.5(a). It represents mediated moderation (described earlier) and also **first-stage moderation** (Edwards & Lambert, 2007), where the *first* path of the indirect effect of  $X$  on  $Y$ , or  $X \rightarrow M$ , depends on the external variable  $W$ . This interactive effect is represented in the figure by the regression of  $M$  on  $X$ ,  $W$ , and  $XW$ . Thus, the direct effect of  $X$  on  $M$  changes over the levels of  $W$ . The same model also represents the hypothesis that the *first* path of the indirect effect of  $W$  on  $Y$ , or  $W \rightarrow M$ , depends on  $X$  (i.e., interaction is symmetrical). Figure 17.5(b) represents **second-stage moderation**, where just the *second* path of the indirect effect of  $X$  on  $Y$ , or  $M \rightarrow Y$ , depends on the level of the external variable  $W$ . This interaction is represented in the figure by the regression of  $Y$  on  $X$ ,  $M$ ,  $W$ , and  $MW$ . In this way, the direct effect of  $M$  on  $Y$  depends on  $W$ .

Other forms of moderated mediation are summarized next (Edwards & Lambert, 2007):

1. **First-and-second-stage moderation** occurs when  $W$  moderates *both* direct paths of the indirect pathway from  $X$  to  $Y$  through  $M$ . A variation is when one variable, such as  $W$ , moderates  $X \rightarrow M$ , and a different variable, such as  $Z$ , moderates  $M \rightarrow Y$ .

2. In **direct effect and first-stage moderation**, both the direct effect of  $X$  on  $Y$  and the *first* path of the indirect effect, or  $X \rightarrow M$ , are moderated by an external variable. In **direct effect and second-stage moderation**, an external variable moderates both the direct effect and just the *second* path of the indirect effect, or  $M \rightarrow Y$ . And in **total effect moderation**, an external variable moderates both paths of the indirect effect and also the direct effect.

Curran, Hill, and Niemiec (2013) evaluated a conditional process model of children's engagement and disaffection with soccer. They found that structure from coaches related positively to engagement and negatively to disaffection, and that these relations are indirect through children's psychological need satisfaction. These indirect effects were appreciable only among children who reported higher levels of autonomy support from their coaches. Preacher et al. (2007) provide SPSS macros for analyzing path models with conditional indirect effects of the kind represented in Figure 17.5 (see footnote 1). Additional computer tools for conditional process analysis by Hayes (2013b) are also freely available over the Internet.<sup>3</sup>

---

<sup>3</sup>[www.afhayes.com](http://www.afhayes.com)

## CAUSAL MEDIATION ANALYSIS

Causal mediation analysis was introduced in Chapter 8. Briefly, direct, indirect, and total effects are defined in terms of counterfactuals for linear or nonlinear models as well as for continuous or dichotomous mediators or outcomes. Interaction between the causal variable  $X$  and the mediator  $M$  is assumed. Conversely, the classical Baron and Kenny (1986) method assumes no interaction and estimates indirect effects for continuous variables in linear models as products of the coefficients for direct effects that make up an indirect pathway (e.g., Table 11.4). If there is no interaction, the results of a causal mediation analysis and the Baron–Kenny method for a linear model are the same; otherwise, the two approaches can generate quite different estimates for the same data. This is why causal mediation analysis extends the Baron–Kenny method for the types of models just described to allow for interaction (Valeri & VanderWeele, 2013). It is also possible to estimate cause–mediator interaction in conditional process modeling (Preacher et al., 2007), but such effects are not defined from a counterfactual perspective.

Figure 17.5(c) is a moderated path model of interaction between causal variable  $X$  and mediator  $M$ . We assume for this discussion a binary variable  $X$  that represents random assignment of cases to either control ( $X = 0$ ) or treatment ( $X = 1$ ) conditions. Both the mediator  $M$  and outcome  $Y$  are continuous. Randomization for  $X$  guarantees over random replication samples no confounding between treatment and mediator and also between treatment and outcome, but it does not rule out confounding between mediator and outcome. Causal mediation analysis generally assumes no mediator–outcome confounding, but this is a strong assumption when the mediator is an individual difference variable. Experimental mediation designs covered in Chapter 8, such as manipulation-of-mediation designs, may reduce the biasing effects of confounders of the mediator and outcome. See MacKinnon and Pirlott (2015) for more information.

If means are analyzed along with covariances, Figure 17.5(c) with cause–mediation interaction generates the two unstandardized regression equations listed next:

$$\begin{aligned}\hat{M} &= \beta_0 + \beta_1 X \\ \hat{Y} &= \theta_0 + \theta_1 X + \theta_2 M + \theta_3 XM\end{aligned}\quad (17.7)$$

where  $\beta_0$  and  $\theta_0$  are the intercepts for, respectively, the regressions of  $M$  on  $X$  and of  $Y$  on  $X$ ,  $M$  and the product term  $XM$ ; the coefficient for  $X$  when predicting  $M$  is  $\beta_1$ ; and  $\theta_1$ – $\theta_3$  are the coefficients for, respectively,  $X$ ,  $M$ , and  $XM$  when predicting  $Y$ .

The controlled direct effect (CDE) of  $X$  is how much outcome  $Y$  would change on average if the mediator were controlled at the same level  $M = m$  for all cases but the treatment were changed from  $X = 0$  (control) to  $X = 1$  (treatment). The natural direct effect (NDE) is how much the outcome would change on average if  $X$  were changed from control to treatment, but the mediator is kept to the level that it would have taken in the control condition. The natural indirect effect (NIE) is the amount the outcome would change on average in the treatment condition, but the mediator changes from as it would from the control condition to the treatment condition. The total effect of  $X$  on  $Y$  is the sum of NDE and NIE.

Given the expressions in Equation 17.7 for the model in Figure 17.5(c), the CDE, NDE, and NIE can be expressed as follows (Valeri & VanderWeele, 2013):

$$\begin{aligned} \text{CDE} &= \theta_1 + \theta_3 m \\ \text{NDE} &= \theta_1 + \theta_3 \beta_0 \\ \text{NIE} &= (\theta_2 + \theta_3) \beta_1 \end{aligned} \tag{17.8}$$

In the equations, note that the CDE is defined for a particular level of the mediator ( $M = m$ ) and that the NDE is defined at the predicted level of the mediator in the control group ( $X = 0$ ). This predicted level is  $\beta_0$ , which is the intercept in the equation for regressing  $M$  on  $X$ . Also note that if there is no interaction, then  $\theta_3 = 0$  (see Equation 17.7). In this case, both the CDE and NDE are equal to the direct effect in the Baron–Kenny approach, or  $\theta_1$ , and the NIE equals the Baron–Kenny product estimator of the indirect effect, or  $\beta_1 \theta_2$ .

Presented next is a numerical example based on one by Petersen et al. (2006), where  $X = 1$  is an antiretroviral therapy for HIV and  $X = 0$  is control; the mediator  $M$  is the blood level of HIV (viral load); and the outcome  $Y$  is the level of CD4 T-cells (helper white blood cells). We assume that the scores are not centered. Suppose that the two unstandardized regression equations are

$$\begin{aligned} \hat{M} &= 1.70 - .20X \\ \hat{Y} &= 450.00 + 50.00X - 20.00M - 10.00XM \end{aligned} \tag{17.9}$$

In words, the predicted viral load in the control group is 1.70, but treatment reduces this count by .20. For control patients with no viral load, the predicted level of CD4 T-cells is 450.00. Treatment increases this count by 50.00 for patients with no viral load, and for every one-point increase in viral load for control patients, the level of CD4 T-cells decreases by 20.00. For treated patients, the slope of the regression line for predicting the level of CD4 T-cells from viral load decreases by 10.00 compared with control patients. These equations imply that

$$\begin{aligned} \beta_0 &= 1.70 \text{ and } \beta_1 = -.20 \\ \theta_0 &= 450.00, \theta_1 = 50.00, \theta_2 = -20.00, \text{ and } \theta_3 = -10.00 \end{aligned}$$

Continuing with the same example, the direct effect of treatment versus control at a given level of viral load  $M = m$  is

$$\text{CDE} = 50.00 - 10.00m$$

The researcher can select a particular value of  $m$  and then estimate the CDE by substituting this value in the formula just listed. Another option is to estimate the direct effect at the weighted average value of  $M$  for the whole sample. The direct effect of treatment estimated at the level of viral load that would have been observed in the control condition is

$$NDE = 50.00 - 10.00 (1.70) = 33.00$$

where 1.70 is the predicted value of the viral load in the control condition ( $X = 0$ ) (see Equation 17.9). The indirect effect of treatment allowing viral load to change as it would from the control condition to the treatment condition is estimated as

$$NIE = (-20.00 - 10.00)(-.20) = 6.00$$

where  $-.20$  is the difference in viral load between the control and treatment conditions (see Equation 17.9). The total effect (TE) of treatment is the sum of the natural direct and indirect effects just calculated, or

$$TE = 33.00 + 6.00 = 39.00$$

That is, antiretroviral therapy increases the level of CD4 T-cells by 39.00 through both its natural direct effect (33.00) and its natural indirect effect through viral load (6.00).

MacKinnon and Pirlott (2015) describe causal mediation analysis and other techniques, such as instrumental variable estimation, for enhancing the causal interpretation of mediator-to-outcome direct effects. Valeri and VanderWeele (2013) describe macros for SPSS and SAS/STAT for causal mediation analysis that allow covariates, such as variables that control for treatment history.<sup>4</sup> The decomposition of the total effect of the cause on the outcome into parts due (1) to neither mediation nor moderation, (2) to just mediation but not moderation, (3) to just moderation but not mediation, and (4) to both mediation and moderation is outlined by VanderWeele (2014). Muthén and Asparouhov (2015) describe causal mediation analysis with latent variables that controls for measurement error in observed variables. Imai, Keele, and Yamamoto (2010) describe the mediation package for R that also conducts sensitivity analyses about the effects of violated assumptions.<sup>5</sup> Special syntax for causal mediation analysis is available in Mplus (Muthén & Muthén, 1998–2014). Hicks and Tingley (2011) describe the MEDIATION module for causal mediation analysis in STATA.<sup>6</sup> See VanderWeele (2015) for more information about causal mediation analysis.

## INTERACTIVE EFFECTS OF LATENT VARIABLES

In the **indican product approach** in SEM, product terms are specified as multiple indicators of latent product variables that represent interactive or curvilinear effects. We will consider only the estimation of interactive effects of latent variables, but the same principles apply to the analysis of curvilinear effects of latent variables. We assume

<sup>4</sup><http://dx.doi.org/10.1037/a0031034.supp>

<sup>5</sup><http://cran.r-project.org/web/packages/mediation>

<sup>6</sup><http://econpapers.repec.org/software/bocbocode/s457294.htm>

that all indicators are continuous. Suppose that factor  $A$  has two indicators,  $X_1$  and  $X_2$ , and factor  $B$  has two indicators,  $W_1$  and  $W_2$ . The reference variable for  $A$  is  $X_1$ , and the corresponding indicator for  $B$  is  $W_1$ . Equations that specify the measurement model for these indicators are:

$$\begin{aligned} X_1 &= A + E_{X_1} & W_1 &= B + E_{W_1} \\ X_2 &= \lambda_{X_2} A + E_{X_2} & W_2 &= \lambda_{W_2} B + E_{W_2} \end{aligned} \quad (17.10)$$

The free parameters of these equations (17.10) include the pattern coefficients for  $X_2$  and  $W_2$ , the variances of the four error terms, and the variances and covariance of factors  $A$  and  $B$ .

The latent product variable  $AB$  represents the linear  $\times$  linear interactive effect of factors  $A$  and  $B$  when an outcome variable is regressed on  $A$ ,  $B$ , and  $AB$ . Its indicators are the four product indicators

$$X_1W_1, X_1W_2, X_2W_1, \text{ and } X_2W_2$$

By taking the products of the corresponding expressions in Equation 17.10 for the nonproduct indicators, the equations of the measurement model for the product indicators are

$$\begin{aligned} X_1W_1 &= AB + AE_{W_1} + BE_{X_1} + E_{X_1}E_{W_1} \\ X_1W_2 &= \lambda_{W_2}AB + AE_{W_2} + \lambda_{W_2}BE_{X_1} + E_{X_1}E_{W_2} \\ X_2W_1 &= \lambda_{X_2}AB + \lambda_{X_2}AE_{W_1} + BE_{X_2} + E_{X_2}E_{W_1} \\ X_2W_2 &= \lambda_{X_2}\lambda_{W_2}AB + \lambda_{X_2}AE_{W_2} + \lambda_{W_2}BE_{X_2} + E_{X_2}E_{W_2} \end{aligned} \quad (17.11)$$

These equations (17.11) show that the product indicators depend on a total of *eight* additional latent product variables besides  $AB$ . For example,  $X_1W_1$  depends on

$$AB, AE_{W_1}, BE_{X_1}, \text{ and } E_{X_1}E_{W_1}$$

The last term just listed is the residual for  $X_1W_1$ . All pattern coefficients in Equation 17.11 are either the constant 1.0 or functions of the coefficients for the nonproduct indicators  $X_2$  and  $W_2$  (Equation 17.10). Thus, no new coefficients need to be estimated for the product indicators.

The only other parameters of the measurement model for the product indicators are the variances and covariances of the latent variables implied by Equation 17.11. Assuming normal distributions for all nonproduct latent variables (Equation 17.10) and centered scores on all nonproduct indicators, Kenny and Judd (1984) showed that (1) covariances among the latent product variables and the nonproduct factors  $A$  and  $B$  are all zero; (2) variances of the latent product variables can be expressed as functions of the variances of the nonproduct latent variables, or

$$\begin{aligned}
 \sigma_{AB}^2 &= \sigma_A^2 \sigma_B^2 + \sigma_{A,B}^2 & \sigma_{E_{X_1} E_{W_1}}^2 &= \sigma_{E_{X_1}}^2 \sigma_{E_{W_1}}^2 \\
 \sigma_{BE_{X_1}}^2 &= \sigma_B^2 \sigma_{E_{X_1}}^2 & \sigma_{E_{X_1} E_{W_2}}^2 &= \sigma_{E_{X_1}}^2 \sigma_{E_{W_2}}^2 \\
 \sigma_{BE_{X_2}}^2 &= \sigma_B^2 \sigma_{E_{X_2}}^2 & \sigma_{E_{X_2} E_{W_1}}^2 &= \sigma_{E_{X_2}}^2 \sigma_{E_{W_1}}^2 \\
 \sigma_{AE_{W_1}}^2 &= \sigma_A^2 \sigma_{E_{W_1}}^2 & \sigma_{E_{X_2} E_{W_2}}^2 &= \sigma_{E_{X_2}}^2 \sigma_{E_{W_2}}^2 \\
 \sigma_{AE_{W_2}}^2 &= \sigma_A^2 \sigma_{E_{W_2}}^2
 \end{aligned} \tag{17.12}$$

where the term  $\sigma_{A,B}^2$  represents the covariance between factors  $A$  and  $B$ . For example, the variance of the latent product factor  $AB$  equals the product of the variances for factors  $A$  and  $B$  plus their covariance. All variances of the other latent product variables are related to variances of the nonproduct latent variables; thus, no new variances need to be estimated, so the measurement model for the product indicators is theoretically identified.

Presented in Figure 17.6 is the entire SR model for the regression of  $Y$  on factors  $A$ ,  $B$ , and  $AB$  in the Kenny–Judd approach. The measurement models for the nonproduct indicators and the product indicators defined by, respectively, Equations 17.10 and 17.11 are also represented in the figure. Among parameters for the structural model in the figure, coefficients for the paths

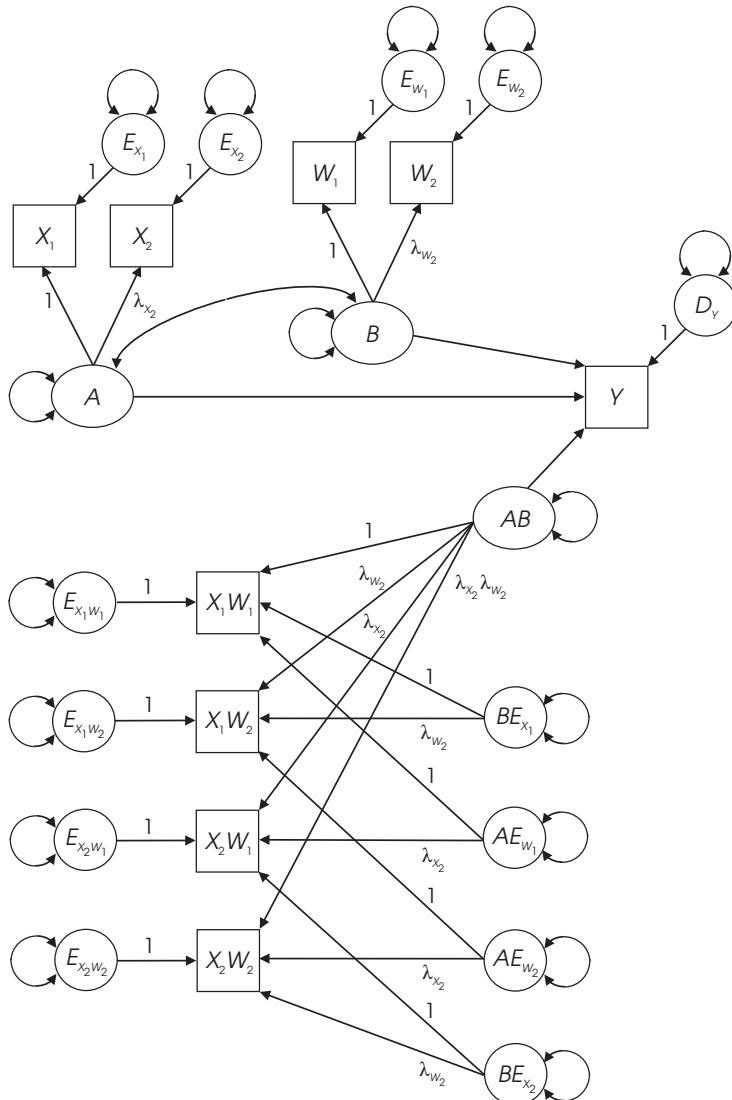
$$A \rightarrow Y, B \rightarrow Y, \text{ and } AB \rightarrow Y$$

estimate, respectively, the linear effects of factors  $A$  and  $B$  and their linear  $\times$  linear interaction, each controlling for the other effects.

### Estimation in the Kenny–Judd Method

Kenny and Judd (1984) were among the first to describe a method for estimating structural equation models with product indicators. The **Kenny–Judd method** is generally applied to observed variables in mean-deviated form; that is, scores on nonproduct indicators are centered before creating product indicators. It has two potential complications:

1. The method requires the imposition of nonlinear constraints in order to estimate some parameters of the measurement model for the product indicators (see Equation 17.12). Not all SEM computer tools support nonlinear constraints. Correctly specifying all such constraints can be tedious and error prone.
2. A product variable is not normally distributed even if each of its components is normally distributed. For example, the Kenny–Judd method assumes that factors  $A$  and  $B$  and the error terms for their nonproduct indicators in Figure 17.6 are normally distributed. But the products of these variables, such as  $AB$ , are not normally distributed,



**FIGURE 17.6.** Model with interactive and lower-order effects of factors  $A$  and  $B$  in the Kenny-Judd method.

which violates the normality requirement of default maximum likelihood estimation. Yang-Wallentin and Jöreskog (2001) demonstrate the estimation of a model with product indicators using a corrected normal theory method that can generate robust standard errors and adjusted model test statistics. Also, minimum sample sizes of 400–500 cases may be needed for large samples when estimating even relatively small models.

Presented in Table 17.3 is a covariance matrix generated by Kenny and Judd (1984) for a hypothetical sample of 500 cases. I fitted the model in Figure 17.6 to the data in

**TABLE 17.3. Input Data (Covariances) for Analysis of a Model with an Interactive Effect of Latent Variables with the Kenny-Judd Method**

Variable	1	2	3	4	5	6	7	8	9
1. $X_1$	2.395								
2. $X_2$	1.254	1.542							
3. $W_1$	.445	.202	2.097						
4. $W_2$	.231	.116	1.141	1.370					
5. $X_1 W_1$	-.367	-.070	-.148	-.133	5.669				
6. $X_1 W_2$	-.301	-.041	-.130	-.117	2.868	3.076			
7. $X_2 W_1$	-.081	-.054	.038	.037	2.989	1.346	3.411		
8. $X_2 W_2$	-.047	-.045	.039	-.043	1.341	1.392	1.719	1.960	
9. $Y$	-.368	-.179	.402	.282	2.556	1.579	1.623	.971	2.174

Note. These data for a hypothetical sample are from Kenny and Judd (1984);  $N = 500$ .

Table 17.3 using the Kenny–Judd method in Mplus (Muthén & Muthén, 1998–2014). You can download from the website for this book all computer files for this analysis. The Mplus syntax file is annotated with comments that explain the specification of nonlinear constraints. Because Kenny and Judd (1984) used a generalized least squares (GLS) estimator in their original analysis of these data, I specified the same estimator in this analysis with Mplus. The input data for this analysis are in matrix form, so it is not possible to use a corrected normal theory method or to analyze a mean structure because such methods require raw data files.

With a total of nine observed variables (four nonproduct indicators, four product indicators, and  $Y$ ; see Figure 17.6), a total of  $9(10)/2$ , or 45 observations are available for this analysis. There are 13 free parameters, including

1. two pattern coefficients for  $X_2$  and  $W_2$ ;
2. seven variances of  $A$ ,  $B$ ,  $E_{X_1}$ ,  $E_{X_2}$ ,  $E_{W_1}$ ,  $E_{W_2}$ , and  $D_Y$ ;
3. the covariance between  $A$  and  $B$ ; and
4. three directs of  $A$ ,  $B$ , and  $AB$  on  $Y$ ,

so  $df_M = 45 - 13 = 32$ . The analysis in Mplus converged to an admissible solution. Values of selected fit statistics are reported next and generally indicate acceptable global fit:

$$\begin{aligned}\chi^2_M(32) &= 41.989, p = .111 \\ \hat{\epsilon} &= .025, 90\% \text{ CI } [0, .044], p_{\epsilon_0 \leq .05} = .988 \\ \text{CFI} &= .988, \text{SRMR} = .046\end{aligned}$$

The Mplus-generated GLS parameter estimates for the model of Figure 17.6 are similar to those reported by Kenny and Judd (1984) in their original analysis. Factors A, B, and AB together explain .868 of the total variance in Y. The unstandardized equation for predicting Y is

$$\hat{Y} = -.169A + .321B + .699AB$$

This prediction equation has no intercept because no means were analyzed. But you could use the same method demonstrated earlier to rearrange this equation to (1) eliminate the product term and (2) generate simple regressions of say, Y on factor B where the slope of the equation varies as a function of factor A. Following these steps will show that the relation between Y and B is positive for levels of A above the mean (i.e.,  $> 0$ ) but negative for levels of A below the mean ( $< 0$ ) for the model in Figure 17.6 and the data in Table 17.3.

### Alternative Estimation Methods

When using the Kenny–Judd method to estimate the interactive effects of latent variables, Jöreskog and Yang (1996) recommend adding a mean structure to the model. (The basic Kenny–Judd method does not analyze means.) They argue that because the means of the indicators are functions of other parameters in the model, their intercepts should be added to the model in order for the results to be more accurate. They also note that a single product indicator is all that is needed for identification, not every possible product indicator as in the Kenny–Judd method. Marsh, Wen, and Hau (2006) recommended analyzing **matched-pairs indicators** in which information from the same indicator is not repeated. For example, given indicators  $X_1$  and  $X_2$  for factor A and indicators  $W_1$  and  $W_2$  for factor B, the pair of product indicators  $X_1W_1$  and  $X_2W_2$  is a set of matched-pairs indicators because no individual indicator appears twice in any product term. The pair  $X_1W_2$  and  $X_2W_1$  is the other set of matched-pairs indicators for this example.

Ping (1996) describes a two-step estimation method that does not require nonlinear constraints, so it can be used with just about any SEM computer tool. In the first step, the model is analyzed without product variables. Parameter estimates from this analysis are used to calculate the parameters of the measurement model for the product indicators. These calculated values are then specified as fixed parameters in the second step where all indicators, product and nonproduct, are analyzed together. Included in the results of the second analysis are estimates of latent interactive effects. Microsoft Excel templates for Ping's method are freely available.<sup>7</sup>

Bollen's (1996) two-stage least squares (2SLS) method for latent variables is another estimation option. This method requires at least one product indicator of a latent product variable and a separate product indicator specified as an instrument. It does not assume normality nor is the method iterative, so 2SLS estimation may be less suscep-

---

<sup>7</sup>[www.wright.edu/~robert.ping](http://www.wright.edu/~robert.ping)

tible to technical problems in the analysis. In a simulation study, Yang-Wallentin (2001) compared default ML estimation with Bollen's (1996) 2SLS method applied to the estimation of latent interactive effects. Neither method performed especially well for sample sizes of  $N < 400$ , but for larger samples differences in bias across the two methods were generally slight.

Wall and Amemiya (2001) describe the **generalized appended product indicator (GAPI) method** for estimating latent curvilinear or interaction effects. As in the Kenny–Judd method, products of the observed variables are specified as indicators of latent product variables, but the GAPI method does not require normality. Consequently, it is not assumed that the latent product variables are independent. Instead, these covariances are estimated as part of the analysis of the model with a mean structure. But all other nonlinear constraints in the Kenny–Judd method are imposed in the GAPI method. A drawback of this method is that its implementation in computer syntax can be complicated (Marsh et al., 2006).

Marsh et al. (2006) describe an **unconstrained approach** to the estimation of latent interactive and curvilinear effects that does not assume multivariate normality. It features product indicators, but it imposes no nonlinear constraints on estimates of the correspondence between product indicators and latent product variables. This unconstrained approach is generally easier to implement in computer syntax than the GAPI method (Marsh et al., 2006). Results of computer simulation studies by Marsh, Wen, and Hau (2004) generally support the unconstrained method for large samples.

Klein and Moosbrugger's (2000) **latent moderated structural equations (LMS) method** uses a special form of ML estimation that assumes normal distributions for nonproduct variables but estimates the degree of non-normality implied by the latent product variables. This method adds a mean structure to the model, and it uses a form of the expectation–maximization (EM) algorithm in estimation. The LMS method directly analyzes raw data (there is no matrix input) from the nonproduct indicators without the need to create any product indicators. Of all the methods described here, the LMS method may be the most precise because it explicitly estimates the degree of non-normality.

The LMS method is computationally intensive. Klein and Muthén (2007) describe a simpler version known as **quasi-maximum likelihood (QML) estimation** that closely approximates the results of the LMS method. A version of the QML method is implemented in Mplus. It relies on numerical integration to generate the parameter estimates. It also features special compact syntax. For example, the keyword "xwith" automatically creates a latent product variable. A drawback is that most traditional SEM fit statistics (including residuals) are not available in the Mplus implementation of the QML method. Instead, the relative fit of different models can be compared using predictive fit indexes such as the AIC or BIC.

Little, Bovaird, and Widaman (2006) describe an extension of residual centering for estimating interactive or curvilinear effects of latent variables. In this approach, the researcher creates every possible product indicator and then regresses each product indicator on its own set of constituent nonproduct indicators. The residuals from this analysis

represent interaction but are uncorrelated with the corresponding nonproduct factors. The only special parameterization in this approach is that error covariances are specified between pairs of residualized product indicators based on common nonproduct indicators. This method can be implemented in basically any SEM computer tool, and it relies on traditional fit statistics in the evaluation of global fit. Based on computer simulation studies by Little, Bovaird, and Widaman (2006), their residualized product indicator method generally yielded similar parameter estimates compared with the LMS/QML method and also the Marsh et al. (2004) unconstrained method used with mean centering.

No single method for estimating the curvilinear or interactive effects of latent variables has emerged as the “best” approach, but this is an active research area—see Marsh, Wen, Nagengast, and Hau (2012) for more information. An empirical example is briefly described next. Klein and Moosbrugger (2000) applied the LMS method in a sample of 304 middle-aged men in order to estimate the latent interactive effects of flexibility in goal adjustment and perceived physical fitness on level of complaining about one’s mental or physical state. They found higher levels of perceived fitness neutralized the effects of goal flexibility, but effects of goal flexibility on complaining were more substantial at lower levels of perceived fitness.

## MULTILEVEL MODELING AND SEM

Multilevel modeling (MLM)—also called hierarchical linear modeling and random coefficient modeling, among other variations—is a set of statistical techniques for analyzing hierarchical (nested) data where (1) scores are clustered into larger units and (2) scores within each unit may not be independent. Repeated measures data are inherently hierarchical in that multiple scores are nested under the same person. Dependence among such scores is explicitly estimated in various statistical techniques for repeated measures data, including SEM (e.g., autocorrelated errors in latent growth models).

In a **complex sampling design**, the levels of at least one higher-order variable are selected prior to sampling cases within each level. Suppose that a sample is composed of 7,000 students who attend 50 different schools. Scores from students who attend the same school may not be independent. This is because such students are exposed to common influences that include school staff, discipline policy, or curriculum that are not exactly repeated in any other school. Score dependence implies that the use of formulas for estimating standard errors in a single-level sample that assume independence (e.g., Equation 3.2) may not yield correct results; specifically, such formulas tend to *underestimate* sampling error in complex samples. Because standard errors are the denominators of significance tests, underestimation leads to rejection of the null hypothesis too often. Thus, one motivation for MLM is to correctly estimate standard errors in complex sampling designs.

Many SEM computer tools have the capability to automatically adjust the standard errors in a complex sample. The researcher typically specifies at least one higher-order variable that in syntax may be designated as a cluster, stratification, or grouping vari-

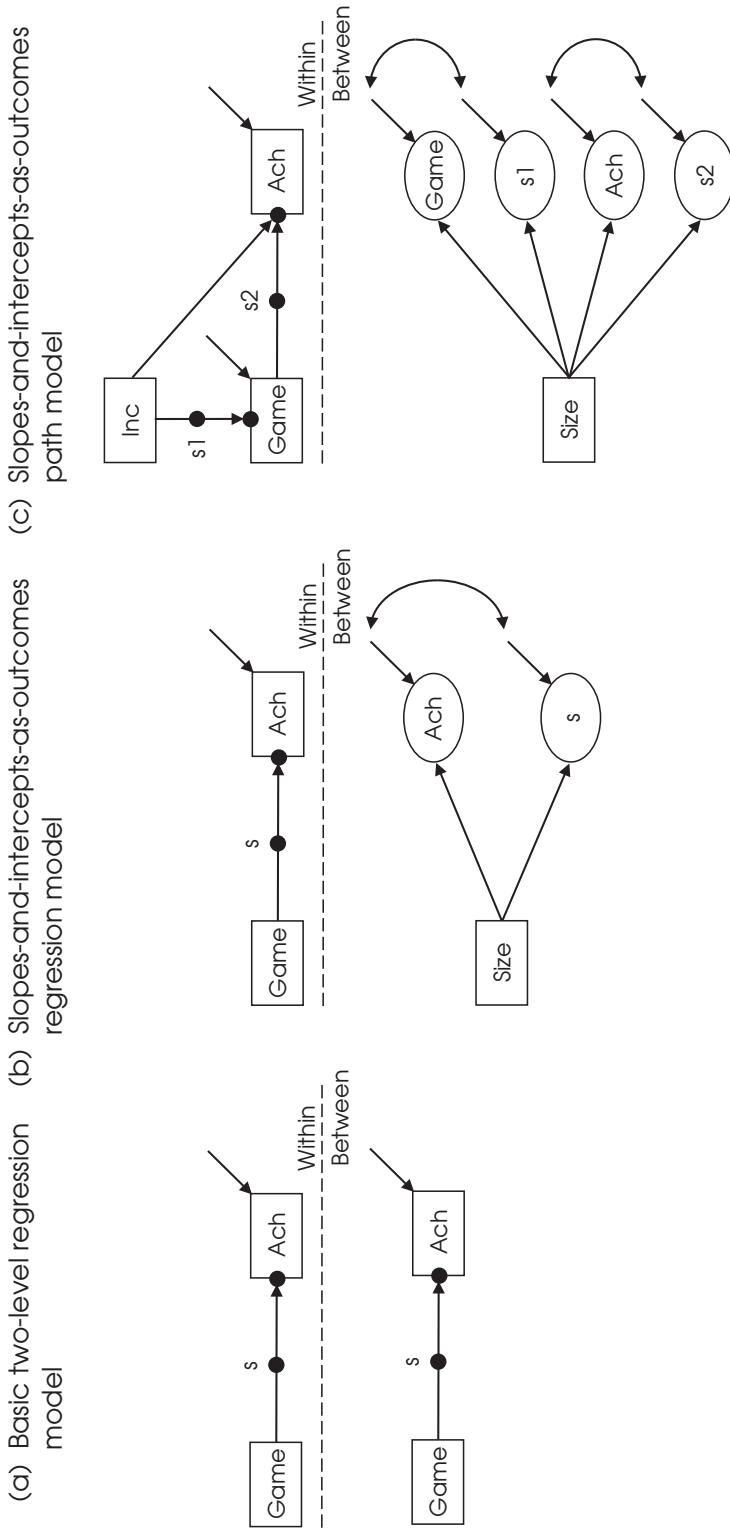
able, such as “school” in the example where students are sampled from different schools. Such variables may not appear in the model, but the computer “knows” how to adjust estimates of standard errors for lack of independence of scores within each level of a cluster. The same computer program may also accept sampling weights that adjust proportions of cases that belong to a particular group to make them conform to known population base rates. If too many high-income households were sampled, for instance, then weights could be applied to reduce the relative contribution of scores from such households. Doing so also increases the relative weight of data from households with lower incomes.

But there is more to MLM than just adjusting standard errors or assigning probability weights in complex sampling designs. An example is the estimation of **contextual effects** of higher-order variables on scores of cases in a hierarchical data set. The goal is to explain within-groups variation with a combination of within- and between-groups predictors. Suppose that the amount of time spent playing computer games (Game) and scholastic achievement (Ach) are measured among students who attend 50 different schools. The schools vary in their numbers of enrolled students, or size. This is a characteristic of schools, not of students.

In a two-level regression analysis, the relation between Game and Ach could be analyzed at two-different levels, within schools and between schools. The within-schools association would be estimated from the pooled within-schools covariance matrix for these variables, and the between-school level would be estimated from the between-schools covariance matrix, which is based on average (aggregated) values in each school. It is possible that the within-schools association between Game and Ach differs from that observed at the between-schools level. For example, the two variables may be unrelated at the level of individual students within schools but negatively related at the between-schools level. This can happen if the within-schools slopes or intercepts differ from those based on averages over schools (Stapleton, 2013).

Presented in Figure 17.7(a) is a diagram for the two-level regression analysis. It features compact symbolism for diagrams of multilevel models associated with Mplus. The slope for the regression of Ach on Game at each level, between or within, is designated in the figure with a closed circle labeled “s” that falls in the middle of the path from Game to Ach. Intercepts are represented by the unlabeled closed circles that lie at the ends of the same paths. A mean structure is assumed in Figure 17.7(a) but is not explicitly represented. Curran and Bauer (2007) describe an alternative symbolism for multilevel models that explicitly represents mean structures.

Figure 17.7(b) represents a random coefficients regression analysis for a **slopes-and-intercepts-as-outcomes model**. At the within level, Ach is regressed on Game, just as in Figure 17.7(a). But school size (Size) is specified as a contextual variable at the between level in Figure 17.7(b), such that the slopes and intercepts from the within level are regressed on Size at the between level. These slopes and intercepts are assumed to vary and covary over schools and thus are specified as random latent variables in the between part of the model in Figure 17.7(b). These specifications correspond to **cross-level interactions** where the regression of Ach on Game at the within level varies as



**FIGURE 17.7.** Representation in Mplus-type compact symbolism of a basic two-level regression model (a), a slopes-and-intercepts-as-outcomes regression model (b), and a slopes-and-intercepts-as-outcomes path model (c). Game, amount of time spent playing computer games; Ach, scholastic achievement; Size, school enrollment size; Inc, family income.

function of school size at the between level. For example, if the slopes for predicting Ach from Game are steeper for larger schools than for smaller schools—that is, the two variables are more strongly related in bigger schools—then there is an interaction between a within variable (Game) and a between variable (Size). It is also possible to specify that slopes only are random (**slopes-as-outcomes model**) or to specify that intercepts only are random (**intercepts-as-outcomes model**). The complexity of the analysis increases quickly in designs with more than two levels, such as students within schools within neighborhoods. This is probably why most applications of random coefficients regression concern just two levels.

Some limitations of traditional MLM are summarized next (Bauer, 2003; Curran, 2003):

1. Scores on within or between predictors, such as Game and Size in Figure 17.7(b), are assumed to be perfectly reliable. This is because there is no direct way in MLM to represent measurement error.
2. There is also no direct way in MLM to represent predictors or outcomes as latent variables with multiple indicators; that is, it is difficult to specify a measurement model.
3. There are methods to estimate indirect effects in MLM, but they can be difficult to apply.
4. There is no single inferential test of the whole model (i.e., global fit) in MLM. Instead, the *relative* predictive power of alternative models can be evaluated in the same sample.

## Convergence of MLM and SEM

The relative weaknesses of MLM correspond to the relative strengths of SEM. Briefly, it is straightforward in SEM to represent error variance for either single or multiple indicators through specification of a measurement model. Latent variables can be specified as either predictors or outcomes in a structural model. The estimation of direct or indirect effects is routine in SEM, and there are inferential tests of global model fit.

Early efforts to extend capabilities of SEM to MLM analyses were based on tricking SEM computer programs into analyzing two-level models (e.g., Duncan, Duncan, Hops, & Alpert, 1997). The trick was to exploit the capability of the software to simultaneously estimate a structural equation model across two groups. But in this case the “groups” corresponded to two different models, a within and between model, each estimated in the same complex sample. The pooled within-groups covariance matrix is the input data for the within model, and the between-groups covariance matrix based on group averages is the input data for the between model. Because older versions of many SEM computer tools had no built-in capabilities for handling data from complex samples, it was usually necessary to calculate these two data matrices separately using an external program, such as SPSS.

Writing the syntax required to trick older versions of SEM programs into analyzing even relatively simple two-level models, such as Figure 17.1(b), quickly “becomes a remarkably complex, tedious, and error-prone task” (Curran, 2003, p. 557)—that is, a data management nightmare. Fortunately, newer versions of some programs, including EQS, LISREL, Mplus, and Stata, feature special syntax that makes it easier to specify and analyze multilevel models in complex samples. This special syntax is more compact to use than programming languages of the kind used to trick older versions of SEM computer programs into analyzing two-level models.

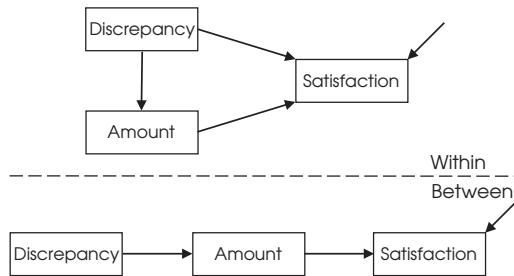
The relatively new capability just described supports the full integration of MLM and SEM in a framework known as **multilevel structural equation modeling** (ML-SEM). For example, it would be no special problem to specify in an SEM computer tool that supports MLM the model in Figure 17.7(c), which is a two-level, slopes-and-intercepts-as-outcomes path model. Family income (Inc) is represented as a cause of both Game and Ach in the within model. In the between model, two pairs of correlated random slopes and intercepts from the within model are regressed on Size. The first pair is from the regression of Game on Inc, and the second pair is from the regression of Ach on Game and Inc, both from the within level. The between model represents the cross-level interactions that involve school size.

Using an SEM computer tool with built-in MLM support makes possible the basic types of ML-SEM analyses summarized next:

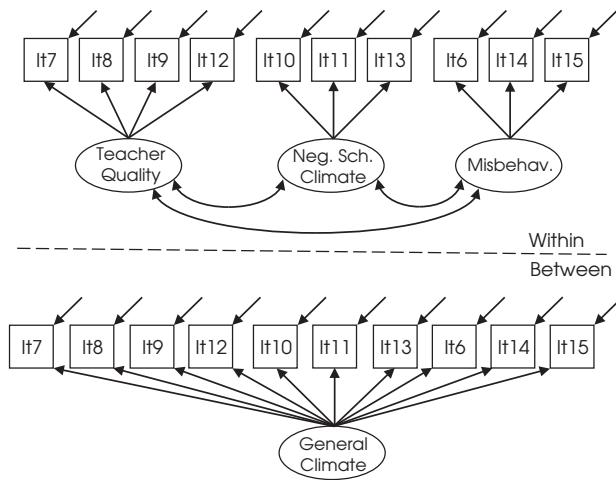
1. Estimation of correct standard errors when fitting a model to data from a complex sample.
2. Analysis of one model at the within level but a different model at the between level. The within model could be identical to the between model (e.g., Figure 17.7(a)), but the between model could also be a different model. Structural models at either level can include indirect effects. Preacher, Zhang, and Zyphur (2011) outline the possible advantages of estimating indirect effects in ML-SEM.
3. Analysis of a slopes-and-intercepts-as-outcomes model where random slopes and intercepts from either observed or latent variables at the within level are regressed on either observed or latent variables that represent contextual variables at the between level.

Two examples of ML-SEM analyses that correspond to the second category are briefly described next. Wu (2008) administered to 333 students questionnaires about life satisfaction, what respondents say they want (*amount*), and the gap between what they have and what they want (*have-want discrepancy*) in 12 different areas (social, financial, etc.). Because such ratings are repeated measures, the various areas are nested within persons, and the between level corresponds to differences across people that affect their satisfaction ratings in the various areas. The final multilevel path model retained by Wu (2008) is presented in Figure 17.8(a). At the within level, *have-want discrepancy* has both

(a) Two-level path analysis



(b) Two-level confirmatory factor analysis

**FIGURE 17.8.** (a) Two-level path analysis model analyzed by Wu (2008). (b) Two-level confirmatory factor-analytic model analyzed by Kaplan (2000).

direct and indirect effects on satisfaction. But at the between level, effects of *have-want discrepancy* are entirely indirect through the *amount* variable. Wu (2008) interpreted the results as suggesting that life satisfaction involves an explicit have-want comparison, but whether the effect is entirely indirect through what people say they want depends on the level of analysis, within or between.

Kaplan (2000, pp. 48–53) describes a multilevel CFA in a large sample of high school students enrolled in different schools. The students rated their perceptions of teacher quality, negative school environment (e.g., safety concerns), and disruptive behavior by other students. The final model retained by Kaplan (2000) is presented in Figure 17.8(b). The within model is a three-factor CFA model where the factors correspond to the three domains rated by students. The between model is simpler in that all indicators depend

on a single school climate factor. Thus, variation in student ratings within schools is differentiated along three dimensions, but a single factor explains variation between schools. See Stapleton (2013) for additional examples of ML-SEM.

## SUMMARY

In moderated path analysis, the interactive effects of observed variables are represented by product terms included in the model along with terms for the constituent variables. Path coefficients for the product terms estimate the corresponding interaction effects. One way to interpret an interactive effect between two continuous variables is to generate the simple regressions of the outcome variable on one predictor at different levels of the other predictor. One of the first approaches to estimate interactive effects of latent variables is the Kenny–Judd method, which features use of all possible product indicators of latent product variables and the imposition of nonlinear constraints. More recent methods are easier to implement. Causal mediation analysis estimates controlled direct effects, natural direct effects, and natural indirect effects, all defined from the perspective of counterfactuals. It offers consistent definitions of these effects, and interactions between causal variables and mediators are routinely estimated. Conditional process modeling also permits the estimation of mediation and moderation in the same analysis in the form of conditional indirect effects and indirect interactive effects. The convergence of multilevel modeling and SEM offers the capability to (1) analyze observed or latent predictors from both the within level and the between level (contextual effects) of observed or latent outcomes at the within level, (2) take account of measurement error, and (3) estimate both direct and indirect effects when structural models are specified. The increasing availability of SEM computer tools that directly support multilevel analyses in complex sampling designs is making it easier for researchers to actually reap these potential benefits.

## LEARN MORE

Marsh, Wen, Nagengast, and Hau (2012) describe options for estimating the interactive effects of latent variables, Stapleton (2013) reviews the rationale of ML-SEM, and Valeri and VanderWeele (2013) offer a clear account of causal mediation analysis.

Marsh, H. W., Wen, Z., Nagengast, B., & Hau, K. T. (2012). Structural equation models of latent interaction. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 436–458). New York: Guilford Press.

Stapleton, L. M. (2013). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (2nd ed., pp. 521–562). Greenwich, CT: IAP.

Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 2, 137–150.

### EXERCISES

1. Center the predictors for the data in Table 17.1 and verify Equation 17.2.
2. Explain the difference between a moderator and a mediator.
3. Compare correlations among  $X$ ,  $W$ , and  $XW$  versus  $x$ ,  $w$ , and  $xw$  for the data in Table 17.1.
4. For the data in Table 17.1, regress  $Y$  on  $X$ ,  $W$ , and  $XW_{\text{res}}$ . Comment on the results.

## 18

# Best Practices in Structural Equation Modeling

---

The techniques that make up the SEM family provide researchers with an extensive set of tools for hypothesis testing. As with any complex procedure, though, its use must be guided by reason. Some issues considered next were mentioned in earlier chapters, but they are discussed altogether here in the form of best practices. They are addressed under categories that involve specification, identification, measures, sample and data, estimation, respecification, tabulation, and interpretation. Other topics include avoiding confirmation bias and bottom-line perspectives, or the most important things about SEM. These categories are not mutually exclusive, but they offer a useful way to focus this discussion. You are encouraged to use these points as a checklist to guide the conduct of your own analyses. By adopting best practices and avoiding common mistakes, you are helping to improve the state of practice. This saying attributed to the psychologist William James is relevant here: Act as if what you do makes a difference; it does.

---

## RESOURCES

Listed in Table 18.1 are citations for works about reporting practices, problems, and guidelines in SEM. Works presented in the top of the table concern recommendations for better reporting. For example, Hoyle and Isherwood (2011) devised a questionnaire that covers all phases of the analysis in SEM. It is intended to supplement more general standards for the reporting of quantitative results in journal articles (American Psychological Association Publication and Communications Board Working Group on Journal Article Reporting Standards, 2008). DiStefano and Hess (2005) and Jackson, Gillaspy, and Purc-Stephenson (2009) offer recommendations for reporting CFA results. Thomp-

**TABLE 18.1. Citations for Works about Reporting Practices and Guidelines for Written Summaries of Results in Structural Equation Modeling**

Citation	Comment
<u>General reporting guidelines</u>	
Boomsma, Hoyle, and Panter (2012)	Guidance about latent-variable analyses and Monte Carlo studies
DiStefano and Hess (2005)	Use of CFA in construct validation
MacCallum and Austin (2000)	Review of shortcomings in reporting and recommendations for better practice
McDonald and Ho (2002)	Recommendations for reporting about data preparation and results of model testing
Hoyle and Isherwood (2011)	Questionnaire that verifies thorough and detailed reports of SEM analyses
Jackson, Gillaspy, and Purc-Stephenson (2009)	Reporting guidelines for CFA studies
Mueller and Hancock (2008)	Summary of best practices and reporting
Thompson (2000)	“Ten commandments” of SEM <sup>a</sup>
<u>Applying SEM in specific disciplines</u>	
Chan, Lee, Lee, Kubota, and Allen (2007)	Rehabilitation counseling
Chin, Peterson, and Brown (2008)	Marketing
Grace (2006)	Natural systems
Khine (2013)	Educational research and practice
Nunkoo, Ramkissoon, and Gursoy (2013)	Travel and tourism
Okech, Kim, and Little (2013)	Social work
Schreiber (2008)	Social and administrative pharmacy
Shah and Goldstein (2006)	Operations management

<sup>a</sup>No small samples; analyze covariance, not correlation matrices; simpler models are better; verify distributional assumptions; consider theoretical and practical significance, not just statistical significance; report multiple fit statistics; use two-step modeling for SR models; consider theoretically plausible alternative models; respecify rationally; acknowledge equivalent models.

son's (2000) "ten commandments" of SEM, summarized in the table footnote, are also pertinent. Citations at the bottom of the table concern the use of SEM in particular disciplines such as education (Khine, 2013) and natural systems (Grace, 2006), among others.

Presented next are recommendations for the conduct of SEM organized by phases of the analysis. Carefully review these suggestions and use them wisely; see also Mueller and Hancock (2008) and Schumaker and Lomax (2010, chap. 11) for related tips.

## SPECIFICATION

Despite all the statistical machinations in SEM, specification is the most important part, but occasionally researchers spend the least amount of time on it. Listed next are several ways to do your homework in this critical area:

- Describe the theoretical framework or body of empirical results that form the basis for specification. Articulate the specific problem addressed in the analysis. Explain why the application of SEM is relevant, including why using a simpler statistical technique is not better.
- For models with latent variables, define the corresponding constructs, especially if that construct is relatively unfamiliar in a particular literature. Avoid vague construct names such as "aggression." Instead, specify the particular kinds or aspects of aggression that correspond to the construct of interest, and use more precise names, such as verbal aggression, defensive aggression, dominance aggression, and so on.
- Multiple-indicator measurement is generally better than single-indicator measurement. An exception is when only one among a set of indicators of the same construct has good psychometric characteristics. In this case, it may be better to rely on the single best indicator.
- If multiple indicators are specified for a factor, then (1) all of those indicators should have good psychometrics and (2) each factor should have at least three indicators. Having only two indicators per factor may lead to problems, including failure of iterative estimation or empirical underidentification. It is also more difficult to estimate error correlations for factors with only two indicators, which can result in specification error.
- Avoid the specification where a single indicator of an exogenous construct is assumed to have no measurement error, especially if this assumption is known to be false. Instead, estimate the score reliability of the single indicator and specify an error term for that indicator, which explicitly controls for score imprecision. An alternative is to specify an instrument for the single indicator, which removes random error from that indicator, but the instrument must have good psychometrics—see Bollen (2012).

- State the rationale for directionality specifications. This includes both the measurement model and the structural model. For example, is reflective measurement appropriate for describing the directionalities of factor–indicator correspondences? Or would the specification of formative measurement make more sense? For the structural model, explain hypotheses about causal priority, especially if your research design has no formal elements, such as time precedence, that directly support causal inference.
- Specify reciprocal causation between a pair of variables in a cross-sectional design, if there is a theoretical rationale for such effects. But do not specify feedback loops as a way to mask uncertainty about directionality. Not only do feedback relations have their own assumptions (e.g., equilibrium), but their presence also makes a structural model nonrecursive, which introduces potential problems (e.g., identification) in analyzing the model.
- Be mindful about the consequences of omitting causes that are correlated with other variables in the model. If an omitted cause is uncorrelated with measured causes, then estimates of direct effects are not affected due to this omission. But it is rare that the types of causes studied by behavioral scientists are independent. Depending on the pattern of correlations between measured and unmeasured causes, estimates of direct effects can be too high or too low.
- Use insights from Pearl’s structural causal model (graph theory) to help you specify the model and plan the study. For example, you can work with directed acyclic graphs where some causes are assumed to be unmeasured in order to enumerate which causal effects are identified versus others that are not identified. Awareness of those not identified should prompt you to think about how to measure at least proxies (indicators) of omitted confounders.
- Specify design-driven correlated residuals, such as correlated disturbances in a structural model or correlated errors in a measurement model, if doing so is theoretically justifiable and identification requirements can be satisfied. Omission of such terms can lead to inaccurate results, especially for latent variables. In some disciplines, such as economics, the specification of correlated residuals is routine. Such specification should not be seen as a necessary evil. The flip side of this advice is to add correlated residuals without a basis in theory or study design (e.g., repeated measures), such as to improve model fit when there is no good reason to expect such effects. Doing so makes the model more complex, which will improve its fit, but at the cost of capitalizing on chance variation.
- Sometimes it is appropriate to expect that an indicator depends on two or more factors, but this specification should come from prior knowledge of that variable. Just like error correlations, the specification that an indicator is complex instead of simple makes a measurement model less parsimonious.
- In reflective measurement models, indicators of the same factor should have positive intercorrelations. It is a specification error if any of those correlations are close

to zero or negative. The specification of formative measurement where indicators are viewed as causing latent variables is an alternative, but you should have good theoretical reasons for specifying formative measurement instead of reflective measurement. It is also more difficult to assess construct validity in formative measurement models.

- Do not rely on empirical tests of whether a set of indicators is causal or reflective. For example, do not automatically conclude that a set of indicators is causal if their intercorrelations are not all positive. Also, do not rely on such tests of whether a variable is a proper instrument for another variable that is involved in a nonrecursive causal relation with a third variable. These kinds of specifications should come from your knowledge of measurement or substantive issues about causation in a particular research area. Empirical tests will just capitalize on chance variation, especially in small samples.
- Respect the parsimony principle: Specify the simplest model possible as your initial model, one that includes the effects of highest priority, given relevant theory. Doing so in single-sample analyses corresponds to model building, where the simplest model in a set of nested models is tested first. But when testing for measurement invariance in multiple-samples CFA, it is usually better to analyze the most complex model first. This is the model of configural invariance, which is then made simpler (it is trimmed) by imposing equality constraints on certain parameters, such as pattern coefficients, intercepts, thresholds, or error variances.
- The previous comments on parsimony are not intended to dissuade you from analyzing complex models per se. This is because a phenomenon that is complex may require a relatively complicated statistical model in order to capture its basic essence. The main point is that the model should be as simple as possible while respecting theory and prior empirical results. Models that are complex without justification are probably so specified in order to maximize fit.
- Declare whether the model includes a mean structure. If so, then describe which variables are included in the mean structure either in text or in the model diagram (e.g., use the symbol  $\triangle$  for the constant).
- If the model includes interaction effects, then explain how those effects were specified. Also state the theoretical rationale for expecting interaction.
- State the expected directions, positive or negative, of presumed causal effects. Give a complete diagram of your model. If possible, represent all error terms and unanalyzed associations (covariances) in the diagram. Make sure that the diagram is consistent with the text.
- Explain the rationale for constraints in parameter estimation. Relate these constraints to requirements for identification, relevant theory, previous empirical results, or aims of your study.
- Outline theoretically plausible alternative models. State the role of these alterna-

tive models in your plan for model testing. Describe this plan, such as comparing hierarchical models versus nonhierarchical models.

- In multiple-samples CFA, state the particular forms of measurement invariance to be tested and in what sequence (i.e., free vs. constrained baseline approach).
- Properly scale latent variables. In multiple-samples SEM, standardizing factors by fixing their variances to 1.0 is incorrect if the groups differ in their variabilities. Fixing the pattern coefficient for a reference variable to 1.0 (i.e., the factor is unstandardized) is preferable, but note that (1) the same pattern coefficient must be fixed in each group and (2) indicators with fixed pattern coefficients are assumed to be invariant across all samples. The effects coding method, where average unstandardized pattern coefficients or intercepts are fixed to equal, respectively, 1.0 or 0 in all groups, is an alternative for indicators of the same factor that also share the same metric. In single-sample analyses, fixing to 1.0 the variances of factors measured over time is also wrong if factor variability is expected to change.
- In complex sampling designs, do not assume that the within model and the between models are the same without verification. A lesson from multilevel modeling is that different models may describe covariance patterns at the within versus between levels of analysis.

## IDENTIFICATION

The problem of identification must be dealt with in virtually all SEM studies. Some recommendations for managing identification are listed next:

- Tally the number of observations and free parameters in your initial model. State (or indicate in a diagram) how latent variables are scaled; that is, demonstrate that necessary but insufficient conditions for identification are met.
- Comment on sufficient requirements that identify the particular kind of structural equation model you are analyzing. For example, if the structural model is non-recursive, is the rank condition sufficient to identify it? If the measurement model has complex indicators and error covariances, does their pattern satisfy the required sufficient conditions?
- If your model is especially complex, you need to ensure your readers that it is actually identified. Remember that it is theoretically possible for the computer to generate a converged, admissible solution for a model that is not actually identified, yet give no warning about the problem. Whatever solution is so computed, it is but one in an infinite number of solutions (i.e., it has no meaningful interpretation), if the model is not really identified.

## MEASURES

Your scores come from your measures, so those measures better be good. Some advice for dealing with the measurement problem in SEM are considered next:

- Explain your operationalizations for constructs of interest; that is, establish the links between construct definitions and specific characteristics or behaviors that are to be measured.
- State the psychometric characteristics of your measures, including evidence for score reliability (i.e., are they precise?) and score validity (e.g., do they actually measure target constructs?). It is best practice to estimate score reliability in your own sample.
- If it is impossible to estimate reliability in your own sample, report coefficients from other samples (reliability induction), but describe whether those other samples are similar to yours.
- If you anticipate missing data—for example, the design is longitudinal and participants can choose to withdraw from the study at any point—then specify and measure auxiliary variables that may predict the data loss pattern. These variables need not be included in the model, but they may be helpful when imputing multiple scores for each missing observation.
- The specification of parcels—average or total scores over sets of items—as continuous indicators in CFA requires the assumption that the items in each parcel are unidimensional, a requirement that should be addressed before analyzing the data in CFA. This is because trying to establish the unidimensionality of parcels in the same analysis is likely to capitalize heavily on chance. If so, then parceling can mask the true absence of unidimensionality and distort the results.
- Avoid jingle-jangle fallacies, which together involve confusion of test names with what those tests may actually measure. Construct validity is established over a series of empirical studies, not by the name given to a test by its author(s).

## SAMPLE AND DATA

The nature of samples and data are critical in any type of statistical analysis, SEM or otherwise. Emphasized next are issues more specific to SEM:

- Check the accuracy of data input or coding. Then check it again. Data entry mistakes are so easy to make, whether in recording the raw data or in typing the values of a correlation or covariance matrix. Even computer-based or automated data entry is not error free (e.g., programming errors lead to calculation of incorrect scores). Mistaken

specification of codes in statistical software is also common (e.g., “9” for missing data instead of “–9”).

- Clearly describe the characteristics of your sample (cases). If the sample is a convenience (ad hoc) sample, then explain how the cases may not be representative of the intended population of interest, given how they were selected.
- Explain on what basis the sample size was determined. Considerations include the results of a power analysis, use of a sample size heuristic (e.g., the  $N:q$  rule, where  $q$  is the number of free parameters), or resource constraints.
- Use a sample size that is large enough for your model and estimation method. As models become more complex relative to the number of cases, the statistical precision of the estimates becomes more doubtful. There is greater capitalization on chance in smaller samples, too. Methods that make fewer distributional assumptions generally require more cases. The analysis of ordinal data may require more cases compared with analyzing continuous data. Convince your readers that the sample size is large enough to do the job. There is no shame in using a simpler type of statistical technique in a smaller sample.
  - If a target sample size was established in a power analysis, state the target power (e.g.,  $\geq .90$ ) and describe the level of the analysis (i.e., whole model vs. individual parameter). State the particular null and alternative hypotheses and other power analysis parameters, such as the level of significance ( $\alpha$ ) and population values of approximate fit indexes.
  - If any data were simulated, state the computer tool and algorithm used, the number and sizes of generated samples, and how many generated samples were lost due to nonconvergence or other problems in the analysis.
  - If the sample is archival—that is, you are fitting a structural equation model within an extant data set—then mention possible specification errors due to omission of relevant causal or outcome variables. Another drawback to archival samples is the realization that the model is not identified. With the data already collected, it may be too late to do anything about identification. Adding exogenous variables is one way to remedy an identification problem for a nonrecursive structural model, and adding indicators can help to identify a measurement model.
  - Do not standardize the raw scores (i.e., convert them to normal deviates,  $z$ ), especially if you plan to use an estimation method that assumes unstandardized variables. Situations when standardizing the scores is especially inappropriate include the analysis of a model across independent samples with different variabilities, longitudinal data characterized by changes in variances or means over time, or a type of SEM analysis that requires the analysis of means, such as a latent growth model, which needs the input of not only means but covariances, too.

- Describe how data-related complications were handled. This includes the extent and strategy for dealing with missing observations or outliers, how extreme collinearity was managed, and the use of transformations, if any, to normalize continuous variables.
- Evaluate whether the pattern of missing data loss is random or systematic. This point assumes that there are more than just a few missing scores. Classical methods for dealing with missing data, such as case deletion or single-imputation methods, generally assume that the data loss pattern is missing completely at random, which is probably unlikely. These classical techniques have little basis in statistical theory and take scant advantage of information in the data. Modern alternatives, including those that impute multiple scores for missing observations based on theoretical predictive distributions, generally assume that the data loss mechanism is missing at random, a less strict assumption. But such methods may generate inaccurate results if the data loss pattern is systematic. If so, then (1) there is no statistical “fix” for the problem, and (2) you need to explicitly qualify the interpretation of the results, given the data loss pattern.
- If the choice of a method to handle missing data or outliers makes a difference in the results, then report those different findings, not just the ones that more closely favor your model. Doing so makes it plain that the results depend on how these common problems in data collection and analysis are managed. This can be an interesting result by itself.
- Verify distributional assumptions of your estimation method, such as multivariate normality for continuous endogenous variables. For example, report values of the skew index and kurtosis index for all continuous outcome variables. Also verify that relations between continuous variables are linear. Curvilinear relations between a causal variable and an outcome variable are no special problem, if (1) the researcher detects them and (2) includes the appropriate power terms in the analysis.
- Report sufficient descriptive statistics—including means, standard deviations, and correlations for continuous variables—so that another researcher could perform a secondary analysis based on your data summaries (e.g., someone else can verify your results). To save space, reliability coefficients and values of skew or kurtosis indexes can be reported in the same place. Give the final sample size in this summary.
- Even better, make the raw data file accessible to other researchers after removing confidential information about cases or settings. This option may be best for SEM analyses that require raw data, such as when the data are ordinal or when using a special estimation method for incomplete data files that corrects for non-normality in continuous outcomes.
- Clearly state the type of data matrix analyzed, which is ordinarily a covariance matrix for continuous data or a matrix of polychoric correlations, along with a matrix of asymptotic covariances for ordinal data or thresholds. If just a correlation matrix is analyzed for continuous data, then use an appropriate estimation method that is intended for analyzing correlation structures.

- Verify that your data matrix is positive definite.
- If the data are nested, such as repeated measures or collected in a complex sampling design, explain how nonindependence of the scores was taken into account (e.g., correlated disturbances are specified for repeated measures variables, a two-level model was analyzed).

## ESTIMATION

Undetected problems at earlier stages may make the problems in the analysis considered next more likely to happen:

- State which SEM computer tool was used (and its version), and list the syntax for your final model in an appendix. If the latter is not feasible due to length limitations, tell your readers how they can access your code (e.g., a website address).
- Specify the estimation method used, even if it is default maximum likelihood. If a different method is used, then clearly state this method and give your rationale for selecting it (e.g., the data are ordinal).
- Use an appropriate method for ordinal data, especially if the number of ordered categories is relatively small (e.g., < 6) and response distributions are asymmetrical. Robust weighted least squares is one option. Another is full information maximum likelihood with some type of method for numerical integration, but very large sample sizes may be needed.
- Carefully check your computer syntax, then check it again. Just as in data entry, it is easy to make an error in computer syntax that misspecifies the model, data, or analysis. Although SEM computer tools have become easier to use, they still cannot detect a mistake that is logical rather than a syntax error. A logical error does not cause the analysis to fail but instead results in an unintended specification, for instance,  $Y_1 \rightarrow Y_2$  is specified when  $Y_2 \rightarrow Y_1$  is intended. Carefully check to see that the model analyzed was actually the one that you intended to specify.
- Say whether estimation converged and whether the solution is admissible. Describe any complications, such as failure of iterative estimation or Heywood cases, and how such problems were handled (e.g., increasing the default limit on the number of iterations). Remember that SEM computer programs do not always print warning or error messages for inadmissible solutions, so you must carefully inspect the whole output. Likewise, do not interpret results from a solution that is not admissible as it is untrustworthy.
- Never retain a model based solely on global fit testing; specifically, do not rely on “golden rules” for approximate fit indexes to justify the retention of the model, especially if that model failed the chi-square test or the endogenous variables are not continuous.

- Always conduct local fit assessment; that is, inspect fit at a more molecular level by examining the residuals, including conditional independences, covariance residuals, correlation residuals, mean residuals, or threshold residuals. Treat significance tests of the residuals just mentioned, such as standardized residuals for covariance residuals, with caution. In smaller samples, such tests can fail to be significant even when the corresponding discrepancy between sample and predicted values is substantial. In very large samples, these significance tests can signal discrepancies that are trivial in magnitude.
- If alternative models are compared (whether nested or not nested), then state the decision rules used to select one model over another. Report the results of the chi-square difference test for relevant comparisons of hierarchical models. Remember that it is perfectly acceptable in SEM to retain no model, if there is no theoretically defensible respecification that leads to satisfactory model–data correspondence.
- In multiple-samples analyses, do not forget that equality constraints for the same parameter usually apply in the unstandardized solution only. It is expected that values of the standardized estimates for the same parameter will be different across the groups. Remember also that standardized estimates are, in general, not directly comparable across groups.
- Check for constraint interaction when testing for equality of unstandardized pattern coefficients across different factors or of direct effects on different endogenous variables. If the results of the chi-square difference test for the equality-constrained parameters depend on how the factors are scaled, there is constraint interaction. One option is to analyze a correlation matrix using constrained estimation, assuming that it makes sense to analyze standardized variables.
- When testing fully latent SR models, establish that the measurement model is consistent with the data before estimating versions with alternative structural models; that is, use two-step modeling, not one-step modeling.
- When testing latent growth models, first analyze a basic change model that includes just the repeated measures variables. Assuming that a change model is retained, next add predictors of change (covariates) to the model.
- Before formally comparing group means on observed variables, determine whether those scores measure the same latent variables in each group. In other words, test for strict invariance, which assumes equality of pattern of coefficients, intercepts or thresholds, and error variances and covariances over groups; otherwise, appreciable differences in the parameters just mentioned can confound group differences on the observed variables.
- In order to formally compare group means on latent variables, strong measurement invariance should be established. This is because appreciable group differences in

pattern coefficients or intercepts say that the indicators do not measure the factors in the same way across the groups. Formal comparison of groups on factor variances or covariances requires only weak invariance, or cross-group equality of the unstandardized pattern coefficients.

- As a relative novice to SEM, you should not be in the position of trying to analyze a model so complex that you are not certain whether it is identified or not identified. There are empirical tests for whether a converged, admissible solution is unique, but such tests are not foolproof. Failing an empirical check—for example, specifying different start values leads to a different solution—proves that the solution is not unique, but passing an empirical check does not prove that the model is really identified.
- Another challenge is empirical underidentification, which can occur due to data-related problems such as extreme collinearity or estimates of key parameters that are close to zero or nearly equal to each other. Measurement models where some factors have just two indicators may be especially susceptible to empirical underidentification. Respecification of a model when the data are the problem may lead to a specification error.

## RESPECIFICATION

Except when working in a strictly confirmatory mode, respecification is part of most SEM analyses. It is critical to get right the things considered next:

- Explain the theoretical basis for respecifying a model; that is, how are the changes justified? Indicate the particular statistics, such as correlation residuals, standardized residuals, or modification indexes, consulted in respecification and how the values relate to theory.
- Plainly differentiate between results from a priori specifications versus those found after fitting the model and otherwise examining the data. A specification search guided entirely by statistical criteria such as modification indexes is unlikely to lead to the correct model. Use your knowledge of theory and empirical results to inform the use of such statistics.
- Clearly state the nature and number of such respecifications such as, how many paths were added or dropped and which ones?
- If the final model is quite different from your initial model, reassure your readers that its specification was not merely the result of chasing sampling error. If there is no such rationale, the model may be overparameterized (good fit is achieved at the cost of too many parameters), and results from such models are unlikely to replicate. It is better to retain no model in this case.

## TABULATION

At the conclusion of the analysis, you must organize the statistical results so that they can be reported. Here are some suggestions for doing so in a clear and thorough way:

- Report the parameter estimates for your model (if a model is retained). This includes the unstandardized estimates, their standard errors, and the standardized estimates. Explain how the standardized solution was derived in a multiple-samples analysis (e.g., common metric vs. within-groups standardized solution). In analyses of latent variable models in a single sample, describe the particular standardized solution reported (e.g., just the factors are standardized vs. all variables are standardized).
- Do not indicate anything about statistical significance for the standardized parameter estimates unless you used a method, such as constrained estimation, or a computer tool that prints in the output correct standard errors for the standardized solution.
- For structural models, report an effect decomposition, which breaks down total effects into direct effects and total indirect effects. Estimate and interpret any individual indirect effects that are theoretically meaningful.
- Estimation of indirect (mediator) effects in the conventional way, or as products of the coefficients for the direct effects that make up the indirect pathway, assumes that the causal variable and the intervening (mediating) variable do not interact. If this assumption is not reasonable, then use an appropriate method, such as one that analyzes controlled direct effects, natural direct effects, and natural indirect effects (causal mediation analysis).
- Report information about the residuals, either in text, in a table, or in an appendix. *Show your readers the details about model fit.* Just reporting values of global fit statistics is inadequate.
- Report information for individual outcome variables about predictive power, such as  $R^2$  or a corrected- $R^2$  for endogenous variables in nonrecursive relations. Remember that  $R^2$  for an ordinal indicator in CFA applies to the corresponding latent response variable, and thus not directly to that indicator. Also remember that  $R^2$  for individual endogenous variables has nothing to do with global model fit. Interpret effect sizes (e.g., unstandardized or standardized path coefficients,  $R^2$ ) in reference to results expected in a particular research area.
- Always report the model chi-square and its degrees of freedom and  $p$  value. If the model fails the chi-square test, then explicitly state this result and tentatively reject the model. If possible, report the values of a minimal set of approximate fit indexes that include the RMSEA and its 90% confidence interval, Bentler CFI, and SRMR. Avoid selective reporting of the values of just those fit statistics that favor the model. If the model has a mean structure, explain the specification of the independence model (e.g.,

all means are assumed to equal zero vs. their sample values), if you report the CFI or a related type of incremental fit index.

## INTERPRETATION

Issues in the interpretation of SEM results for various kinds of effects and models are considered next:

- Comment on whether the signs and magnitudes of the parameter estimates make theoretical sense. Look for “surprises” that may indicate suppression or other unexpected results.
- Do not make hair-splitting distinctions among  $p$  values from significance tests for indirect effects. These tests, including the Sobel test, may be inaccurate because they make assumptions that are usually untenable. Bootstrapped significance tests for indirect effects may also be inaccurate, especially if the sample size is not large. Rely more on whether the magnitudes of indirect effects are substantively meaningful, given the research context.
- Do not confuse statistical significance with effect size or whether results are clinically, theoretically, or practically significant. Be careful not to commit one of many kinds of cognitive errors about statistical significance (e.g., the false belief that “significant” results are not due to chance). Do not be dazzled by asterisks (i.e., statistical significance), for they do not light the path to truth in SEM—or in any other kind of statistical analysis.
- Do not refer to indirect effects as “mediation” unless your research design includes time precedence between a causal variable, a presumed mediator, and an outcome variable. If the causal variable is experimental but the mediator is an individual difference variable, be especially wary that omitted common causes of the mediator and the outcome could bias the results.
- Do not automatically interpret “closer to fit” as “closer to truth.” Close model–data correspondence could reflect any of the following (not all mutually exclusive) possibilities: (1) the model accurately reflects reality; (2) the model is an equivalent or near-equivalent version of one that corresponds to reality but itself is incorrect; (3) the model fits the data in a nonrepresentative sample but has poor fit in the population; or (4) the model has so many freely estimated parameters that it cannot have poor fit even if it were grossly misspecified. In a single study, it is usually impossible to determine which of these scenarios explains the acceptable fit of the researcher’s model. If the analysis is never replicated, then we will never know. This is another way of saying that SEM is more useful for rejecting a false model than for somehow “confirming” whether a given model is actually true, especially without replication. For the same reasons, close fit to the data does not “prove” the directionality specifications (causal effects) represented in the model.

- Do not commit the naming fallacy, or the false belief that naming a factor means that it is understood. Factor names are not explanations. For example, if a three-factor CFA model fits the data, this does not prove that the verbal labels assigned by the researcher to the factors are correct. Alternative explanations of factors are often possible in many, if not most, factor analyses. Do not reify factors, or believe that constructs in your model *must* correspond to things in the real world. Perhaps they do, but do not assume it.
- If there are appreciable interaction effects, then explain their patterns. For example, by *how much* does a moderator variable change the association between two other variables? Report effects sizes for interaction effects, not just whether they are statistically significant or not.

## AVOID CONFIRMATION BIAS

Perhaps the most serious form of confirmation bias in SEM involves the failure to address the existence of equivalent or near-equivalent models:

- Explicitly acknowledge the issue of equivalent models. Generate some plausible equivalent versions of your final model and give reasons why your preferred model should be favored over those equivalent versions. Without such arguments, there can be no preference.
- It may also be possible to consider alternative models that are not equivalent but are based on the same variables and fitted to the same data matrix. Among alternative models that are near equivalent, give reasons why your model should be preferred.
- If you compare the relative fits of alternative-but-not-nested models with predictive fit indexes, such as the AIC or BIC, do not forget that the particular rank order indicated by the statistic is subject to sampling error; that is, the model preferred by the index may not be the “real” model in the population. The amount of this sampling error in model selection also *increases* along with the sample size instead of getting smaller. These problems explain why replication is a gold standard in science, not statistical prediction about the model that is most likely to replicate in hypothetical future studies.

## BOTTOM LINES AND STATISTICAL BEAUTY

The points summarized next deal with the role of SEM as a tool for science:

- The technique of SEM is about testing theories, not just models. The model analyzed represents predictions based on a particular body of work, but outside of this role, the model has little intrinsic value. This means that it provides a vehicle for testing

ideas, and the real goal of SEM is to evaluate these ideas in a meaningful and valid way. Whether or not a model is retained is incidental to this purpose.

- If no model is retained, then explain the implications for theory. For example, in what way(s) could theory be incorrect, based on your results?
- If a model is retained, explain to your readers just what was learned as a result of your study; that is, what is the *substantive significance* of your findings? How has the state of knowledge in your area been advanced? What new questions are posed? What comes next?
- If your sample is not large enough to randomly split and cross-validate your analyses, then clearly state this as a limitation. If so, replication is a necessary “what comes next” activity.
- A strong analytical method such as SEM cannot compensate for poor study design or slipshod ideas. For example, expressing poorly thought out hypotheses in a path diagram does not give them credibility. The specification of direct or indirect effects cannot be viewed as a substitute for an experimental or longitudinal design. Inclusion of an error term for a test with poor psychometrics cannot somehow transform it into a good measure. Applying SEM in the absence of good design, measures, and ideas is like using a chain saw to slice butter: one will accomplish the task, but without a more substantial base, one is just as likely to make a mess.

## SUMMARY

So concludes this journey of discovery about SEM. As on any guided tour, you may have found some places along the way more interesting than others. You may decide to revisit certain sites by using particular techniques in your own work. In any event, I hope that reading this book has given you new ways of looking at your data and hypotheses. Use SEM to address new questions or to provide new perspectives on older ones, but use it guided by good sense and strong domain knowledge. Use it also as a way to reform methods of data analysis by focusing more on models instead of specific effects analyzed with traditional significance tests. As Garrison Keillor says at the conclusion of *The Writer's Almanac*, the long-running radio program about poetry and literature: Be well, do good work, and keep in touch.

## LEARN MORE

McCoach, Black, and O'Connell (2007) outline sources of inference error, Tomarken and Waller (2005) survey common misunderstandings, and Tu (2009) addresses the use of SEM in epidemiology and reminds us of its limitations.

- McCoach, D. B., Black, A. C., & O'Connell, A. A. (2007). Errors of inference in structural equation modeling. *Psychology in the Schools*, 44, 461–470.
- Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, 1, 31–65.
- Tu, Y.-K. (2009). Commentary: Is structural equation modelling a step forward for epidemiologists? *International Journal of Epidemiology*, 38, 549–551.

## Suggested Answers to Exercises

### CHAPTER 2

- Given the descriptive statistics and with slight rounding error:

$$B_X = .686 \left( \frac{10.870}{3.007} \right) = 2.479$$

$$A_X = 102.950 - 2.479 (16.900) = 61.054$$

- Given  $M_X = 16.900$ , mean-centered scores ( $x$ ) are

-.90, -2.90, -.90, -4.90, 1.10,  
1.10, -3.90, -.90, 1.10, 5.10,  
1.10, 2.10, -.90, -.90, 5.10,  
-4.90, 3.10, -2.90, 4.10, .10

and  $M_x = 0$ ,  $SD_x = 3.007$ ,  $r_{xy} = .686$ , so with slight rounding error

$$B_X = .686 \left( \frac{10.870}{3.007} \right) = 2.479$$

$$A_x = 102.950 - 2.479 (0) = 102.950$$

- Given  $\hat{Y} = 2.479X + 61.054$ , the predicted scores  $\hat{Y}$  are

100.719, 95.761, 100.719, 90.803, 105.677,  
105.677, 93.282, 100.719, 105.677, 115.593,  
105.677, 108.156, 100.719, 100.719, 115.593,  
90.803, 110.635, 95.761, 113.114, 103.198

and the residual scores  $\hat{Y} - Y$  are

$$\begin{aligned} & -.719, -3.761, -12.719, 4.197, -7.677, \\ & -4.677, 3.718, -2.719, 4.323, 8.407, \\ & -3.677, 6.844, -8.719, 1.281, -11.593, \\ & -5.803, 7.365, 9.239, -2.114, 18.802 \end{aligned}$$

With slight rounding error,

$$\begin{aligned} s_Y^2 &= s_{\hat{Y}}^2 + s_{Y-\hat{Y}}^2 = 55.570 + 62.586 = 118.155 \\ r_{XY}^2 &= s_Y^2/s_{\hat{Y}}^2 = 55.570/118.155 = .470, \text{ so } r_{XY} = .686 \end{aligned}$$

4. Given the descriptive statistics and with slight rounding error:

$$b_X = \frac{.686 - .499(.272)}{1 - .272^2} = .594 \quad \text{and} \quad B_X = .594 \left( \frac{10.870}{3.007} \right) = 2.147$$

$$b_W = \frac{.499 - .686(.272)}{1 - .272^2} = .337 \quad \text{and} \quad B_W = .337 \left( \frac{10.870}{2.817} \right) = 1.302$$

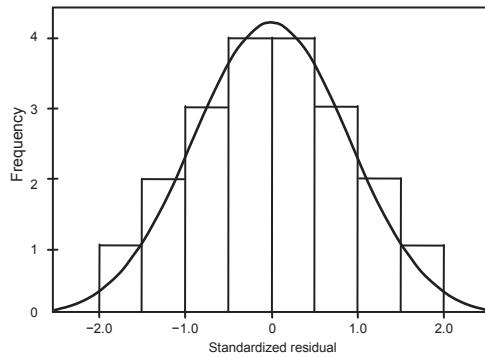
$$A_{X,W} = 102.950 - 2.147(16.900) - 1.302(49.400) = 2.340$$

$$R_{Y,X,W}^2 = .595(.686) + .337(.499) = .576$$

5. For  $N = 20$ ,  $k = 2$  and  $R_{Y,X,W}^2 = .576$ :

$$\hat{R}_{Y,X,W}^2 = 1 - (1 - .576) \left( \frac{20 - 1}{20 - 2 - 1} \right) = .526$$

6. Presented next is the distribution of standardized residuals for the regression of  $Y$  on both  $X$  and  $W$  generated in SPSS with a superimposed normal curve:



7. For  $r_{XY} = .686$ ,  $r_{WY} = .499$ ,  $r_{XW} = .272$ , and  $R_{Y,X,W}^2 = .576$  with slight rounding error:

$$r_{Y(W \cdot X)}^2 = .576 - .686^2 = \frac{(.499 - .686(.272))^2}{1 - .272^2} = .105$$

$$r_{WY \cdot X}^2 = \frac{.576 - .686^2}{1 - .686^2} = \frac{(.499 - .686(.272))^2}{(1 - .272^2)(1 - .686^2)} = .199$$

Respectively, variable  $W$  uniquely explains about 10.5% of the total variance in  $Y$ , and of variance in  $Y$  not already explained by  $X$ , predictor  $W$  accounts for about 19.9% of the rest.

## CHAPTER 3

Comments about the selected quotes:

1. Representativeness is determined by how cases are selected, which has nothing to do with statistical significance. If “reliability” means “repeatability,” then statistical significance does not directly indicate the likelihood of replication. But if “reliability” means “sampling error,” then, yes, there is less sampling error over larger random samples. Also,  $p$  is not the probability of error, which is virtually 1.0 for sample results, and neither is  $p$  the probability that the null hypothesis is true.
2. This is a statement of the odds against chance fallacy. A  $p$  value does not indicate the likelihood that a particular result is due to chance, nor does  $1 - p$  measure the probability that the data are due to any “real” effect. All sample results are affected by error.
3. The probability of sampling error is virtually 1.0 and thus cannot be specified in advance. The level of  $\alpha$  is specified by the researcher in advance, but there is actually no requirement to specify an arbitrary criterion level of statistical significance. The rest of the quote is correct, including the claim that significance testing assumes random sampling.
4. I used the NDC calculator for this problem. We can say that  $F(2, 47) = 31.925$  falls at
  - a. 97.5th percentile in the noncentral  $F(2, 47, 28.573)$  distribution; and the same observed  $F$  falls at the
  - b. 2.5th percentile in the noncentral  $F(2, 17, 109.201)$  distribution.

So the 95% confidence interval for  $\lambda$  is [28.573, 109.201]. Using Equation 3.4 to convert the lower and upper bounds of this interval to  $\rho^2$  units for  $N = 50$  gives us the noncentral 95% confidence interval based on  $R^2 = .576$ , which is [.364, .686]. As expected, this interval is narrower than the corresponding interval based on  $N = 20$ , which is [.173, .722].

## CHAPTER 4

1. There is slight rounding error in these calculations based on the statistics in Table 4.1:

$$s_x^2 = 3.0070^2 = 9.0422, s_w^2 = 2.8172^2 = 7.9366, \text{ and } s_y^2 = 10.8699^2 = 118.0895$$

$$\text{cov}_{xw} = .2721 (3.0070) (2.8172) = 2.3050$$

$$\text{cov}_{xy} = .6858 (3.0070) (10.8699) = 22.4159$$

$$\text{cov}_{wy} = .4991 (2.8172) (10.8699) = 15.2838$$

2. Given  $\text{cov}_{xy} = 13.00$ ,  $s_x^2 = 12.00$ , and  $s_y^2 = 10.00$ , the covariance is

$$\text{cov}_{xy} = r_{xy} \sqrt{12.00 \times 10.00} = r_{xy} (10.9545) = 13.00$$

Solving for the correlation gives a value that is out of bounds:

$$r_{xy} = 13.00 / 10.9545 = 1.19$$

3. For the data in Figure 4.2,  $\hat{\gamma}_1 = 3.10$  and  $\hat{\gamma}_2 = 15.73$ . Before applying a transformation to these data, add the constant  $-9.0$  to each score so that the lowest score is  $1.0$ . For a square root transformation,  $\hat{\gamma}_1 = 2.31$  and  $\hat{\gamma}_2 = 9.95$ . Even greater reduction in non-normality is afforded by the transformation  $\ln X$ , for which  $\hat{\gamma}_1 = 1.655$  and  $\hat{\gamma}_2 = 5.788$ .
4. The covariance matrix with effective sample sizes derived using pairwise deletion for the data in Table 4.3 is presented next:

		X	Y	W
X	cov	86.400	-22.500	15.900
	N	6	4	5
Y	cov	-22.500	8.200	-9.667
	N	4	5	4
W	cov	15.900	-9.667	5.200
	N	5	4	6

I submitted the whole covariance matrix (without the sample sizes) to an online matrix calculator. The eigenvalues are  $(95.937, 7.074, -3.211)$ , and the determinant is  $-2,178.864$ . The matrix is clearly nonpositive definite. The correlation matrix implied by the covariance matrix for pairwise deletion is presented next in lower diagonal form:

	X	Y	W
X	1.00		
Y	-.85	1.00	
W	.75	-1.48	1.00

5. Presented next for these five items are their intercorrelations at four-decimal accuracy shown without 1.0s in the diagonal:

	I1	I2	I3	I4	I5
I1	—				
I2	.3333	—			
I3	.1491	.1491	—		
I4	.3333	.3333	.1491	—	
I5	.3333	.3333	.1491	.3333	—

Calculations for  $\alpha_C$  are as follows:

$$\bar{r}_{ij} = \frac{6 (.3333) + 4 (.1491)}{10} = .2596 \quad \text{and} \quad \alpha_C = \frac{5 (.2596)}{1+(5-1).2596} = .6368$$

which, within the limits of rounding error, is equivalent to  $\alpha_C = .63$  computed by the Reliability Analysis procedure of SPSS for these data.

## CHAPTER 6

1. One option is to specify a set of two dummy codes,  $d_1$  and  $d_2$ , to represent both degrees of freedom for group membership, as follows:

Group	$d_1$	$d_2$
1	1	0
2	0	1
3	0	0

Code  $d_1$  specifies the contrast of groups 1 and 3, and  $d_2$  specifies the contrast of groups 2 and 3. Specify  $d_1$  and  $d_2$  as a pair of correlated exogenous variables in a path model.

2. Increasing the measurement error in  $Y$  of Figure 6.3 would decrease the score reliability coefficient,  $r_{YY}$ , increase the disturbance variance, and decrease  $R^2$ .
3. Coefficient  $a$  in Figure 6.4(a) for the model  $X \rightarrow Y$  and coefficient  $b$  in Figure 6.4(b) for the model  $Y \rightarrow X$  in unstandardized form are just different rearrangements of the elements in

$$\text{cov}_{XY} = r_{XY} SD_X SD_Y$$

When regressing  $Y$  on  $X$ , the unstandardized coefficient is

$$r_{XY} (SD_Y / SD_X)$$

and when regressing  $X$  on  $Y$ , the unstandardized coefficient is

$$r_{XY} (SD_X / SD_Y)$$

In standardized form, the coefficient for both  $X \rightarrow Y$  and  $Y \rightarrow X$  is  $r_{XY}$ , so in this way the path models in Figures 6.4(a) and 6.4(b) are clearly equivalent.

4. In Figure 6.6(b), there are four direct effects on endogenous variables. There are a total of four exogenous variables, two measured ( $X_1, X_2$ ) and two unmeasured ( $D_1, D_2$ ). The variances (4) and covariances (2) of all the variables just listed are free parameters. The total number of free parameters for Figure 6.6(b) is 10. There are three direct effects on endogenous variables in Figure 6.6(c), four variances of exogenous variables, and two covariances between pairs of exogenous variables for a total of nine free parameters.
5. Part of the relation of  $Y_1$  and  $Y_2$  in Figure 6.6(c) is causal due to the direct effect of the former variable on the latter. There are also two aspects of their association that are spurious: (1) they share at least one common but measured cause and (2) their direct causes,  $X_1$  and  $X_2$ , are correlated. Variable  $X_1$  indirectly affects  $Y_2$  through the intermediary  $Y_1$ .
6. With 6 observed variables, there are  $6(7)/2$ , or 21 observations. There are a total of six direct effects on endogenous variables. Additional free parameters include six variances and three covariances between pairs of exogenous variables, so the total number is  $6 + 6 + 3$ , or 15. Thus,  $df_M = 21 - 15 = 6$ .

**CHAPTER 7**

- For Figure 7.1(a), the number of observations is  $4(5)/2 = 10$ . Free parameters include a total of 4 variances (of  $X_1$ ,  $X_2$ ,  $D_1$ , and  $D_2$ ), 2 covariances ( $X_1 \curvearrowright X_2$  and  $D_1 \curvearrowright D_2$ ), and 4 direct effects on variables  $Y_1$  and  $Y_2$  from other measured variables for a total of 10, so  $df_M = 0$ . For Figure 7.1(b), the number of observations is  $6(7)/2 = 21$ . There are a total of 18 parameters: 6 variances (of  $X_1-X_3$  and  $D_1-D_3$ ), 6 covariances, and 6 direct effects on  $Y_1-Y_3$ , so  $df_M = 3$ .
- The number of endogenous variables in Figure 7.1(a) is 2, so the rank of the system matrix for each equation must be at least  $2 - 1 = 1$ .

Evaluation for  $Y_1$ :

$$\blacktriangleright \begin{array}{l} Y_1 \\ Y_2 \end{array} \left[ \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \\ \pm & \theta & \pm & \pm \\ \theta & 1 & \pm & \pm \end{array} \right] \rightarrow \left[ \begin{array}{c} 1 \end{array} \right] \rightarrow \text{Rank} = 1$$

Evaluation for  $Y_2$ :

$$\blacktriangleright \begin{array}{l} Y_1 \\ Y_2 \end{array} \left[ \begin{array}{cccc} X_1 & X_2 & Y_1 & Y_2 \\ 1 & \theta & \pm & \pm \\ \theta & \pm & \pm & \pm \end{array} \right] \rightarrow \left[ \begin{array}{c} 1 \end{array} \right] \rightarrow \text{Rank} = 1$$

Both equations pass the rank condition, so the model is identified.

- There are no variables excluded from the equation for  $Y_1$  in Figure 7.4(b), so the order condition is failed. Evaluation of the rank condition follows:

Evaluation for  $Y_1$ :

$$\blacktriangleright \begin{array}{l} Y_1 \\ Y_2 \end{array} \left[ \begin{array}{ccc} X & Y_1 & Y_2 \\ \pm & \pm & \pm \\ \theta & \pm & \pm \end{array} \right] \rightarrow \left[ \begin{array}{c} \end{array} \right] \rightarrow \text{Rank} = 0$$

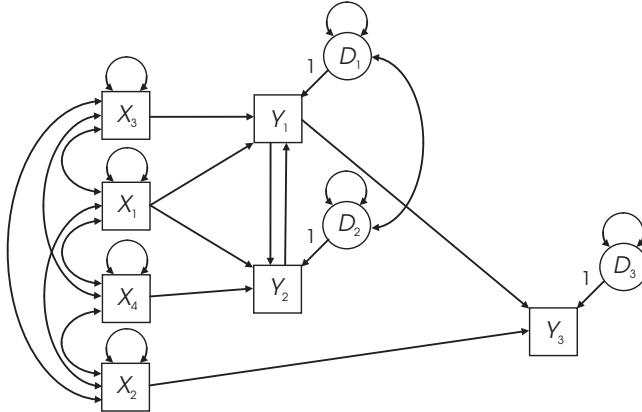
Evaluation for  $Y_2$ :

$$\blacktriangleright \begin{array}{l} Y_1 \\ Y_2 \end{array} \left[ \begin{array}{ccc} X & Y_1 & Y_2 \\ 1 & \pm & \pm \\ \theta & \pm & \pm \end{array} \right] \rightarrow \left[ \begin{array}{c} 1 \end{array} \right] \rightarrow \text{Rank} = 1$$

The equation for  $Y_1$  fails the rank condition, but Figure 7.4(b) is nevertheless identified.

- The model in Figure 7.4(b) with no disturbance correlation is identified but ( $df_M = 0$ ). In order to estimate the disturbance correlation, a unique instrument would be needed for  $Y_2$ . The respecified model just described corresponds to Figure 7.1(a). That model is also just-identified, but the  $z$  test for the disturbance covariance evaluates the hypothesis that the corresponding parameter equals zero.

5. Shown next is the revised version of Figure 7.4(a):



Because each variable in the feedback loop now has a unique instrument ( $X_3$  for  $Y_1$ ,  $X_4$  for  $Y_2$ ), their equations are identified. The recursive block with  $Y_3$  is also identified; thus, the respecified model is identified.

## CHAPTER 8

- The graph  $Y \rightarrow W \rightarrow X$  implies the same conditional independence, or  $X \perp Y | W$ , as Figures 8.1(a) and 8.1(b), so all three graphs are d-separation equivalent.
- Figure 8.2(c) implies the 24 conditional independences listed next:

Nonadjacent pair	Conditional independences	
$X, C$	$X \perp C   B$	$X \perp C   (B, A)$
	$X \perp C   (B, Y)$	$X \perp Y   (B, A, Y)$
$X, Y$	$X \perp Y   B$	$X \perp Y   (B, C)$
	$X \perp Y   (B, A)$	$X \perp Y   (B, C, A)$
$A, B$	$A \perp B   X$	$A \perp B   (X, C)$
	$A \perp B   (X, Y)$	$A \perp B   (X, C, Y)$
$A, C$	$A \perp C   B$	$A \perp C   X$
	$A \perp C   (B, X)$	$A \perp C   (B, Y)$
	$A \perp C   (X, Y)$	$A \perp C   (B, X, Y)$
$A, Y$	$A \perp Y   B$	$A \perp Y   X$
	$A \perp Y   (B, X)$	$A \perp Y   (B, C)$
	$A \perp Y   (X, C)$	$A \perp Y   (B, C, X)$

3. There are five nonadjacent pairs of variables in Figure 8.2(c) that can be d-separated, so the size of the basis set is 5. Listed next are the parents of each nonadjacent pair and the associated conditional independences of the basis set:

Nonadjacent pair	Parents	Conditional independence
$X, C$	$B$	$X \perp C \mid B$
$X, Y$	$B, C$	$X \perp Y \mid (B, C)$
$A, B$	$X$	$A \perp B \mid X$
$A, C$	$B, X$	$A \perp C \mid (B, X)$
$A, Y$	$B, C, X$	$A \perp Y \mid (B, C, X)$

4. Listed next are the four paths between  $X_2$  and  $Y_1$  in Figure 8.3(b):

$$\begin{aligned} & X_2 \rightarrow Y_2 \rightarrow Y_3 \leftarrow Y_1 \\ & X_2 \rightarrow Y_2 \leftarrow U_3 \rightarrow Y_1 \\ & X_2 \leftarrow U_1 \rightarrow X_1 \rightarrow Y_1 \\ & X_2 \leftarrow U_1 \rightarrow X_1 \leftarrow U_2 \rightarrow Y_1 \end{aligned}$$

The first, second, and fourth paths just listed are blocked by colliders, but the third path is open. Conditioning on  $X_1$  would close the third path, but doing so would open the fourth path, where  $X_1$  is a collider. Controlling also for  $U_1$  would close the fourth path opened by controlling for  $X_1$  alone, but  $U_1$  is unmeasured, so it cannot be part of any conditioning set.

5. In Figure 8.4(a), there are two back-door paths between the pair  $D$  and  $Y$ , including

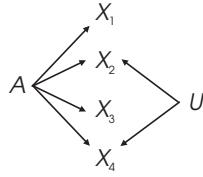
$$\begin{aligned} & D \leftarrow A \rightarrow X \rightarrow E \rightarrow Y \\ & D \leftarrow A \rightarrow X \leftarrow C \rightarrow Y \end{aligned}$$

The second path just listed is blocked by the collider  $X$ . The set  $(A)$  is sufficient because it would block the first back-door path just listed but leave the second back-door path blocked. This set is also minimally sufficient because its only proper subset is the empty set  $\emptyset$ , which is not sufficient itself. The set  $(C, X)$  is also sufficient because conditioning on this set would also close the open back-door path but does not open the blocked back-door path. This set is also minimally sufficient because neither  $(C)$  nor  $(X)$  is sufficient to close both paths.

6. The sets  $(A, X)$ ,  $(D, E)$ , and  $(D, X)$  each d-separate variables  $C$  and  $Y$  in a modified version of Figure 8.4(a) where the path  $C \rightarrow Y$  is deleted. Each set just listed is also a minimally sufficient set that identifies the direct effect of  $C$  on  $Y$ . There are other sufficient sets that identify the same direct effect, such as  $(A, D, X)$ , but none are minimally sufficient.

**CHAPTER 9**

1. Presented next is Figure 9.1(b) respecified as a causal DAG:



Because  $U$  is not a substantive latent variable, it cannot appear in any conditioning set. Thus, the back-door path with  $U$  as a common cause of  $X_2$  and  $X_4$  cannot be blocked by conditioning, so the figure implies:

$$\begin{array}{lll} X_1 \perp X_2 \mid A & X_1 \perp X_3 \mid A & X_1 \perp X_4 \mid A \\ X_2 \perp X_3 \mid A & X_3 \perp X_4 \mid A & \end{array}$$

2. If we constrain the average unstandardized pattern coefficient for the indicators of factor  $B$  in Figure 9.4, we obtain the equation

$$\frac{\lambda_{42} + \lambda_{52} + \lambda_{62}}{3} = 1.0$$

which implies all three relations listed next:

$$\lambda_{42} = 3 - \lambda_{52} - \lambda_{62}, \quad \lambda_{52} = 3 - \lambda_{42} - \lambda_{62}, \quad \text{and} \quad \lambda_{62} = 3 - \lambda_{42} - \lambda_{52}$$

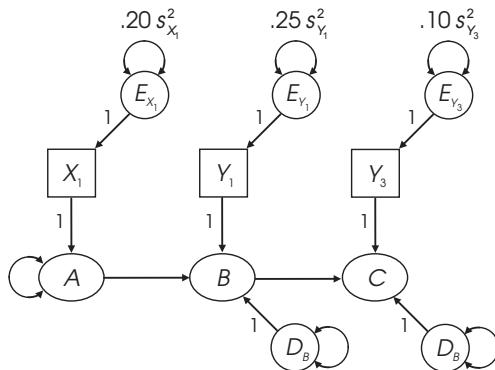
Select any of the formulas just listed and represent the corresponding linear constraint in the syntax of an SEM computer tool in order to scale factor  $B$  using the effects coding method.

3. With two indicators in Figure 9.5(a), there are  $2(3)/2 = 3$  observations. There are four free parameters, including the variances of 3 exogenous variables (of  $A$  and  $E_1-E_2$ ) and 1 pattern coefficient for  $X_2$ , so  $df_M = 3 - 4 = -1$ .
4. The number of observations for Figure 9.5(b) is  $3(4)/2 = 6$ . Free parameters include the variances of 4 exogenous variables (of  $A$  and  $E_1-E_3$ ) and 2 pattern coefficients for  $X_2-X_3$ , so  $df_M = 6 - 6 = 0$ .
5. There are  $4(5)/2 = 10$  observations for Figure 9.5(c). The total number of free parameters is 9, including 6 variances (of  $A$ ,  $B$ ,  $E_1-E_4$ ), 1 factor covariance, and 2 pattern coefficients for indicators  $X_2$  and  $X_4$ , so  $df_M = 10 - 9 = 1$ .
6. Yes, the model in Figure 9.6(f) but with the error correlation  $E_3 \curvearrowleft E_5$  identified. This is because the respecified model satisfies Rule 9.4 in Table 9.2. Specifically, the respecified model satisfies Rule 9.3 (and by implication Rule 9.2; see Table 9.1), and there is at least one simple indicator of both factors  $A$  and  $B$  with which the complex indicator  $X_3$  does not share an error correlation (e.g.,  $X_2 \rightarrow A$ ,  $X_4 \rightarrow B$ ).
7. Figure 9.7 is a standard CFA model that satisfies Rule 9.1 concerning the minimum number of indicators for each factor. Each factor and error term is scaled through a ULI constraint.

With 8 indicators, there are  $8(9)/2 = 36$  observations available to estimate a total of 17 free parameters, including 10 variances of exogenous variables (of the two factors and 8 error terms), 1 factor covariance, and a total of 6 pattern coefficients (for all indicators except the marker variables), so  $df_M = 36 - 17 = 19$ . Given all these facts, Figure 9.7 is identified.

## CHAPTER 10

- There are  $6(7)/2 = 21$  observations for Figure 10.1(b). Free parameters include the variances of 9 exogenous variables (of  $A$ , errors for  $X_1-X_2$  and  $Y_1-Y_4$ , and disturbances for factors  $B$  and  $C$ ) and 5 direct effects on endogenous variables ( $X_2$ ,  $Y_2$ ,  $Y_4$ ,  $B$ , and  $C$ ) for a total of 14. Thus,  $df_M = 21 - 14 = 7$ .
- There are  $5(6)/2 = 15$  observations for both Figures 10.2(a) and 10.3(a). Free parameters in Figure 10.2(a) include the variances of 7 exogenous variables (of  $X_1$ , errors for  $Y_1-Y_4$ , disturbances for factors  $B$  and  $C$ ) and a total of 4 direct effects on endogenous variables ( $Y_2$ ,  $Y_4$ ,  $B$ , and  $C$ ), so  $df_M = 15 - 11 = 4$ . Free parameters in Figure 11.3(a) include the variances of 7 exogenous variables (of  $A$ , errors for  $Y_1-Y_4$ , disturbances for factors  $B$  and  $C$ ) and the same 4 direct effects on endogenous variables, so  $df_M = 15 - 11 = 4$ .
- Given reliabilities of .80, .75, and .90 for, respectively,  $X_1$ ,  $Y_1$ , and  $Y_3$  in Figure 10.1(a), we specify the SR model that controls for measurement error shown next:



- There are  $12(13)/2 = 78$  observations for Figure 10.6. Free parameters include the variances of 16 exogenous variables (of constructive thinking, 12 error terms for the indicators, and 3 disturbances for endogenous factors) and the coefficients for 12 direct effects on endogenous variables (including 8 pattern coefficients [2 per factor] in the measurement model and 4 direct effects in the structural model) for a total of 28 free parameters. Thus,  $df_M = 78 - 28 = 50$ .
- For the continuance organizational commitment indicator in Figure 10.7,  $s^2 = .98^2 = .9604$  and  $r_{XX} = .70$  (Table 10.1). Thus, the unstandardized error variance is fixed to equal

$$.30 (.9604) = .2881$$

6. With 8 observed variables in Figure 10.7, there are a total  $8(9)/2$ , or 36 observations available to estimate the 32 free parameters, including 8 variances (of 6 exogenous factors and 2 disturbances), 16 covariances (1 disturbance covariance and 15 covariances among the 6 exogenous factors), and 8 direct effects in the structural model, so  $df_M = 36 - 32 = 4$ .

## CHAPTER 11

1. Given the results in Table 11.2, a 1-point increase in hardiness predicts a decrease in stress of .203 points. This amount of change is statistically significant because

$$z = -.203/.045 = -4.51, p < .01$$

An increase in hardiness of a full standard deviation predicts a decrease in stress of .230 standard deviations.

2. A total of three different sufficient sets, (Exercise), (Hardiness), and (Fitness), satisfy Rule 8.4 and thus identify the direct effect of stress on illness. Given results in Table 11.2, the three different estimators for stress range from .574 to .628 in the unstandardized solution and from .307 to .337 in the standardized solution. Now we consider only results that control for both parents of illness: A 1-point increase in stress predicts an increase in illness of .574 points, and a one standard deviation increase in stress leads to an increase in illness of .307 standard deviations, both controlling for fitness.
3. Given the results in Table 11.3, fitness and stress explain .177 of the total variance for illness, so the standardized disturbance variance is  $1 - .177$ , or .823. The observed variance is 3,903.75, so the unstandardized disturbance variance is .823 (3,903.95), or 3,212.786.
4. Given the results in Figure 11.1 for the indirect effect of hardiness on illness through stress, the unstandardized estimate is  $-.203$  (.574), or  $-.117$ , so a 1-point increase in hardiness leads to a decrease in illness of .117 points through the intervening variable of stress. The standardized estimate is  $-.230$  (.307), or  $-.071$ , so an increase in hardiness of one standard deviation predicts a decrease in illness of .071 standard deviations through the intervening variable of stress.
5. The product estimates of the unstandardized and standardized indirect effects of hardiness on illness through stress are, respectively,  $-.117$  and  $-.071$ . Given the results in Table 11.2, the approximate standard error of the unstandardized product estimator is

$$SE_{ab} = \sqrt{.574^2(.045^2) + (-.203)^2.089^2} = .032$$

(see Table 11.4) and  $z = -.117/.032 = 3.71, p < .05$ . The other two estimates in Table 11.4 of the same indirect effect are based on covariate adjustment, and both are statistically significant. Over all three results, the standardized estimates range from  $-.163$  to  $-.071$ , and the unstandardized estimates range from  $-.267$  to  $-.117$ .

6. There is a single noncausal path between fitness and stress (Figure 7.5); it is

Fitness  $\longleftrightarrow$  Exercise  $\curvearrowleft$  Hardiness  $\rightarrow$  Stress

The product of the standardized coefficients (Table 11.5) for this path is

$$.390 (-.030) (-.230) = .003$$

so the predicted correlation is .003. The observed correlation is -.130 (Table 4.2), so the correlation residual equals  $-.130 - .003$ , or  $-.133$ .

## CHAPTER 12

- Given the results in Table 12.1, the RMSEA is calculated as follows:

$$\hat{\Delta}_M = 11.107 - 5 = 6.107, N = 373, df_M = 5$$

$$RMSEA = \sqrt{\frac{6.107}{5(372)}} = .057$$

- The CFI is calculated from the results in Table 12.1 as

$$\hat{\Delta}_M = 11.107 - 5 = 6.107, \hat{\Delta}_B = 172.289 - 10 = 162.289$$

$$CFI = 1 - \frac{6.107}{162.289} = .962$$

- In Table 12.1, SRMR = .051. The average absolute correlation residual excluding the diagonal entries in the top part of Table 11.8 is

$$(.057 + .015 + .082 + .092 + .133 + .041 + .033)/10 = .045$$

The value of the SRMR varies with that of the average absolute correlation residual, but the two may not be exactly equal.

- Summarized next are values of fit statistics generated by the student version of LISREL when fitting the model in Figure 7.5 to the data in Table 4.2 at two different sample sizes:

Statistic	<u>N = 373</u>	<u>N = 5,000</u>
$\chi^2_M(5)$	11.107, $p = .049$	148.894, $p < .001$
RMSEA [90% CI]	.057 [.003, .103]	.076 [.066, .087]
CFI	.962	.937
SRMR	.051	.051

As expected, the value of the model chi-square is greater and that of the corresponding  $p$  value is smaller in the analysis in the larger sample. The value of RMSEA is somewhat higher in the larger sample, but it is more precise (its 90% confidence interval is narrower). The value of CFI is also somewhat worse in the larger sample, owing to an increase in the model chi-square, but the value of the SRMR is unchanged.

5. The scaled chi-square difference statistic is calculated for these data as follows:

$$df_D = 17 - 12 = 5$$

$$\chi^2_D(5) = 57.50 - 18.10 = 39.40$$

$$c_1 = \frac{57.50}{28.35} = 2.028 \quad \text{and} \quad c_2 = \frac{18.10}{11.55} = 1.567$$

$$\hat{\chi}^2_D(5) = \frac{39.40}{[2.028(17) - 1.567(12)]/5} = \frac{39.40}{3.134} = 12.57, p = .028$$

6. For both models,  $N = 469$ . Given the data in Table 12.4 for the *psychosomatic model*:

$$\chi^2_M(5) = 40.402, q = 10$$

$$AIC_1 = 40.402 + 2(10) = 60.402$$

$$BIC = 40.401 + 10 [\ln (469)] = 101.908$$

and for the *conventional medical model*:

$$\chi^2_M(3) = 3.238, q = 12$$

$$AIC_1 = 3.238 + 2(12) = 27.238$$

$$BIC = 3.238 + 12 [\ln (469)] = 77.045$$

## CHAPTER 13

- With 8 indicators, there are  $8(9)/2 = 36$  observations. If we assume that the Hand Movements task is the reference variable, then free parameters for a single-factor model include 7 pattern coefficients, 8 residual variances, and the factor variance for a total of 16. Thus,  $df_M = 36 - 16 = 20$ .
- Listed next in the lower part of the matrix are the standardized residuals for the Mplus analysis of single-factor model of the KABC-I. Values that are statistically significant at the .05 level are shown in boldface. Reported in the upper part of the matrix are correlation residuals computed in EQS. Absolute values  $> .10$  are shown in boldface. These results indicate many problems with local fit:

Indicator	1	2	3	4	5	6	7	8
1. Hand	—	<b>.101</b>	.047	-.056	-.069	.028	.046	-.034
2. Number	<b>2.062</b>	—	<b>.397</b>	<b>-.130</b>	-.081	-.045	.010	-.092
3. Word	1.026	<b>6.218</b>	—	-.091	-.077	-.071	-.025	-.030
4. Gestalt	-1.231	<b>-2.727</b>	-1.953	—	.057	-.009	.025	.068
5. Triangles	<b>-2.201</b>	<b>-2.364</b>	<b>-2.355</b>	1.378	—	.019	.003	.066
6. Spatial	.723	-1.188	<b>-1.996</b>	-.210	.595	—	.011	.018
7. Matrix	1.086	.236	-.601	.544	.088	.313	—	-.034
8. Photo	-1.240	<b>-3.420</b>	-1.037	1.833	<b>2.178</b>	.675	-.036	—

3. Structure coefficients calculated with slight rounding error (see Table 13.4) are shown next:

<u>Indicator</u>	<u>Simultaneous</u>	<u>Indicator</u>	<u>Sequential</u>
HM	.497 (.557) = .277	GC	.503 (.557) = .280
NR	.807 (.557) = .449	Tr	.726 (.557) = .404
WO	.808 (.557) = .450	SM	.656 (.557) = .365
		MA	.588 (.557) = .328
		PS	.782 (.557) = .436

4. Given the results in Table 13.3, the CR coefficient for the simultaneous processing factor is calculated as follows:

$$\sum \hat{\lambda}_i = (1.000 + 1.445 + 2.029 + 1.212 + 1.727) = 7.413$$

$$\hat{\phi} = 1.835$$

$$\sum \hat{\theta}_{ii} = (5.419 + 3.425 + 9.998 + 5.104 + 3.483) = 27.429$$

$$CR = \frac{7.413^2(1.835)}{7.413^2(1.835) + 27.429} = .786$$

5. The threshold  $\hat{\tau}_2 = .25$  for the item illustrated in Figure 13.5 is the value of the normal deviate that corresponds to the 60th percentile in a normal distribution. It marks the point on the continuous variable  $X^*$  where the responses on  $X$  shift from “2” for *neutral* to “3” for *agree*.

## CHAPTER 14

1. Results of the  $z$  test for the factor covariances in Table 16.4 are summarized next:

<u>Covariance</u>	<u>Test</u>
Constructive ↗ Dysfunctional	$z = -.028/.017 = -1.65, p = .099$
Constructive ↗ Subjective Well-Being	$z = .024/.014 = 1.17, p = .242$
Constructive ↗ Job Satisfaction	$z = .060/.029 = 2.07, p = .039$
Dysfunctional ↗ Subjective Well-Being	$z = -.088/.017 = -5.18, p < .001$
Dysfunctional ↗ Job Satisfaction	$z = -.132/.030 = -4.40, p < .001$
Subjective Well-Being ↗ Job Satisfaction	$z = .139/.027 = 5.15, p < .001$

2. Listed next is the standardized effect decomposition for Figure 14.2. The total indirect effect of constructive thinking on job satisfaction consists of three indirect pathways:

Endogenous	Effect	Constructive	Dysfunctional	Subjective
Dysfunctional	Direct	-.124	—	—
	Total indirect	—	—	—
	Total	-.124	—	—
Subjective	Direct	.082	-.470	—
	Total indirect	-.058	—	—
	Total	.024	-.470	—
Satisfaction	Direct	.093	-.149	.382
	Total indirect	.072	-.180	—
	Total	.165	-.329	.382

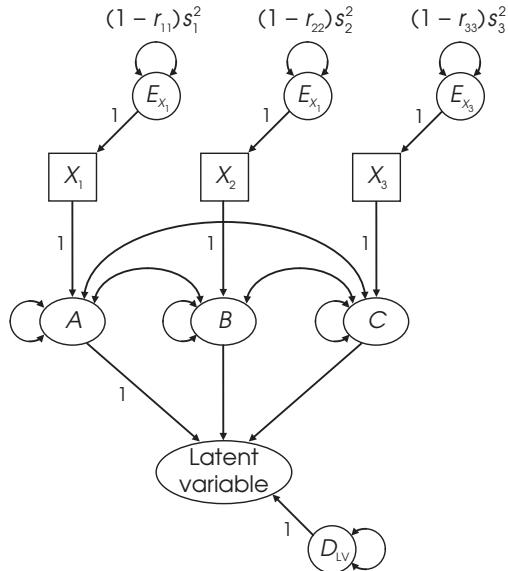
3. Presented next is lavaan syntax that imposes the equality constraint by assigning the same label to both direct effects (see Figure 10.7):

$$\text{OrgTI} \sim a * \text{OccTI}$$

$$\text{OccTI} \sim a * \text{OrgTI}$$

In the reanalysis with the equality constraint,  $\chi^2_M(4) = 2.221$ ,  $p = .695$ , and both unstandardized reciprocal direct effects equal .155. But the standardized direct effect of OccTI on OrgTI is .169, and the standardized direct effect of OrgTI on OccTI is .142.

4. A diagram for the respecification of Figure 14.3(b) but with three indicators where  $r_{11}$ ,  $r_{22}$ , and  $r_{33}$  are reliability coefficients and  $s_1^2$ ,  $s_2^2$ , and  $s_3^2$  are the sample variances for the causal indicators is presented next. This model is not identified in isolation, but it shows how to control for measurement error at the indicator level for a formative factor:



5. There are  $9(10)/2 = 45$  observations for Figure 14.4. Free parameters include 12 variances (of 3 causal indicators, 3 disturbances, and 6 errors for effect indicators), three covariances (among the 3 causal indicators), two direct effects in the structural model, and a total of six pattern coefficients in the measurement model for a total of 23, so  $df_M = 45 - 23 = 22$ .

## CHAPTER 15

1. In Figure 15.3, the total effect of the constant on the fourth trial is

$$\begin{aligned}\Delta \rightarrow \text{Initial} \rightarrow \text{Trial 4} &= \kappa_{In} \times 1.0 = \kappa_{In} \\ \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 4} &= \kappa_{Sh} \times \lambda_4\end{aligned}$$

The predicted mean for the fourth trial equals  $\kappa_{In} + \lambda_4 (\kappa_{Sh})$ , where  $\lambda_4$  indicates the proportion of the average improvement over the first two trials that must be added to the initial mean in order to generate the predicted mean for the fourth trial.

2. From Table 15.5,  $\hat{\lambda}_5 = 2.171$ , so the predicted mean on trial 5 equals the sum of the estimated mean for the Initial factor, or 11.763, and 2.171 times the estimated mean improvement from trial 1 to trial 2, or 9.597. The result  $\hat{\lambda}_6 = 2.323$  for trial 6 has the same interpretation except that the proportion of the improvement in performance over trials 1–2 is 2.323.
3. The observed means on trials 4–6 are, respectively, 31.02, 32.58, and 34.20 (Table 15.3). Given the results in Table 15.5, the predicted means are calculated as follows:

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 4} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 3} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 3} \\ &= 11.763 (1.0) + 9.597 (2.015) = 31.101\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 5} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 5} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 5} \\ &= 11.763 (1.0) + 9.597 (2.171) = 32.598\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 6} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 6} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 6} \\ &= 11.763 (1.0) + 9.597 (2.323) = 34.057\end{aligned}$$

4. In symbolic form, the predicted mean of the Initial factor in Figure 15.5 is

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Initial} &= \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} + \\ &\quad \Delta \rightarrow \text{Initial} \\ &= \kappa_{Ab} (\gamma_{In}) + \alpha_{In}\end{aligned}$$

In words, the expected average of this factor is the sum of (1) the mean of the ability variable ( $\kappa_{Ab}$ ) times the unstandardized coefficient for the regression of the Initial factor on ability ( $\gamma_{In}$ ) and (2) the unstandardized intercept for the regression analysis just described, or  $\alpha_{In}$ .

5. The observed means for trials 1–2 and 4–6 are, respectively, 11.77, 21.39, 31.02, 32.58, and 34.20 (Table 15.3). Given the results in Table 15.6, predicted means for trials 1–2 and 4–6 are calculated as follows:

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 1} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 1} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} \rightarrow \text{Trial 1} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 1} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} \rightarrow \text{Trial 1} \\ &= 11.287 (1.0) + .700 (.678) (1.0) + \\ &\quad 9.608 (0) + .700 (-.096) (0) \\ &= 11.762\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 2} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 2} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} \rightarrow \text{Trial 2} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 2} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} \rightarrow \text{Trial 2} \\ &= 11.287 (1.0) + .700 (.678) (1.0) + \\ &\quad 9.608 (1.0) + .700 (-.096) (1.0) \\ &= 21.302\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 4} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 4} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} \rightarrow \text{Trial 4} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 4} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} \rightarrow \text{Trial 4} \\ &= 11.287 (1.0) + .700 (.678) (1.0) + \\ &\quad 9.608 (2.027) + .700 (-.096) (2.027) \\ &= 31.101\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 5} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 5} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} \rightarrow \text{Trial 5} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 5} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} \rightarrow \text{Trial 5} \\ &= 11.287 (1.0) + .700 (.678) (1.0) + \\ &\quad 9.608 (2.185) + .700 (-.096) (2.185) \\ &= 32.608\end{aligned}$$

$$\begin{aligned}\text{Total effect of } \Delta \text{ on Trial 6} &= \Delta \rightarrow \text{Initial} \rightarrow \text{Trial 6} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Initial} \rightarrow \text{Trial 6} + \\ &\quad \Delta \rightarrow \text{Shape} \rightarrow \text{Trial 6} + \\ &\quad \Delta \rightarrow \text{Ability} \rightarrow \text{Shape} \rightarrow \text{Trial 6} \\ &= 11.287 (1.0) + .700 (.678) (1.0) + \\ &\quad 9.608 (2.338) + .700 (-.096) (2.338) \\ &= 34.068\end{aligned}$$

**CHAPTER 16**

- Suppose that the model in Figure 16.1 is evaluated over samples drawn from two different populations. The population means on the single factor are  $\kappa_1$  and  $\kappa_2$ . We do not assume that  $\kappa_1 = \kappa_2$ . Suppose that  $\lambda_{1A}$  and  $v_{1A}$  are, respectively, the unstandardized pattern coefficient and the intercept for regressing  $X_1$  on the common factor in population A. The terms  $\lambda_{1B}$  and  $v_{1B}$  represent the corresponding quantities in population B. Given the model, the mean on  $X_1$  can be expressed in both populations as follows:

$$\begin{aligned}\mu_{1A} &= \kappa_1 (\lambda_{1A}) + v_{1A} \\ \mu_{1B} &= \kappa_2 (\lambda_{1B}) + v_{1B}\end{aligned}$$

Unless both  $\lambda_{1A} = \lambda_{1B}$  and  $v_{1A} = v_{1B}$ , the contrast  $\mu_{1A} - \mu_{1B}$  will reflect something other than the difference between the factor means,  $\kappa_1$  and  $\kappa_2$ . That is, any differences in unstandardized pattern coefficients or intercepts are confounded with the factor mean contrast.

- Given Figure 16.1, these unstandardized results in Table 16.4 are freely estimated in both samples:

$$\begin{aligned}\hat{\kappa}_{\text{Eng}} &= 1.843, \hat{v}_{1\text{Eng}} = .323, \hat{v}_{2\text{Eng}} = .346 \\ \hat{\kappa}_{\text{Spa}} &= 1.532, \hat{v}_{1\text{Spa}} = .486, \hat{v}_{2\text{Spa}} = .202\end{aligned}$$

These unstandardized estimates in Tables 16.3 and 16.4 are equal across both samples:

$$\begin{aligned}\hat{\lambda}_1 &= 1.062, \hat{\lambda}_2 = .635, \hat{\lambda}_3 = 1.144, \hat{\lambda}_4 = .988, \hat{\lambda}_5 = 1.171 \\ \hat{v}_3 &= -.051, \hat{v}_4 = -.144, \hat{v}_5 = -.474\end{aligned}$$

Using Equation 16.1, the predicted indicator means for both samples are calculated as follows:

$$\begin{array}{ll}\hat{\mu}_{1\text{Eng}} = 1.843 (1.062) + .323 = 2.280 & \hat{\mu}_{1\text{Spa}} = 1.532 (1.062) + .486 = 2.113 \\ \hat{\mu}_{2\text{Eng}} = 1.843 (.635) + .346 = 1.516 & \hat{\mu}_{2\text{Spa}} = 1.532 (.635) + .202 = 1.175 \\ \hat{\mu}_{3\text{Eng}} = 1.843 (1.144) - .051 = 2.057 & \hat{\mu}_{3\text{Spa}} = 1.532 (1.144) - .051 = 1.702 \\ \hat{\mu}_{4\text{Eng}} = 1.843 (.988) - .144 = 1.677 & \hat{\mu}_{4\text{Spa}} = 1.532 (.988) - .144 = 1.370 \\ \hat{\mu}_{5\text{Eng}} = 1.843 (1.171) - .474 = 1.684 & \hat{\mu}_{5\text{Spa}} = 1.532 (1.171) - .474 = 1.320\end{array}$$

These predicted means are similar to the observed means in both samples (see Table 16.1).

- Given  $\hat{\kappa}_{\text{Eng}} = 1.843$ ,  $\hat{\sigma}_{\text{Eng}}^2 = .412$ ,  $n_1 = 193$ ,  $\hat{\kappa}_{\text{Spa}} = 1.532$ ,  $\hat{\sigma}_{\text{Spa}}^2 = .235$ , and  $n_2 = 257$ , the Welch–James test is calculated with slight rounding error as follows:

$$\begin{aligned}t(df_{WJ}) &= \frac{1.843 - 1.532}{\hat{\sigma}_{WJ}} \\ df_{WJ} &= \frac{\left(\frac{.412}{193} + \frac{.235}{257}\right)^2}{\frac{(.412)^2}{193^2(192)} + \frac{(.235)^2}{257^2(256)}} = 344.33\end{aligned}$$

$$\hat{\sigma}_{WJ} = \sqrt{\frac{.412}{193} + \frac{.235}{257}} = .0552$$

$$t(344.33) = \frac{.311}{.0552} = 5.63, p < .001$$

4. Thresholds for  $X_1$  in the white sample are .772, 1.420, and 1.874 (Table 16.5), and percentile equivalents in the normal curve are, respectively, 77.99, 92.22, and 96.95. Thresholds in the African American sample for  $X_1$  are .674, 1.487, and 1.849, and percentile equivalents in the normal curve are, respectively, 74.98, 93.15, and 96.78. Proportions of responses in each category (0, 1, 2, 3) for each sample are listed next:

Sample	0	1	2	3
White	.7799	.1423	.0443	.0335
African American	.7498	.1817	.0363	.0322

5. The  $R^2$  values reported next apply to the latent response variables, not to the original items:

Variable	White	African American
$X_1^*$	$1 - .634 = .366$	$1 - .631 = .369$
$X_2^*$	$1 - .587 = .413$	$1 - .584 = .416$
$X_3^*$	$1 - .372 = .628$	$1 - .672 = .328$
$X_4^*$	$1 - .630 = .370$	$1 - .627 = .373$
$X_5^*$	$1 - .410 = .590$	$1 - .407 = .593$

## CHAPTER 17

1. Listed next are the centered scores on the predictors, their product, and scores on the criterion for the data in Table 17.1:

x	w	xw	y
-5.125	-3.375	17.2969	5
-1.125	-1.375	1.5469	9
.875	-.375	-.3281	11
3.875	-3.375	-13.0781	11
-3.125	10.625	-33.2031	11
-.125	5.625	-.7031	10
.875	4.625	4.0469	7
3.875	11.625	45.0469	5

The unstandardized regression equation for predicting  $Y$  from  $x$  and  $w$  is

$$\hat{Y} = .112x - .064w + 8.625$$

(i.e., Equation 17.2) where the regression coefficients for the centered predictors  $x$  and  $w$  are the same as those for the original (not centered) predictors (see Equation 17.1).

2. A moderator changes the relation between two variables, either dampening or strengthening that association, depending on the level of the moderator. In causal modeling, a moderator

changes how one variable affects a third variable. Moderation is symmetrical, so the role of moderator versus cause (focal variable) can be exchanged. A mediator is an intervening variable that transmits the causal effect of one variable to a third variable. A mediator must be a variable that can change as a result of another variable's influence on it before changes in the mediator affect the outcome (Little, 2013). Mediation is *not* symmetrical; that is, the role of the cause and mediator cannot be exchanged. This is because mediation implies a specific causal order (an indirect effect). Both mediators and moderators are causal variables in that each is assumed to affect the outcome. In causal mediation analysis, the causal variable and mediator are assumed to interact. In this case, the variable that is a mediator is also a moderator.

3. Bivariate correlations among predictors and product terms for the original scores versus centered scores in Table 17.1 are listed next:

	$X$	$W$	$XW$		$x$	$w$	$xw$
$X$	—			$x$	—		
$W$	.156	—		$w$	.156	—	
$XW$	.747	.706	—	$xw$	.284	.113	—

Centering reduces correlations between the predictors and the corresponding product term (e.g., .747 vs. .284). This reduction reflects nonessential multicollinearity due to the scales of  $X$  and  $W$ .

4. For the data in Table 17.1, the unstandardized equation for regressing  $XW$  on  $X$  and  $W$  is

$$\hat{Y}_{XW} = 15.3372X + 7.3892W - 111.0248$$

Residuals are calculated as

$$XW_{\text{res}} = XW - \hat{Y}_{XW}$$

and are displayed next for individual cases:

$XW$	Predicted $XW$	$XW_{\text{res}}$
20	-6.4589	26.4589
72	69.6681	2.3319
104	107.7317	-3.7317
110	131.5758	-21.5758
96	127.6636	-31.6636
133	136.7294	-3.7294
144	144.6774	-.6774
275	242.4130	32.5870

You should verify that the bivariate correlations between  $XW_{\text{res}}$  and each of  $X$  and  $W$  are zero within slight rounding error. The equation for regressing  $Y$  on  $X$ ,  $W$ , and  $XW_{\text{res}}$  is

$$\hat{Y} = .112X - .064W - .108XW_{\text{res}} + 8.873$$

where the coefficients for  $X$ ,  $W$ , and the intercept equal their counterparts in Equation 17.1 with no product term. The coefficients for  $X$  and  $W$  in their analysis with  $XW_{\text{res}}$  estimate the *unconditional* linear relations of each predictor to  $Y$  controlling for the other predictor.

# References

- Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, 22, 327–339.
- Acock, A. C. (2013). *Discovering structural equation modeling using Stata 13*. College Station, TX: Stata Press.
- Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: Wiley.
- Aguinis, H. (1995). Statistical power with moderated multiple regression in management research. *Journal of Management*, 21, 1141–1158.
- Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, 13, 515–539.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112, 545–557.
- American Psychological Association Publication and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839–851.
- Amos Development Corporation. (1983–2013). *IBM SPSS Amos* (Version 22.0) [computer software]. Meadville, PA: Author.
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411–423.
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalivé, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21, 1086–1120.
- Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, 23, 321–327.
- Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, 21, 495–508.
- Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92–108). New York: Guilford Press.
- Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 625–666). Charlotte, NC: IAP.

- 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.
- Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, 42, 815–824.
- Bartholomew, D. J. (2002). Old and new approaches to latent variable modeling. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 1–13). Mahwah, NJ: Erlbaum.
- Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28, 135–167.
- Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory and Rasch models for speech–language pathologists. *American Journal of Speech–Language Pathology*, 20, 243–259.
- Beauducel, A., & Wittman, W. (2005). Simulation study on fit indices in confirmatory factor analysis based on data with slightly distorted simple structure. *Structural Equation Modeling*, 12, 41–75.
- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York: Routledge.
- Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31, 419–456.
- Bentler, P. M. (1987). Drug use and personality in adolescence and young adulthood: Structured models with nonnormal variables. *Child Development*, 58, 65–79.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler, P. M. (2000). Rites, wrongs, and gold in model testing. *Structural Equation Modeling*, 7, 82–91.
- Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, 74, 137–143.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–600.
- Bentler, P. M., & Raykov, T. (2000). On measures of explained variance in nonrecursive structural equation models. *Journal of Applied Psychology*, 85, 125–131.
- Benyamin, Y., Ein-Dor, T., Ginzburg, K., & Solomon, Z. (2009). Trajectories of self-rated health among veterans: A latent growth curve analysis of the impact of posttraumatic symptoms. *Psychosomatic Medicine*, 71, 345–352.
- Bergsma, W., Croon, M. A., & Hagenaars, J. A. (2009). *Marginal models: For dependent, clustered, and longitudinal categorical data*. New York: Springer.
- Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin*, 105, 467–477.
- Berry, W. D. (1984). *Nonrecursive causal models*. Beverly Hills, CA: Sage.
- Bishop, J., Geiser, C., & Cole, D. A. (2015). Modeling latent growth with multiple indicators: A comparison of three approaches. *Psychological Methods*, 20, 43–62.
- Blalock, H. M. (1961). Correlation and causality: The multivariate case. *Social Forces*, 39, 246–251.
- Blest, D. C. (2003). A new measure of kurtosis adjusted for skewness. *Australian and New Zealand Journal of Statistics*, 45, 175–179.
- Block, J. (1995). On the relation between IQ, impulsivity, and delinquency: Remarks on the Lynam, Moffitt, and Stouthamer-Loeber (1993) interpretation. *Journal of Abnormal Psychology*, 104, 395–398.
- Blunch, N. (2013). *Introduction to structural equation modeling using IBM SPSS Statistics and Amos* (2nd ed.). Thousand Oaks, CA: Sage.
- Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, 76, 306–317.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A. (1996). A limited-information estimator for LISREL models with and without heterosce-

- dastic errors. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 227–241). Mahwah, NJ: Erlbaum.
- Bollen, K. A. (2000). Modeling strategies: In search of the Holy Grail. *Structural Equation Modeling*, 7, 74–81.
- Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53, 605–634.
- Bollen, K. A. (2007). Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods*, 12, 219–228.
- Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38, 37–72.
- Bollen, K. A., & Bauldry, S. (2010). Model identification and computer algebra. *Sociological Methods and Research*, 39, 127–156.
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, 16, 265–284.
- 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*. Hoboken, NJ: Wiley.
- Bollen, K. A., & Hoyle, R. H. (2012). Latent variable models in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 56–67). New York: Guilford Press.
- Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (TSLS) and maximum likelihood (ML) estimators. *Sociological Methods and Research*, 36, 48–86.
- Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 301–328). New York: Springer.
- Bollen, K. A., & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 111–135). Newbury Park, CA: Sage.
- Bollen, K. A., & Ting, K. (1993). Confirmatory tetrad analysis. In P. M. Marsden (Ed.), *Sociological Methodology 1993* (pp. 147–175). Washington, DC: American Sociological Association.
- Boomsma, A., Hoyle, R. H., & Panter, A. T. (2012). The structural equation modeling research report. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 341–358). New York: Guilford Press.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791–799.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 26, 211–252.
- Breitling, L. P. (2010). dagR: A suite of R functions for directed acyclic graphs. *Epidemiology*, 21, 586–587.
- Breivik, E., & Olsson, U. H. (2001). Adding variables to improve fit: The effect of model size on fit assessment in LISREL. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 169–194). Lincolnwood, IL: Scientific Software International.
- Brito, C., & Pearl, J. (2002). A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9, 459–474.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: Guilford Press.
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.
- Browne, M. W. (1984). Asymptotically distribution-free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62–83.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

- Browne, M. W., & Du Toit, S. H. C. (1991). Models for learning data. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 47–68). Washington, DC: American Psychological Association.
- Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling*, 19, 372–398.
- Bryant, F. B., & Satorra, A. (2013). EXCEL macro file for conducting scaled difference chi-square tests via LISREL 8, LISREL 9, EQS, and Mplus. Retrieved from [www.econ.upf.edu/~satorra](http://www.econ.upf.edu/~satorra)
- Budtz-Jørgensen, E., Keiding, N., Grandjean, P., & Weihe, P. (2002). Estimation of health effects of prenatal methylmercury exposure using structural equation models. *Environmental Health: A Global Access Science Source*, 1(2). Retrieved from [www.ehjournal.net](http://www.ehjournal.net)
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, 98, 550–558.
- Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods and Research*, 5, 3–52.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.
- Byrne, B. M. (2010). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.
- Byrne, B. M. (2012a). Choosing structural equation modeling computer software: Snapshots of LISREL, EQS, Amos, and Mplus. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–324). New York: Guilford Press.
- Byrne, B. M. (2012b). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466.
- Byrne, D. (2009). *Bicycle diaries*. New York: Viking.
- Cameron, L. C., Ittenbach R. F., McGrew, K. S., Harrison, P., Taylor, L. R., & Hwang, Y. R. (1997). Confirmatory factor analysis of the K-ABC with gifted referrals. *Educational and Psychological Measurement*, 57, 823–840.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2014). *The Nineteenth Mental Measurements Yearbook*. Lincoln: Buros Institute of Mental Measurements, University of Nebraska.
- Chan, F., Lee, G. K., Lee, E.-J., Kubota, C., & Allen, C. A. (2007). Structural equation modeling in rehabilitation counseling research. *Rehabilitation Counseling Bulletin*, 51(1), 53–66.
- Chang, H.-T., Chi, N. W., & Miao, M. C. (2007). Testing the relationship between three-component organizational/occupational commitment and organizational/occupational turnover intention using a non-recursive model. *Journal of Vocational Behavior*, 70, 352–368.
- Chen, B., & Pearl, J. (2015). Graphical tools of linear structural equation modeling. Retrieved from [http://ftp.cs.ucla.edu/pub/stat\\_ser/r432.pdf](http://ftp.cs.ucla.edu/pub/stat_ser/r432.pdf)
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, 29, 468–508.
- Chen, F., Curran, P. J., Bollen, K. A., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research*, 36, 462–494.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464–504.
- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41, 189–225.
- Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-

- cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31, 187–212.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.
- Chin, W. W. (2001). *PLS-Graph user's guide*. Houston, TX: Soft Modeling.
- Chin, W. W., Peterson, R. A., & Brown, S. P. (2008). Structural equation modeling in marketing: Some practical reminders. *Journal of Marketing Theory and Practice*, 16, 287–298.
- Choi, J., Fan, W., & Hancock, G. R. (2009). A note on confidence intervals for two-group latent mean effect size measures. *Multivariate Behavioral Research*, 44, 396–406.
- Choi, Y. Y., Song, J.-I., Chun, J. S., Lee, K. O., & Song, W. K. (2013). A structural equation modeling approach for the estimation of genetic and environmental effects from twin fMRI data. *International Journal of Bioscience, Biochemistry and Bioinformatics*, 3, 167–169.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York: Routledge.
- Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, 12, 381–398.
- 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.
- Cornoni-Huntley, J., Barbano, H. E., Brody, J. A., Cohen, B., Feldman, J. J., Kleinman, J. C., et al. (1983). National Health and Nutrition Examination I—Epidemiologic followup survey. *Public Health Reports*, 98, 245–251.
- Crawford, J. R. (2007). SBIDIFF.EXE [computer software]. Retrieved from <http://homepages.abdn.ac.uk/j.crawford/pages/dept/sbendiff.htm>
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349–354.
- Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, 105, 317–327.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38, 529–569.
- Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods*, 12, 283–297.
- Curran, T., Hill, A. P., & Niemiec, C. P. (2013). A conditional process model of children's behavioral engagement and behavioral disaffection in sport based on self-determination theory. *Journal of Sport and Exercise Psychology*, 35, 30–43.
- Dawson, J. F., & Richter, A. W. (2006). Probing three-way interactions in moderated multiple regression: Development and application of a slope difference test. *Journal of Applied Psychology*, 91, 917–926.
- Deshon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, 46, 137–149.
- Diamantopoulos, A. (Ed.). (2008). Formative indicators [Special issue]. *Journal of Business Research*, 61(12).
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61, 1203–1218.
- Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL: A guide for the uninitiated*. Thousand Oaks, CA: Sage.
- Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38, 269–277.

- Dillman Carpentier, F. R., Mauricio, A. M., Gonzales, N. A., Millsap, R. E., Meza, C. M., Dumka, L. E., et al. (2008). Engaging Mexican origin families in a school-based preventive intervention. *Journal of Primary Prevention*, 28, 521–546.
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9, 327–346.
- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, 23, 225–241.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research and Evaluation*, 14(20). Retrieved from <http://pareonline.net/pdf/v14n20.pdf>
- Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, 74, 119–137.
- Duncan, O. D. (1975). *Introduction to structural equation models*. New York: Academic Press.
- Duncan, T. E., Duncan, S. C., Hops, H., & Alpert, A. (1997). Multi-level covariance structure analysis of intra-familial substance use. *Drug and Alcohol Dependence*, 46, 167–180.
- Duncan, T. E., Duncan, S. C., Strycker, L. A., Li, F., & Alpert, A. (1999). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum.
- 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.
- Edwards, J. R. (2009). Seven deadly myths of testing moderation in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 143–164). New York: Taylor & Francis.
- Edwards, J. R. (2010). The fallacy of formative measurement. *Organizational Research Methods*, 14, 370–388.
- Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods*, 12, 1–22.
- Edwards, M. C., Wirth, R. J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 195–208). New York: Guilford Press.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.
- Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait–multimethod data: Different models for different types of methods. *Psychological Methods*, 13, 230–253.
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York: Cambridge University Press.
- Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). New York: Springer.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.
- Epskamp, S. (2014). Package *semPlot*. Retrieved from <http://cran.r-project.org/web/packages/semPlot/semPlot.pdf>
- Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63, 591–601.
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. New York: Oxford University Press.
- Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do they have in common? *Structural Equation Modeling*, 4, 65–79.
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of the two-index strategy revisited. *Structural Equation Modeling*, 12, 343–367.
- Finkel, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, CA: Sage.
- Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation model-

- ing. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269–314). Greenwich, CT: IAP.
- 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: IAP.
- Flora, D. B. (2008). Specifying piecewise latent trajectory models for longitudinal data. *Structural Equation Modeling*, 15, 513–533.
- 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*, 16, 625–641.
- Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, 13, 465–486.
- Fox, J. (2012). Structural equation modeling in R with the sem package. Retrieved from <http://socserv.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-SEMs.pdf>
- Frees, E. W. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences*. New York: Cambridge University Press.
- Friendly, M. (2006). SAS macro programs: boxcox. Retrieved from [www.math.yorku.ca/SCS/sasmac/boxcox.html](http://www.math.yorku.ca/SCS/sasmac/boxcox.html)
- Friendly, M. (2009). SAS macro programs: csmpower. Retrieved from [www.datavis.ca/sasmac/csmpower.html](http://www.datavis.ca/sasmac/csmpower.html)
- Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York: Basic.
- Garson, G. D. (Ed.). (2013) *Hierarchical linear modeling: Guide and applications*. Thousand Oaks, CA: Sage.
- Geary, R. C. (1947). Testing for normality. *Biometrika*, 34, 209–242.
- Geiser, C. (2013). *Data analysis with Mplus*. New York: Guilford Press.
- George, R. (2006). A cross-domain analysis of change in students' attitudes toward science and attitudes about the utility of science. *International Journal of Science Education*, 28, 571–589.
- Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Newbury Park, CA: Sage.
- Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science Quarterly*, 50, 21–43.
- Glymour, C., Scheines, R., Spirtes, P., & Ramsey, J. (2014). TETRAD V [computer software]. Available from [www.phil.cmu.edu/tetrad/current.html](http://www.phil.cmu.edu/tetrad/current.html)
- Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In M. Oakes & J. Kaufman (Eds.), *Methods in social epidemiology* (pp. 387–422). San Francisco: Jossey-Bass.
- Gnamb, T. (2013). Required sample size and power for SEM. Retrieved from <http://timo.gnamb.at/en/scripts/powerforsem>
- Goldman, B. A., & Mitchell, D. F. (2007). *Directory of unpublished experimental mental measures* (vol. 9). Washington, DC: American Psychological Association.
- Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, 32, 252–286.
- Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, 6, 258–269.
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *Journal of Experimental Education*, 74, 251–266.
- Grace, J. B. (2006). *Structural equation modeling and natural systems*. New York: Cambridge University Press.
- Grace, J. B., & Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: The role of composite variables. *Environmental and Ecological Statistics*, 15, 191–213.

- Graham, J. M., Guthrie, A. C., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. *Structural Equation Modeling*, 10, 142–153.
- Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277–295). New York: Guilford Press.
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, 44(Suppl. 3), S78–S94.
- Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. New York: Cambridge University Press.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online*, 7(1), 1–17. Retrieved from [www.dgps.de/fachgruppen/methoden/mpr-online](http://www.dgps.de/fachgruppen/methoden/mpr-online)
- Hallquist, M., & Wiley, J. (2015). MplusAutomation: Automating Mplus model estimation and interpretation. R package version 0.6-3 [computer software]. Retrieved from <http://cran.r-project.org/web/packages/MplusAutomation>
- Hancock, G. R., & Freeman, M. J. (2001). Power and sample size for the Root Mean Square Error of Approximation of not close fit in structural equation modeling. *Educational and Psychological Measurement*, 61, 741–758.
- Hancock, G. R., & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 117–159). Charlotte, NC: IAP.
- Hancock, G. R., & Liu, M. (2012). Bootstrapping standard errors and data–model fit statistics in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 296–306). New York: Guilford Press.
- Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural Equation Modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.
- Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology*, 12(184). Retrieved from [www.biomedcentral.com/1471-2288/12/184](http://www.biomedcentral.com/1471-2288/12/184)
- Harrington, D. (2009). *Confirmatory factor analysis*. New York: Oxford University Press.
- Hayduk, L. A. (1996). *LISREL issues, debates and strategies*. Baltimore, MD: Johns Hopkins University Press.
- Hayduk, L. A. (2006). Blocked-error-R<sup>2</sup>: A conceptually improved definition of the proportion of explained variance in models containing loops or correlated residuals. *Quality and Quantity*, 40, 629–649.
- Hayduk, L. A. (2014a). Seeing perfectly-fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. *Educational and Psychological Measurement*, 74, 905–926.
- Hayduk, L. A. (2014b). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology*, 14(124). Retrieved from [www.biomedcentral.com/1471-2288/14/124](http://www.biomedcentral.com/1471-2288/14/124)
- Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulian, S. (2007). Testing! testing! one, two, three—Testing the theory in structural equation models! *Personality and Individual Differences*, 42, 841–850.
- Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., et al. (2003). Pearl's d-separation: One more step into causal thinking. *Structural Equation Modeling*, 10, 289–311.
- Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling*, 7, 1–35.
- Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, 12(159). Retrieved from [www.biomedcentral.com/1471-2288/12/159](http://www.biomedcentral.com/1471-2288/12/159)

- Hayduk, L. A., Pazderka-Robinson, H., Cummings, G. C., Levers, M.-J. D., & Beres, M. A. (2005). Structural equation model testing and the quality of natural killer cell activity measurements. *BMC Medical Research Methodology*, 5(1). Retrieved from [www.ncbi.nlm.nih.gov/pmc/articles/PMC546216](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC546216)
- Hayes, A. F. (2013a). Conditional process modeling: Using structural equation modeling to examine contingent causal processes. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 219–266). Greenwich, CT: IAP.
- Hayes, A. F. (2013b). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.
- Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, 41, 924–936.
- Henningsen, A., & Hamann, J. D. (2007). *systemfit: A package for estimating systems of simultaneous equations in R*. *Journal of Statistical Software*, 23(4). Retrieved from [www.jstatsoft.org/v23/i04/paper](http://www.jstatsoft.org/v23/i04/paper)
- Hershberger, S. L. (1994). The specification of equivalent models before the collection of data. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 68–105). Thousand Oaks, CA: Sage.
- Hershberger, S. L. (2006). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 13–41). Greenwich, CT: IAP.
- Hershberger, S. L., & Marcoulides, G. A. (2013). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 3–39). Charlotte, NC: IAP.
- Hicks, R., & Tingley, D. (2011). Causal mediation analysis. *Stata Journal*, 11, 605–619.
- Hirschfeld, G., & von Brachel, R. (2014). Multiple-group confirmatory factor analysis in R—A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research and Evaluation*, 19(7). Retrieved from <http://pareonline.net/getvn.asp?v=19&n=7>
- Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2013). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3. Retrieved from [www.frontiersin.org/article/10.3389/fpsyg.2012.00137/full](http://www.frontiersin.org/article/10.3389/fpsyg.2012.00137/full)
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.
- Holsta, K. K., & Budtz-Jørgensen, E. (2012). Linear latent variable models: The lava package. *Computational Statistics*, 28, 1385–1452.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117–144.
- Houghton, J. D., & Jinkerson, D. L. (2007). Constructive thought strategies and job satisfaction: A preliminary examination. *Journal of Business Psychology*, 22, 45–53.
- Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, 12, 205–218.
- Hoyle, R. H. (2012). Model specification in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 126–144). New York: Guilford Press.
- Hoyle, R. C., & Isherwood, J. C. (2011). Reporting results from structural equation modeling analyses in *Archives of Scientific Psychology*. *Archives of Scientific Psychology*, 1, 14–22.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Huck, S. W. (1992). Group heterogeneity and Pearson's *r*. *Educational and Psychological Measurement*, 52, 253–260.
- Hunter, J. E., & Gerbing, D. W. (1982). Unidimensional measurement, second order factor analysis, and causal models. *Research in Organizational Behavior*, 4, 267–320.
- Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theory framework and rise of the neoFisherian. *Annales Zoologici Fennici*, 46, 311–349.

- Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25, 51–71.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8): e124. Retrieved from [www.plosmedicine.org](http://www.plosmedicine.org)
- Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the  $N:q$  hypothesis. *Structural Equation Modeling*, 10, 128–141.
- Jackson, D. L., Gillaspy, J. A., Jr., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, 14, 6–23.
- James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38, 324–329.
- James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69, 307–321.
- Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, 30, 199–218.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Lang (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.
- Jöreskog, K. G. (2004). *On chi-squares for the independence model and fit measures in LISREL*. Retrieved from [www.ssicentral.com/lisrel/techdocs/ftb.pdf](http://www.ssicentral.com/lisrel/techdocs/ftb.pdf)
- Jöreskog, K. G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny–Judd model with interaction effects. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 57–88). Mahwah, NJ: Erlbaum.
- Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-participant designs. *Psychological Methods*, 6, 115–134.
- Jung, S. (2013). Structural equation modeling with small sample sizes using two-stage ridge least-squares estimation. *Behavior Research Methods*, 45, 75–81.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
- Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, 74, 657–690.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.
- Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: Sage.
- Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). New York: Guilford Press.
- Kaplan, D., Harik, P., & Hotchkiss, L. (2001). Cross-sectional estimation of dynamic structural equation models in disequilibrium. In R. Cudeck, S. Du Toit, and D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 315–339). Lincolnwood, IL: Scientific Software International.
- Karami, H. (2012). An introduction to differential item functioning. *International Journal of Educational and Psychological Assessment*, 11, 56–76.
- Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC administration and scoring manual*. Circle Pines, MN: American Guidance Service.
- Keith, T. Z. (1985). Questioning the K-ABC: What does it measure? *School Psychology Review*, 14, 9–20.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
- Kenny, D. A. (2011a). Estimation with instrumental variables. Retrieved from <http://davidakenny.net/cm/iv.htm>
- Kenny, D. A. (2011b). Terminology and basics of SEM. Retrieved from <http://davidakenny.net/cm/basics.htm>
- Kenny, D. (2013). Moderator variables: Introduction. Retrieved from <http://davidakenny.net/cm/moderation.htm>
- Kenny, D. A. (2014a). Measuring model fit. Retrieved from <http://davidakenny.net/cm/fit.htm>

- Kenny, D. A. (2014b). Mediation. Retrieved from <http://davidakenny.net/cm/mediate.htm#CI>
- Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201–210.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165–172.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, 4th ed., pp. 233–265). Boston: McGraw-Hill.
- Kenny, D. A., & Milan, S. (2012). Identification: A nontechnical discussion of a technical issue. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 145–163). New York: Guilford Press.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of education researchers: An analysis of the ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350–368.
- Khine, M. S. (Ed.). (2013). *Application of structural equation modeling in educational research and practice*. Rotterdam, The Netherlands: Sense.
- Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, 12, 368–390.
- Klein, A., & Moosbrugger, A. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65, 457–474.
- Klein, A. G., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, 42, 647–673.
- Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). New York: Guilford Press.
- Kline, R. B. (2013a). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.
- Kline, R. B. (2013b). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatsschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 171–207). New York: Routledge.
- Kline, R. B., Snyder, J., & Castellanos, M. (1996). Lessons from the Kaufman Assessment Battery for Children (K-ABC): Toward a new assessment model. *Psychological Assessment*, 8, 7–17.
- Knight, C. R., & Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 275–299). New York: Springer.
- Knüppel, S., & Stang, A. (2010). DAG program: Identifying minimal sufficient adjustment sets. *Epidemiology*, 21, 159.
- Kühnel, S. (2001). The didactical power of structural equation modeling. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 79–96). Lincolnwood, IL: Scientific Software International.
- Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory and Psychology*, 22, 67–90.
- Lance, C. E. (1988). Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interaction effects. *Applied Psychological Measurement*, 12, 163–175.
- Lee, S., & Hershberger, S. L. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research*, 25, 313–334.
- Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, 48, 339–358.
- Lei, P.-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–179). New York: Guilford Press.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York: Guilford Press.

- Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*, 13, 497–519.
- Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When “good” indicators are bad and “bad” indicators are good. *Psychological Methods*, 4, 192–211.
- 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.
- Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, NJ: Erlbaum.
- Lynam, D. R., Moffitt, T., & Stouthamer-Loeber, M. (1993). Explaining the relation between IQ and delinquency: Class, race, test motivation, or self-control? *Journal of Abnormal Psychology*, 102, 187–196.
- Maasen, G. H., & Bakker, A. B. (2001). Suppressor variables in path models: Definitions and interpretations. *Sociological Methods and Research*, 30, 241–270.
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100, 107–120.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51, 201–236.
- MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin*, 114, 533–541.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Erlbaum.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, 1, 173–181.
- MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, 19, 30–43.
- Madans, J. H., Kleinman, J. C., Cox, C. S., Barbano, H. E., Feldman, J. J., Cohen, B., et al. (1986). 10 years after NHANES I—Report of initial followup, 1982–84. *Public Health Reports*, 101, 465–473.
- Maddox, T. (2008). *Tests: A comprehensive reference for assessments in psychology, education and business* (6th ed.). Austin, TX: PRO-ED.
- Mair, P., Wu, E., & Bentler, P. M. (2010). EQS goes R: Simulations for SEM using the package REQS. *Structural Equation Modeling*, 17, 333–349.
- Malone, P. S., & Lubansky, J. B. (2012). Preparing data for structural equation modeling: Doing your homework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 263–276). New York: Guilford Press.
- Marcoulides, G. A., & Ing, M. (2012). Automated structural equation modeling strategies. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 690–704). New York: Guilford Press.
- Mardia, K. V. (1985). Mardia's test of multinormality. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 5, pp. 217–221). New York: Wiley.
- Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modeling. *Personality and Individual Differences*, 42, 851–858.
- Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analysis of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, 15, 47–70.
- Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumaker (Eds.), *Advanced structural equation modeling* (pp. 315–353). Mahwah, NJ: Erlbaum.

- Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling* (pp. 177–198). Thousand Oaks, CA: Sage.
- Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 252–284). Thousand Oaks, CA: Sage.
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, 11, 320–341.
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: Integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85–110.
- Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, 9, 275–300.
- Marsh, H. W., Wen, Z., & Hau, K. T. (2006). Structural equation modeling of latent interaction and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 225–265). Greenwich, CT: IAP.
- Marsh, H. W., Wen, Z., Nagengast, B., & Hau, K. T. (2012). Structural equation models of latent interaction. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 436–458). New York: Guilford Press.
- Masyn, K. E., Petras, H., & Liu, W. (2014). Growth curve models with categorical outcomes. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of Criminology and Criminal Justice* (pp. 2013–2025). New York: Springer Verlag.
- MathWorks. (2013). MATLAB (Version 8.2) [computer software]. Natick, MA: Author.
- Matsueda, R. L. (2012). Key advances in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 3–16). New York: Guilford Press.
- Mauro, R. (1990). Understanding L.O.V.E. (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, 108, 314–329.
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12, 23–44.
- 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.
- McCoach, D. B., Black, A. C., & O'Connell, A. A. (2007). Errors of inference in structural equation modeling. *Psychology in the Schools*, 44, 461–470.
- McDonald, R. A., Behson, S. J., & Seifert, C. F. (2005). Strategies for dealing with measurement error in multiple regression. *Journal of Academy of Business and Economics*, 5, 80–97.
- McDonald, R. P. (1989). An index of goodness of fit based on noncentrality. *Journal of Classification*, 6, 97–103.
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82.
- McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, 107, 247–255.
- McIntosh, C. N., Edwards, J. R., & Antonakis, J. (2014). Reflections on partial least squares path modeling. *Organizational Research Methods*, 17, 210–251.
- McKnight, P. E., McKnight, K. M., Sidani, S., & Figueiredo, A. J. (2007). *Missing data: A gentle introduction*. New York: Guilford Press.
- Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, 14, 611–635.
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93, 568–592.
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7, 361–388.

- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55, 107–122.
- Messick, S. (1995). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156–166.
- Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, 42, 869–874.
- Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, 8, 1–17.
- Millsap, R. E. (2007). Structural equation modeling made difficult. *Personality and Individual Differences*, 42, 875–881.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.
- Millsap, R. E., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380–392). New York: Guilford Press.
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, 39, 479–515.
- Molina, K. M., Alegría, M., & Mahalingam, R. (2013). A multiple-group path analysis of the role of everyday discrimination on self-rated physical health among Latina/os in the U.S. *Annals of Behavioral Medicine*, 45(1), 33–44.
- Monecke, A. (2014). Package semPLS [computer software]. Retrieved from <http://cran.r-project.org/web/packages/semPLS/semPLS.pdf>
- Mooijaart, A., & Satorra, A. (2009). On insensitivity of the chi-square model test to non-linear misspecification in structural equation models. *Psychometrika*, 74, 443–455.
- Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks, CA: Sage.
- Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioral Research*, 22, 267–305.
- Mulaik, S. A. (2009a). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Mulaik, S. A. (2009b). *Linear causal modeling with structural equations*. New York: CRC Press.
- Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling*, 7, 36–73.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115–132.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, 22, 376–398.
- Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides and R. E. Schumaker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahwah, NJ: Erlbaum.
- Muthén, B. O. (2011). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. Retrieved from [www.statmodel.com/download/causalmediation.pdf](http://www.statmodel.com/download/causalmediation.pdf)
- Muthén, B., & Asparouhov, T. (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling*, 22, 12–23.
- Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Retrieved from [www.statmodel.com/bmuthen/articles/Article\\_075.pdf](http://www.statmodel.com/bmuthen/articles/Article_075.pdf)
- Muthén, L. K., & Muthén, B. O. (1998–2014). Mplus (Version 7.3) [computer software]. Los Angeles: Author.
- Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, 5, 297–310.
- Narayanan, A. (2012). A review of eight software packages for structural equation modeling. *American Statistician*, 66, 129–138.

- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2004). *Mx: Statistical modeling* (6th ed.). Richmond: Virginia Commonwealth University, Virginia Institute for Psychiatric and Behavioral Genetics.
- Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *Journal of Experimental Education*, 68, 251–268.
- Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, 8, 353–377.
- Newsom, J. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York: Routledge.
- Nimon, K., & Reio, T., Jr. (2011). Measurement invariance: A foundational principle for quantitative theory building. *Human Resource Development Review*, 10, 198–214.
- Nunkoo, R., Ramkissoon, H., & Gursoy, D. (2013). Use of structural equation modeling in tourism research: Past, present, and future. *Journal of Travel Research*, 52, 759–771.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- O'Brien, R. M. (1994). Identification of simple measurement models with multiple latent variables and correlated errors. *Sociological Methodology*, 24, 137–170.
- Okech, D., Kim, J., & Little, T. D. (2013). Recent developments in structural equation modeling research in social work publications. *British Journal of Social Work*. Advance access publication. Retrieved from <http://bjsw.oxfordjournals.org>
- Oliveri, M. E., Olson, B. D., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, 12, 203–223.
- Olsson, U. H., Foss, T., & Breivik, E. (2004). Two equivalent discrepancy functions for maximum likelihood estimation: Do their test statistics follow a non-central chi-square distribution under model misspecification. *Sociological Methods and Research*, 32, 453–500.
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and non-normality. *Structural Equation Modeling*, 7, 557–595.
- O'Rourke, R., & Hatcher, L. (2013). *A step-by-step approach to using SAS for factor analysis and structural equation modeling* (2nd ed.). Cary, NC: SAS Institute.
- Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research and Evaluation*, 8(6). Retrieved from <http://pareonline.net/getvn.asp?v=8&n=6>
- Osborne, J. W. (2010). Improving your data transformations: Applying the Box–Cox transformation. *Practical Assessment, Research and Evaluation*, 15(12). Retrieved from <http://pareonline.net/getvn.asp?v=15&n=12>
- Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research and Evaluation*, 17. Retrieved from <http://pareonline.net/pdf/v17n15.pdf>
- Park, I., & Schutz, R. W. (2005). An introduction to latent growth models: Analysis of repeated measures physical performance data. *Research Quarterly for Exercise and Sport*, 76, 176–192.
- Paxton, P., Hipp, J. R., & Marquart-Pyatt, S. T. (2011). *Nonrecursive models: Endogeneity, reciprocal relationships, and feedback loops*. Thousand Oaks, CA: Sage.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.
- Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146.
- Pearl, J. (2009b). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.
- Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York: Guilford Press.
- Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, 19, 459–481.
- Pearl, J., & Meshkat, P. (1999). Testing regression models with fewer regressors. In D. Heckerman &

- J. Whittaker (Eds.), *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics* (pp. 255–259). San Francisco: Morgan Kaufmann.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Peters, C. L. O., & Enders, C. (2002). A primer for the estimation of structural equation models in the presence of missing data. *Journal of Targeting, Measurement and Analysis for Marketing*, 11, 81–95.
- Petersen, M. L., Sinisi, S. E., & van der Laan, M. J. (2006). Estimation of direct causal effects. *Epidemiology*, 17, 276–284.
- Ping, R. A. (1996). Interaction and quadratic effect estimation: A two-step technique using structural equation analysis. *Psychological Bulletin*, 119, 166–175.
- Pinter, J. (1996). Continuous global optimization software: A brief review. *Optima*, 52, 1–8.
- Pornprasertmanit, S., Miller, P., Schoemann, A., Quick, C., & Jorgensen, T. (2014). Package simsem. Retrieved from <http://cran.r-project.org/web/packages/simsem/simsem.pdf>
- Pornprasertmanit, S., Miller, P., Schoemann, A., Rosseel, Y., Quick, C., Garnier-Villarreal, M., et al. (2014). Package semTools. Retrieved from <http://cran.r-project.org/web/packages/semTools/semTools.pdf>
- Porter, K., Poole, D., Kisynski, J., Sueda, S., Knoll, B., Mackworth, A., et al. (1999–2009). Belief and Decision Network Tool (Version 5.1.10) [computer software]. Retrieved from <http://aispace.org/bayes>
- Preacher, K. J., & Coffman, D. L. (2006). Computing power and minimum sample size for RMSEA. Retrieved from [www.quantpsy.org/rmsea/rmsea.htm](http://www.quantpsy.org/rmsea/rmsea.htm)
- 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.
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16, 93–115.
- Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17, 1–14.
- Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42, 185–227.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.
- Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, 18, 161–182.
- Provalis Research. (1995–2011). SimStat (Version 2.6.1) [Computer software]. Montreal, Quebec, Canada: Author.
- Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general populations. *Applied Psychological Measurement*, 1, 385–401.
- Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163.
- Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy*, 35, 299–331.
- Raykov, T. (2011). *Introduction to psychometric theory*. New York: Routledge.
- Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure? *Structural Equation Modeling* 8, 142–149.
- Rensvold, R. B., & Cheung, G. W. (1999). Identification of influential cases in structural equation models using the jackknife method. *Organizational Research Methods*, 2, 293–308.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17, 354–373.
- Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A tale of three perspectives examining

- post hoc statistical techniques for detection and correction of common method variance. *Organizational Research Methods*, 12, 762–800.
- Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30, 359–383.
- Rigdon, E. E. (1997). Not positive definite matrices—Causes and cures. Retrieved from [www2.gsu.edu/~mkteer/npdmatr.html](http://www2.gsu.edu/~mkteer/npdmatr.html)
- Rigdon, E. E. (2013). Partial least squares path modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 81–116). Charlotte, NC: IAP.
- Rigdon, E. E. (2014, May). *Factor indeterminacy and factor-based structural equation modeling*. Paper presented at the second Modern Modeling Methods Conference. Retrieved from [www.modeling.uconn.edu/archive/2014/slides-from-paper-symposia](http://www.modeling.uconn.edu/archive/2014/slides-from-paper-symposia)
- Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods and Research*, 13, 109–119.
- Ringle, C. M., Wende, S., & Becker, J.-M. (2014). SmartPLS 3 [computer software]. Retrieved from [www.smartpls.com](http://www.smartpls.com)
- Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, 65, 1–12.
- Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, 42, 59–66.
- Rogosa, D. R. (1988). *Ballad of the casual modeler*. Retrieved from [www.stanford.edu/class/ed260/ballad](http://www.stanford.edu/class/ed260/ballad)
- Romney, D. M., Jenkins, C. D., & Bynner, J. M. (1992). A structural analysis of health-related quality of life dimensions. *Human Relations*, 45, 165–176.
- Rosenberg, J. F. (1998). Kant and the problem of simultaneous causation. *International Journal of Philosophical Studies*, 6, 167–188.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). Retrieved from [www.jstatsoft.org/v48/i02/paper](http://www.jstatsoft.org/v48/i02/paper)
- Roth, D. L., Wiebe, D. J., Fillingham, R. B., & Shay, K. A. (1989). Life events, fitness, hardiness, and health: A simultaneous analysis of proposed stress-resistance effects. *Journal of Personality and Social Psychology*, 57, 136–142.
- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.
- Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25, 127–141.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322–331.
- Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28, 1420–1423.
- Ryder, A. G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S. J., et al. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, 117, 300–313.
- Saris, W. E., & Alberts, C. (2003). Different explanations for correlated disturbance terms in MTMM studies. *Structural Equation Modeling*, 10, 193–213.
- Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.
- SAS Institute. (2014). SAS/STAT (Version 9.4) [computer software]. Cary, NC: Author.
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling*, 21, 167–180.
- Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *American Statistical Association 1988 Proceedings of the Business and Economic Statistics Section* (pp. 308–313). Alexandria, VA: American Statistical Association.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors on covariance

- structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 399–419). Thousand Oaks, CA: Sage.
- 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 chi-square test statistic. *Psychometrika*, 75, 243–248.
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, 21, 149–160.
- Schreiber, J. B. (2008). Core reporting practices in structural equation modeling. *Research in Social and Administrative Pharmacy*, 4, 83–97.
- Schumaker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). Mahwah, NJ: Erlbaum.
- Schumaker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interaction and nonlinear effects in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Scientific Software International. (2006). LISREL (Version 8.8) [computer software]. Skokie, IL: Author.
- Scientific Software International. (2013). LISREL (Version 9.1) [computer software]. Skokie, IL: Author.
- Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research in Human Development*, 6, 144–164.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.
- Shah, R., & Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: Looking back and forward. *Journal of Operations Management*, 24, 148–169.
- Shapiro, A., & Browne, M. W. (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association*, 82, 1092–1097.
- Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and Psychological Measurement*, 66, 435–447.
- Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, 7, 206–218.
- Silvia, E. S. M., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research*, 23, 297–326.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.
- Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290–312). San Francisco: Jossey-Bass.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Spirites, P. (1995). Directed cyclic graphical representations of feedback models. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 491–498). San Francisco: Morgan Kaufmann.
- Spirites, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.
- Stapleton, L. M. (2013). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (2nd ed., pp. 521–562). Greenwich, CT: IAP.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1292–1306.
- StataCorp. (1985–2015). *Stata statistical software: Release 14* [computer software]. College Station, TX: Author.
- StatPoint Technologies, Inc. (1982–2013). Statgraphics Centurion (Version 16.2.04). [Computer software]. Warrenton, VA: Author.

- StatSoft. (2013). *STATISTICA Advanced* (Version 12) [computer software]. Tulsa, OK: Author.
- Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25, 78–107.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173–180.
- Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association*, 96, 331–338.
- 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.
- Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, 42, 893–898.
- Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.
- Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis* (pp. 136–178). Chicago: University of Chicago Press.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80, 99–103.
- Steinmetz, H. (2011). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, 9, 1–12.
- Stone-Romero, E. F., & Rosopa, P. J. (2011). Experimental tests of mediation models: Prospects, problems, and some solutions. *Organizational Research Methods*, 14, 631–646.
- Systat Software. (2009). *Systat* (Version 13.1) [computer software]. Chicago: Author.
- Textor, J., Hardt, J., & Knüppel, S. (2011). DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, 5, 745.
- Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55, 525–534.
- Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261–283). Washington, DC: American Psychological Association.
- Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford Press.
- Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, 60, 174–195.
- Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Upper Saddle River, NJ: Pearson.
- Tomarken, A. J., & Waller, N. G. (2003). Potential problems with “well-fitting”models. *Journal of Abnormal Psychology*, 112, 578–598.
- Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, 1, 31–65.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1–2.
- Tu, Y.-K. (2009). Commentary: Is structural equation modelling a step forward for epidemiologists? *International Journal of Epidemiology*, 38, 549–551.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization. *Measurement and Evaluation in Counseling and Development*, 44, 159–168.
- Valeri, L., & VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 2, 137–150.
- 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.
- VanderWeele, T. J. (2014). A unification of mediation and interaction: A 4-way decomposition. *Epidemiology*, 25, 749–761.
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York: Oxford University Press.
- van Prooijen, J.-W., & van der Kloot, W. A. (2001). Confirmatory analysis of exploratively obtained factor structures. *Educational and Psychological Measurement*, 61, 777–792.
- Vernon, P. A., & Eysenck, S. B. G. (Eds.). (2007). Structural equation modeling [Special issue]. *Personality and Individual Differences*, 42(5).
- Vieira, A. L. (2011). *Interactive LISREL in practice: Getting started with a SIMPLIS approach*. New York: Springer.
- Voelkle, M. C. (2008). Reconsidering the use of autoregressive latent trajectory (ALT) model. *Multivariate Behavioral Research*, 43, 564–591.
- von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2015). Structural equation modeling with  $\Omega$ nyx. *Structural Equation Modeling*, 22, 148–161.
- Vriens, M., & Melton, E. (2002). Managing missing data. *Marketing Research*, 14, 12–17.
- Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics*, 26, 1–29.
- Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. West Sussex, UK: Wiley.
- 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. 209–246). New York: Guilford Press.
- Westland, C. J. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, 9, 476–487.
- Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 2, 440–451.
- Whitaker, B. G., & McKinney, J. L. (2007). Assessing the measurement invariance of latent job satisfaction ratings across survey administration modes for respondent subgroups: A MIMIC modeling approach. *Behavior Research Methods*, 39, 502–509.
- Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, 75, 1182–1189.
- Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indexes in structural equation modeling. *Psychological Methods*, 8, 16–37.
- Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.
- 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.
- Williams, L. J. (2012). Equivalent models: Concepts, problems, alternatives. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 247–260). New York: Guilford Press.
- Williams, T. H., McIntosh, D. E., Dixon, F., Newton, J. H., & Youman, E. (2010). A confirmatory factor analysis of the Stanford-Binet Intelligence Scales, fifth edition, with a high-achieving sample. *Psychology in the Schools*, 47, 1071–1083.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58–79.
- Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (Vol. 2, pp. 1–54). Amsterdam: North-Holland.
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73, 913–934.

- Wolfle, L. M. (2003). The introduction of path analysis to the social sciences, and some emergent themes: An annotated bibliography. *Structural Equation Modeling*, 10, 1–34.
- Wong, C.-S., & Law, K. S. (1999). Testing reciprocal relations by nonrecursive structural equation models using cross-sectional data. *Organizational Research Methods*, 2, 69–87.
- Worland, J., Weeks, G. G., Janes, C. L., & Strock, B. D. (1984). Intelligence, classroom behavior, and academic achievement in children at high and low risk for psychopathology: A structural equation analysis. *Journal of Abnormal Child Psychology*, 12, 437–454.
- Wothke, W. (1993). Nonpositive definite matrices in structural equation modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20, 557–585.
- Wright, S. (1923). The theory of path coefficients: A reply to Niles' criticism. *Genetics*, 20, 239–255.
- Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.
- Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research and Evaluation*, 12(3). Retrieved from <http://pareonline.net/pdf/v12n3.pdf>
- Wu, C. H. (2008). The role of perceived discrepancy in satisfaction evaluation. *Social Indicators Research*, 88, 423–436.
- Yang, C., Nay, S., & Hoyle, R. H. (2010). Three approaches to using lengthy ordinal scales in structural equation models: Parceling, latent scoring, and shortening scales. *Applied Psychological Measurement*, 34, 122–142.
- Yang-Wallentin, F. (2001). Comparisons of the ML and TSLS estimators for the Kenny–Judd model. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 425–442). Lincolnwood, IL: Scientific Software International.
- Yang-Wallentin, F., & Jöreskog, K. G. (2001). Robust standard errors and chi-squares for interaction models. In G. A. Marcoulides & R. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 159–171). Mahwah, NJ: Erlbaum.
- Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, 40, 115–148.
- Yuan, K.-H., Hayashi, K., & Bentler, P. (2007). Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses. *Journal of Multivariate Analysis*, 9, 1262–1282.
- Ziliak, S., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.

# Author Index

- Abbott, J. L., 57, 489  
Achen, C. H., 34, 38, 489  
Ackerman, P. L., 375, 498  
Acock, A. C., 110, 489  
Agresti, A., 47, 79, 489  
Aguinis, H., 57, 430, 489  
Aiken, L. S., 31, 34, 47, 424, 428, 430, 493  
Akaike, H., 286, 489  
Alberts, C., 323, 505  
Alegria, M., 395, 502  
Allen, C. A., 453, 492  
Allison, P. D., 82, 489  
Alpert, A., 23, 447, 494  
Altman, D. G., 44, 47, 505  
Amemiya, Y., 443, 508  
Anderson, J. C., 267, 338, 489, 495  
Angert, C., 57, 489  
Antonakis, J., 132, 142, 182, 323, 361, 489, 501  
Armstrong, J. S., 57, 489  
Asparouhov, T., 402, 437, 489, 502  
Austin, J. T., 16, 22, 129, 263, 296, 453, 500  
  
Bailey, M., 323, 500  
Bakker, A. B., 247, 500  
Balla, J. R., 267, 500  
Bandalos, D. L., 62, 290, 489  
Barbano, H. E., 328, 493, 500  
Baron, R. M., 430, 432, 435, 490  
Barrett, P., 16, 268, 490  
Bartholomew, D. J., 18, 490  
Bauer, D. J., 375, 401, 445, 447, 490, 493, 501  
Bauldry, S., 197, 304, 354, 355, 362, 491  
Baumgartner, H., 401, 507  
Baylor, C., 95, 490  
Beauducel, A., 268, 490  
Beaujean, A. A., 108, 490  
Becker, J.-M., 361, 505  
  
Behson, S. J., 34, 501  
Bendahan, S., 132, 142, 182, 323, 489  
Bentler, P. M., 23, 77, 103, 104, 267, 269, 272, 276, 277, 282, 299, 314, 330, 335, 340, 365, 490, 497, 499, 500, 505, 506, 509  
Benyamin, Y., 395, 490  
Beres, M. A., 274, 497  
Bergsma, W., 10, 490  
Bernstein, I. H., 196, 257, 490, 503  
Berry, W. D., 150, 154, 490  
Bishop, J., 391, 490  
Black, A. C., 467, 468, 501  
Blalock, H. M., 23, 490  
Blest, D. C., 74, 490  
Block, J., 125, 490  
Blunch, N., 103, 490  
Boadu, K., 10, 268, 270, 271, 496  
Boker, S. M., 108, 490, 503  
Bolger, N., 157, 202, 203, 204, 499  
Bollen, K. A., 19, 20, 21, 24, 27, 126, 149, 153, 188, 189, 197, 207, 216, 217, 234, 235, 237, 239, 259, 274, 276, 304, 315, 340, 352, 354, 355, 359, 360, 362, 374, 391, 392, 393, 442, 454, 490, 491, 492, 495  
Bonett, D. G., 277, 490  
Bonnet, G., 111, 495  
Boomsma, A., 120–121, 453, 491  
Boulian, S., 10, 268, 270, 271, 496  
Bovaird, J. A., 430, 444, 500  
Box, G. E. P., 79, 263, 491  
Bradbury, R. B., 38, 508  
Braddy, P. W., 271, 400, 401, 501  
Brandmaier, A. M., 106, 236, 508  
Breitling, L. P., 112, 491  
Breivik, E., 276, 360, 491, 497, 503  
Brett, J. M., 434, 498  
Brick, T., 108, 490  
Briggs, N. E., 374–375, 392, 393, 504

- Brito, C., 136, 181, 491  
 Brody, J. A., 328, 493  
 Brosseau-Liard, P. É., 258, 504  
 Brown, S. P., 453, 493  
 Brown, T. A., 195, 314, 333, 491  
 Browne, M. W., 254, 256, 272, 274, 279, 290, 291, 299,  
   354, 358, 375, 376, 491, 492, 500, 506  
 Bryant, F. B., 282, 492  
 Budtz-Jørgensen, E., 78, 79, 492  
 Bullock, J. G., 182, 184, 492  
 Burt, R. S., 339, 492  
 Bynner, J. M., 287, 288, 289, 505  
 Byrne, B. M., 103, 104, 106, 113, 248, 401, 492  
  
 Cameron, L. C., 207, 492  
 Campbell, D. T., 51, 93, 124, 321, 492, 506  
 Card, N. A., 200, 404, 406, 500  
 Carlson, J. F., 88, 492  
 Castellanos, J., 312, 499  
 Chan, F., 453, 492  
 Chang, H.-T., 222, 223, 349, 492  
 Chen, B., 251, 492  
 Chen, F., 235, 237, 274, 276, 491, 492  
 Chen, F. F., 319, 321, 401, 492  
 Chernyshenko, O. S., 399, 400, 506  
 Cheung, G. W., 264, 271, 387, 398, 400, 401, 492, 493,  
   504  
 Chi, N. W., 222, 223, 349, 492  
 Chin W. W., 361, 453, 493  
 Choi, J., 411, 493  
 Choi, Y. Y., 111, 493  
 Chun, J. S., 111, 493  
 Ciesla, J. A., 196, 493  
 Clark, S. L., 15, 508  
 Coffman, D. L., 82, 87, 88, 96, 290, 496, 504  
 Cohen, B., 328, 493, 500  
 Cohen, J., 31, 34, 47, 424, 427, 430, 493  
 Cohen, P., 31, 34, 47, 424, 427, 430, 493  
 Cole, D. A., 127, 135, 140, 196, 225, 323, 391, 490,  
   493, 494, 501  
 Cook, T. D., 51, 124, 506  
 Cornoni-Huntley, J., 328, 493  
 Cox, C. S., 328, 500  
 Cox, D. R., 79, 491  
 Crawford, J. R., 282, 493  
 Cribbie, R. A., 51, 499  
 Croon, M. A., 10, 490  
 Crowne, D. P., 323, 493  
 Cudeck, R., 274, 491  
 Cudek, R., 253, 493  
 Cumming, G., 60, 493  
 Cummings, G., 10, 19, 165, 167, 169, 170, 268, 270,  
   271, 496  
 Cummings, G. C., 274, 497  
 Curran, P. J., 235, 237, 274, 276, 374, 375, 391, 392,  
   393, 445, 447, 448, 491, 492, 493  
 Curran, T., 434, 493  
  
 Dawson, J. F., 430, 493  
 Depaoli, S., 23, 498  
 Deshon, R. P., 399, 493  
  
 Diamantopoulos, A., 105, 354, 359, 493  
 Dillman Carpentier, F. R., 403, 494  
 DiStefano, C., 193, 257, 258, 259, 260, 326, 327, 328,  
   452, 453, 494, 495  
 Dixon, F., 319, 508  
 Donahue, B., 51, 499  
 Donovan, N. J., 95, 490  
 Dosman, D., 19, 165, 167, 169, 170, 496  
 Doyle, P. J., 95, 490  
 Drasgow, F., 399, 400, 506  
 Du Toit, S. H. C., 258, 375, 376, 492, 502  
 Dumka, L. E., 403, 494  
 Duncan, O. D., 23, 155, 494  
 Duncan, S. C., 23, 447, 494  
 Duncan, T. E., 23, 447, 494  
  
 Edwards, J. R., 359, 361, 362, 427, 430, 432, 434, 494,  
   501  
 Edwards, M. C., 326, 331, 332, 333, 421, 494, 508  
 Efron, B., 60, 494  
 Eid, M., 323, 494  
 Ein-Dor, T., 395, 490  
 Ellis, P. D., 53, 494  
 Elwert, F., 166, 167, 170, 185, 494  
 Enders, C. K., 82, 83, 87, 96, 236, 494, 504  
 Epskamp, S., 109, 494  
 Erceg-Hurn, D. M., 51, 54, 494  
 Ercikan, K., 420, 503  
 Eysenck, S. B. G., 298, 508  
  
 Fabrigar, L. R., 195, 207, 209, 296, 494, 500  
 Fan, W., 411, 493  
 Fan, X., 17, 277, 494  
 Feldman, J. J., 328, 493, 500  
 Figueiredo, A. J., 82, 501  
 Fillingham, R. B., 82, 159, 239, 505  
 Finkel, S. E., 137, 494  
 Finney, S. J., 258, 259, 260, 326, 327, 328, 494, 495  
 Fiske, D. W., 93, 321, 492  
 Fitzpatrick, D. C., 198, 503  
 Flora, D. B., 391, 495  
 Forero, C. G., 330, 495  
 Foss, T., 238, 257, 276, 503  
 Fouladi, R. T., 9, 265, 507  
 Fox, J., 107, 495  
 Freckleton, R. P., 38, 508  
 Freeman, M. J., 290, 496  
 Frees, E. W., 139, 495  
 French, B. F., 292, 496  
 Friendly, M., 79, 290, 495  
  
 Gagné, P., 62, 489  
 Gallardo-Pujol, D., 330, 495  
 Ganambs, T., 290, 495  
 Gardner, H., 300–301, 495  
 Garnier-Villarral, M., 290, 504  
 Geiser, C., 106, 323, 391, 490, 494, 495  
 Geisinger, K. F., 88, 492  
 George, R., 391, 495  
 Gerbing, D. W., 137, 267, 338, 489, 495, 497  
 Gignac, G. E., 321, 495

- Gillaspy, J. A., Jr., 452, 453, 498  
 Ginzburg, K., 395, 490  
 Glaser, D. N., 339, 340, 496  
 Glymour, C., 170, 185, 315, 495, 506  
 Goldman, B. A., 89, 495  
 Goldstein, H., 111, 495  
 Goldstein, S. M., 16, 22, 263, 453, 506  
 Gollwitzer, N., 323, 494  
 Gonzales, N. A., 403, 494  
 Gonzalez, R., 337, 495  
 Goodwin, L. D., 42, 495  
 Grace, J. B., 352, 355, 359, 453, 454, 495  
 Graham, J. M., 82, 87, 88, 96, 194, 302, 496  
 Grandjean, P., 78, 79, 492  
 Grayson, D., 323, 501  
 Green, D. P., 182, 184, 492  
 Greenland, S., 165, 170, 505  
 Gregorich, S. E., 397, 398, 496  
 Griffin, D., 337, 495  
 Grygoryev, K., 19, 165, 167, 169, 170, 496  
 Gursoy, D., 453, 503  
 Guthrie, A. C., 194, 302, 496
- Ha, S. E., 182, 184, 492  
 Hagaars, J. A., 10, 18, 490, 496  
 Haller, H., 55, 496  
 Hallquist, M., 106, 496  
 Hamann, J. D., 108, 497  
 Hancock, G. R., 62, 76, 239, 276, 290, 292, 313, 411, 453, 454, 493, 496, 502, 503  
 Hardt, J., 84, 112, 179, 496, 507  
 Harik, P., 137, 364, 498  
 Harrington, D., 333, 496  
 Harrington, K. M., 15, 508  
 Harrison, P., 207, 492  
 Hatcher, L., 109, 503  
 Hau, K.-T., 267, 268, 303, 442, 443, 444, 450, 500, 501  
 Hayashi, K., 276, 509  
 Hayduk, L. A., 10, 19, 165, 167, 169, 170, 217, 225, 265, 268, 270, 271, 272, 274, 275, 339, 340, 496, 497  
 Hayes, A. F., 245, 429, 430, 432, 434, 435, 497, 504  
 Heine, S. J., 397, 505  
 Henningsen, A., 108, 497  
 Herke, M., 84, 496  
 Hershberger, S. L., 293, 296, 317, 348, 354, 497, 499  
 Hess, B., 452, 453, 494  
 Hicks, R., 437, 497  
 Hill, A. P., 434, 493  
 Hipp, J. R., 157, 160, 503  
 Hirschfeld, G., 420, 497  
 Ho, M.-H. R., 453, 501  
 Hoekstra, R., 54, 497  
 Holland, P. W., 126, 497  
 Holsta, K. K., 108, 497  
 Hops, H., 447, 494  
 Horn, J. L., 396, 497  
 Hotchkiss, L., 137, 364, 498  
 Houghton, J. D., 220, 341, 342, 344, 345, 346, 348, 497  
 Houts, C. R., 326, 331, 333, 494  
 Howell, R. D., 238, 257, 360, 497, 503  
 Hoyle, R. C., 452, 453, 497
- Hoyle, R. H., 120–121, 142, 207, 332, 453, 491, 497, 509  
 Hu, L., 267, 277, 497  
 Huberty, C. J., 51, 499  
 Huck, S. W., 42, 497  
 Hula, W., 95, 490  
 Hunter, J. E., 137, 497  
 Hurlbert, S. H., 57, 497  
 Hwang, Y. R., 207, 492
- Imai, K., 437, 498  
 Ing, M., 285, 500  
 Ioannidis, J. P. A., 56, 498  
 Isherwood, J. C., 452, 453, 497  
 Ittenbach, R. F., 207, 492
- Jackson, D. L., 16, 452, 453, 498  
 Jacquart, P., 132, 142, 182, 323, 489  
 James, G. S., 423, 498  
 James, L. R., 434, 498  
 Janes, C. L., 355, 356, 509  
 Jarvis, C. B., 359, 498  
 Jenkins, C. D., 287, 288, 289, 505  
 Jinkerson, D. L., 220, 341, 342, 344, 345, 346, 348, 497  
 Johnson, A., 54, 497  
 Johnson, E. C., 271, 400, 401, 501  
 Jonson, J. L., 88, 492  
 Jöreskog, K. G., 11, 299, 440, 442, 498, 509  
 Jorgensen, T., 109, 504  
 Judd, C. M., 430, 438, 439, 440, 441, 442, 498, 499  
 Jung, S., 14, 498
- Kane, M. T., 93, 498  
 Kanfer, R., 375, 498  
 Kaplan, D., 22, 23, 137, 238, 239, 256, 269, 284, 364, 377, 391, 498  
 Karami, H., 398, 421, 498  
 Kashy, D. A., 157, 202, 203, 204, 323, 499  
 Kaufman, A. S., 206, 305, 498  
 Kaufman, N. L., 206, 305, 498  
 Kaur, G., 220, 225, 501  
 Keele, L., 437, 498  
 Keiding, N., 78, 79, 492  
 Keith, T. Z., 207, 304, 312, 498  
 Kelley, K., 247, 504  
 Kendall, D., 95, 490  
 Kenny, D. A., 19, 27, 147, 148, 149, 152, 157, 160, 195, 201, 202, 203, 204, 233, 250, 277, 306, 315, 317, 323, 374, 430, 432, 435, 438, 439, 440, 441, 442, 490, 498, 499  
 Keselman, H. J., 51, 499  
 Khine, M. S., 453, 454, 499  
 Kiers, H. A. L., 54, 497  
 Kim, J., 453, 503  
 Kim, K. H., 292, 499  
 Kirby, J. B., 235, 237, 491, 492  
 Kisynski, J., 112, 171, 504  
 Klein, A. G., 443, 444, 499  
 Kleinman, J. C., 328, 493, 500  
 Kline, R. B., 17, 42, 54, 56, 62, 63, 142, 207, 209, 312, 410, 499

- Knight, C. R., 124, 499  
 Knoll, B., 112, 171, 504  
 Knüppel, S., 112, 179, 499, 507  
 Kohlhausen, D., 57, 489  
 Krauss, S., 55, 496  
 Krull, J. L., 247, 500  
 Kubota, C., 453, 492  
 Kühnel, S., 8, 499
- Lalive, R., 132, 142, 182, 323, 489  
 Lambdin, C., 54, 62, 63, 499  
 Lambert, L. S., 434, 494  
 Lance, C. E., 399, 403, 432, 499, 507  
 Lash, T. L., 165, 170, 505  
 Lautenschlager, G. J., 396, 501  
 Law, K. S., 137, 138, 509  
 Lee, E.-J., 453, 492  
 Lee, G. K., 453, 492  
 Lee, K. O., 111, 493  
 Lee, S., 293, 330, 499  
 Leech, N. L., 42, 495  
 Lei, P.-W., 259, 260, 327, 499  
 Leite, W., 290, 489  
 Leonhart, R., 84, 496  
 Levers, M.-J. D., 274, 497  
 Lewis, C., 276, 507  
 Li, F., 23, 494  
 Li, Z., 396, 399, 422, 509  
 Lindenberger, U., 91, 92, 93, 94, 500  
 Lischetzke, T., 323, 494  
 Little, R. J. A., 84, 86, 499  
 Little, T. D., 79, 83, 84, 88, 91, 92, 93, 94, 99, 100, 101, 124, 134, 138, 141, 200, 262, 277, 332, 392, 396, 399, 404, 406, 430, 443, 444, 453, 499, 500, 503  
 Littvay, L., 217, 225, 496  
 Liu, M., 62, 496  
 Liu, W., 390, 501  
 Lix, L. M., 51, 499  
 Lockwood, C. M., 247, 500  
 Loehlin, J. C., 16, 39, 500  
 Lomax, R. G., 454, 506  
 Lombardi, C. M., 57, 497  
 Lubansky, J. B., 96, 500  
 Lynam, D. R., 125, 500
- Maasen, G. H., 247, 500  
 MacCallum, R. C., 16, 22, 129, 263, 284, 285, 290, 291, 296, 354, 358, 374–375, 392, 393, 453, 500, 504, 506  
 MacKenzie, S. B., 359, 498  
 MacKinnon, D. P., 19, 182, 247, 435, 437, 500  
 Mackworth, A., 112, 171, 504  
 Madans, J. H., 328, 500  
 Maddox, T., 88, 500  
 Maes, H. H., 108, 490, 503  
 Mahalingam, R., 395, 502  
 Mair, P., 104, 500  
 Malone, P. S., 96, 500  
 Marcoulides, G. A., 23, 128, 285, 293, 296, 317, 348, 497, 500, 504, 506  
 Mardia, K. V., 74, 500  
 Markland, D., 268, 270, 272, 500
- Marks, M., 57, 507  
 Marlowe, D., 323, 493  
 Marquart-Pyatt, S. T., 157, 160, 503  
 Marsh, H. W., 220, 225, 267, 268, 276, 303, 323, 442, 443, 444, 450, 500, 501  
 Masyn, K. E., 390, 501  
 Matsueda, R. L., 24, 501  
 Matthes, J., 430, 497  
 Mauricio, A. M., 403, 494  
 Mauro, R., 35, 501  
 Maxwell, S. E., 135, 140, 493, 501  
 Maydeu-Olivares, A., 330, 495  
 McArdle, J. J., 106, 396, 497, 501  
 McClelland, G. H., 430, 498  
 McCloskey, D. N., 54, 62, 63, 509  
 McCoach, D. B., 467, 468, 501  
 McCutcheon, A. L., 18, 496  
 McDonald, R. A., 34, 501  
 McDonald, R. P., 106, 276, 400, 453, 501  
 McGrew, K. S., 207, 492  
 McIntosh, C. N., 361, 501  
 McIntosh, D. E., 319, 508  
 McKinney, J. L., 396, 508  
 McKnight, K. M., 82, 501  
 McKnight, P. E., 82, 501  
 Meade, A. W., 271, 396, 400, 401, 501  
 Melton, E., 86, 87, 508  
 Meredith, W., 377, 502  
 Merkle, E. C., 289, 504  
 Meshkat, P., 174, 503  
 Messick, S., 93, 502  
 Meza, C. M., 403, 494  
 Miao, M. C., 222, 223, 349, 492  
 Micceri, T., 51, 502  
 Milan, S., 27, 148, 149, 160, 315, 499  
 Miles, J., 266, 502  
 Miller, M. W., 15, 508  
 Miller, P., 109, 290, 504  
 Millsap, R. E., 11, 198, 263, 264, 339, 400, 403, 405, 408, 411, 413, 415, 420, 421, 494, 502  
 Míndrailă, D., 193, 494  
 Mitchell, D. F., 89, 495  
 Moffitt, T., 125, 500  
 Molina, K. M., 395, 502  
 Monecke, A., 361, 502  
 Mooijaart, A., 271, 502  
 Moosbrugger, A., 443, 444, 499  
 Morin, A. J. S., 220, 225, 501  
 Mueller, R. O., 313, 453, 454, 496, 502  
 Mulaik, S. A., 11, 123, 124, 151, 189, 190, 193, 236, 266, 267, 268, 269, 274, 339, 502  
 Muthén, B. O., 18, 23, 105, 258, 304, 324, 328, 357, 402, 406, 437, 441, 443, 489, 499, 502  
 Muthén, L. K., 105, 304, 328, 357, 402, 406, 437, 441, 502  
 Myers, T. A., 87, 502
- Nagengast, B., 444, 450, 501  
 Narayanan, A., 113, 502  
 Nay, S., 332, 509  
 Neale, M. C., 108, 490, 503  
 Nesselroade, J. R., 91, 92, 93, 94, 500

- Nevitt, J., 76, 239, 276, 503  
Newsom, J., 138, 412, 503  
Newton, J. H., 319, 508  
Niewander, W. A., 42, 505  
Niemiec, C. P., 434, 493  
Nimmo, M., 19, 165, 167, 169, 170, 496  
Nimon, K., 421, 503  
Nunkoo, R., 453, 503  
Nunnally, J. C., 196, 503  
Nussbeck, F. W., 323, 494
- O'Brien, R. M., 202, 503  
O'Connell, A. A., 467, 468, 501  
Okech, D., 453, 503  
Olejnik, S., 51, 499  
Olivera-Aguilar, M., 400, 403, 408, 502  
Oliveri, M. E., 420, 503  
Olson, B. D., 420, 503  
Olsson, U. H., 238, 257, 276, 491, 503  
O'Rourke, R., 109, 503  
Osborne, J. W., 78, 79, 198, 503
- Panter, A. T., 120–121, 453, 491  
Park, I., 390, 391, 392, 393, 503  
Park, J. H., 57, 489  
Parker, P. D., 220, 225, 501  
Paxton, P., 157, 160, 237, 274, 276, 492, 503  
Paxton, P. M., 235, 491  
Pazderka-Robinson, H., 10, 268, 270, 271, 274, 496, 497  
Pearl, J., 9, 19, 20, 21, 24, 27, 122, 126, 128, 131, 136, 164, 165, 169, 170, 174, 177, 179, 180, 181, 185, 232, 251, 293, 296, 491, 492, 503  
Pedhazur, E. J., 8, 504  
Peters, C. L. O., 82, 87, 504  
Petersen, M. L., 183, 184, 436, 504  
Peterson, R. A., 453, 493  
Petras, H., 390, 501  
Ping, R. A., 442, 504  
Pinter, J., 236, 504  
Pirlott, A. G., 182, 435, 437, 500  
Podsakoff, P. M., 359, 498  
Poole, D., 112, 171, 504  
Poon, W. Y., 330, 499  
Pornprasertmanit, S., 109, 290, 504  
Porter, K., 112, 171, 504  
Preacher, K. J., 127, 141, 225, 245, 247, 289, 290, 374–375, 392, 393, 429, 432, 434, 435, 448, 493, 504, 506  
Purc-Stephenson, R., 452, 453, 498
- Quick, C., 109, 290, 504
- Rabe-Hesketh, S., 18, 506  
Radloff, L. S., 328, 414, 504  
Raftery, A. E., 287, 504  
Ramkissoon, H., 453, 503  
Ramsey, J., 315, 495  
Raykov, T., 90, 128, 313, 317, 365, 490, 504  
Reio, T., Jr., 421, 503  
Rensvold, R. B., 264, 271, 387, 398, 400, 401, 492, 493, 504
- Rhemtulla, M., 258, 504  
Richardson, H. A., 323, 504  
Richter, A. W., 430, 493  
Riefler, P., 354, 493  
Rigdon, E. E., 69, 154, 156, 160, 212, 361, 362, 505  
Rindskopf, D., 158, 505  
Ringle, C. M., 361, 505  
Rocher, T., 111, 495  
Rodgers, J. L., 12, 17, 505  
Rogosa, D. R., 9, 505  
Romney, D. M., 287, 288, 289, 505  
Rosenberg, J. F., 123, 505  
Rosopa, P. J., 182, 507  
Rosseel, Y., 107, 349, 505  
Roth, D. L., 82, 159, 239, 505  
Roth, K. P., 354, 493  
Rothman, K. J., 165, 170, 505  
Royston, P., 44, 47, 505  
Rubin, D. B., 18, 19, 123, 505  
Rucker, D. D., 429, 432, 434, 435, 504  
Ryder, A. G., 397, 505
- Sairs, W. E., 290, 323, 505  
Satorra, A., 271, 272, 282, 290, 299, 492, 502, 505, 506  
Sauerbrei, W., 44, 47, 505  
Savalei, V., 238, 258, 504, 506  
Sayer, A. G., 375, 377, 508  
Scheines, R., 315, 495, 506  
Schmelkin, L. P., 8, 504  
Schönemann, P. H., 212, 507  
Schosemann, A., 109, 290, 504  
Schreiber, J. B., 453, 506  
Schumaker, R. E., 23, 454, 506  
Schutz, R. W., 390, 391, 392, 393, 503  
Seifert, C. F., 34, 501  
Selig, J. P., 141, 506  
Shadish, W. R., 51, 124, 506  
Shah, R., 16, 22, 263, 453, 506  
Shapiro, A., 256, 506  
Shay, K. A., 82, 159, 239, 505  
Shevlin, M., 266, 502  
Shieh, G., 36, 47, 506  
Shipley, B., 174, 177, 506  
Sidani, S., 82, 501  
Siguaw, J. A., 105, 493  
Silvia, E. S. M., 284, 285, 506  
Simmering, M. J., 323, 504  
Sinisi, S. E., 183, 184, 436, 504  
Sivo, S. A., 277, 494  
Skrondal, A., 18, 506  
Slegers, D. W., 200, 404, 406, 500  
Snyder, J., 312, 499  
Sobel, M. E., 245, 506  
Solomon, Z., 395, 490  
Song, J.-I., 111, 493  
Song, W. K., 111, 493  
Sousa, K. H., 319, 321, 492  
Spearman, C., 23, 189, 315, 506  
Spiegel, M., 108, 490  
Spirtes, P., 186, 315, 495, 506  
Spisic, D., 258, 502  
Stang, A., 112, 179, 499

- Stapleton, L. M., 445, 450, 506  
 Stark, S., 399, 400, 506  
 Steenkamp, J.-B. E. M., 401, 507  
 Steiger, J. H., 9, 98, 196, 212, 254, 265, 269, 312, 336,  
   341, 363, 493, 507  
 Steinmetz, H., 402, 507  
 Stephens, P. A., 38, 508  
 Stine, R. A., 239, 491  
 Stone-Romero, E. F., 182, 507  
 Stouthamer-Loeber, M., 125, 500  
 Stratkotter, R., 19, 165, 167, 169, 170, 496  
 Streiner, D. L., 91, 507  
 Strock, B. D., 355, 356, 509  
 Strycker, L. A., 23, 494  
 Sturman, M. C., 323, 504  
 Sudea, S., 112, 171, 504  
 Sugawara, H. M., 290, 291, 500
- Taylor, A. B., 298, 508  
 Taylor, L. R., 207, 492  
 Teng, G., 257, 490  
 Textor, J., 112, 179, 507  
 Thompson, B., 38, 58, 90, 194, 302, 452, 453, 454,  
   496, 507  
 Thompson, J. S., 277, 508  
 Thorndike, R. M., 90, 507  
 Thorndike-Christ, T. M., 90, 507  
 Ting, K., 315, 491  
 Tingley, D., 437, 497  
 Tisak, J., 377, 502  
 Tomarken, A. J., 264, 298, 467, 468, 507  
 Trafimow, D., 57, 507  
 Troye, S. V., 238, 257, 503  
 Tsang, S., 106, 236, 508  
 Tu, Y.-K., 467, 468, 507  
 Tucker, L. R., 276, 507
- Uchino, B. N., 296, 500
- Vacha-Haase, T., 90, 507  
 Valeri, L., 435, 436, 437, 450, 451, 507  
 van der Kloot, W. A., 198, 508  
 van der Laan, M. J., 183, 184, 436, 504  
 van Prooijen, J.-W., 198, 508  
 Vandenberg, R. J., 399, 507  
 VanderWeele, T. J., 435, 436, 437, 450, 451, 507, 508  
 Vernon, P. A., 298, 508  
 Vieira, A. L., 105, 508  
 von Brachel, R., 420, 497  
 von Oertzen, T., 106, 236, 508  
 Vriens, M., 86, 87, 508
- Wall, M. M., 443, 508  
 Waller, N. G., 264, 298, 467, 468, 507  
 Wang, J., 106, 508  
 Wang, X., 106, 508  
 Weeks, G. G., 355, 356, 509  
 Wegener, D. T., 296, 500
- Weihe, P., 78, 79, 492  
 Wen, Z., 268, 442, 443, 444, 450, 501  
 Wende, S., 361, 505  
 Werner, S., 57, 489  
 West, S. G., 31, 34, 47, 298, 319, 321, 424, 427, 430,  
   492, 493, 508  
 Westland, C. J., 16, 508  
 Wherry, R. J., 33, 508  
 Whitaker, B. G., 396, 508  
 Whittingham, M. J., 38, 508  
 Wichman, A. L., 374–375, 392, 393, 504  
 Widaman, K. F., 277, 430, 444, 500, 508  
 Wiebe, D. J., 82, 159, 239, 505  
 Wilcox, J. B., 360, 497  
 Wilcox, R. R., 73, 508  
 Wilde, M., 108, 490  
 Wiley, J., 106, 496  
 Willett, J. B., 375, 377, 508  
 Williams, L. J., 296, 508  
 Williams, T. H., 319, 508  
 Winklhofer, H. M., 359, 493  
 Winship, C., 124, 499  
 Wirth, R. J., 326, 331, 332, 333, 421, 494, 508  
 Wittman, W., 268, 490  
 Wold, H., 360, 508  
 Wolf, E. J., 15, 508  
 Wolfle, L. M., 23, 24, 509  
 Wong, C.-S., 137, 138, 509  
 Worland, J., 355, 356, 509  
 Woithke, W., 69, 71, 313, 509  
 Wright, S., 23, 122, 250, 509  
 Wu, A. D., 396, 399, 422, 509  
 Wu, C. H., 448, 449, 509  
 Wu, E., 104, 500  
 Wu, Q., 259, 260, 327, 499  
 Wu, W., 298, 508
- Xi, N., 326, 331, 333, 494  
 Xie, G., 108, 503
- Yamamoto, T., 437, 498  
 Yang, C., 332, 509  
 Yang, F., 442, 498  
 Yang, J., 397, 505  
 Yang-Wallentin, F., 440, 443, 509  
 Yao, S., 397, 505  
 Yi, J., 397, 505  
 Yorkston, K., 95, 490  
 Youman, E., 319, 508  
 Yuan, K.-H., 268, 276, 277, 509  
 Yun-Tein, J., 411, 413, 415, 420, 502
- Zhang, Z., 448, 504  
 Zhu, M., 193, 494  
 Zhu, X., 397, 505  
 Ziliak, S., 54, 62, 63, 509  
 Zumbo, B. D., 396, 399, 420, 422, 503, 509  
 Zyphur, M. J., 448, 504

# Subject Index

*f, t, n* following a page number indicates figure, table, or footnote.

- Absolute fit indexes, 266, 273–278
- Accept–support test, 265
- Adaptive quadrature, 258
- Adjacent vertices, 165
- Akaike Information Criterion (AIC), 286–287, 289
- Alignment method, 402
- All-Y notation, in LISREL, 227–228
- Alternate-forms reliability, 92
- Alternative models
  - discussion of, 11
  - model specification and, 456–457
- Amos program
  - Amos Graphics, 102–103
  - Amos Program Editor, 103
  - analysis of CFA models, 330
  - characteristics of, 102t
  - identification of multivariate outliers, 73
  - incomplete data procedure, 87
  - latent growth models
    - analysis of a basic change model, 380–384
    - analysis of a prediction model, 385–387
  - modification indexes, 283
  - overview and description, 102–103
  - using with SEM, 9
- Analysis of covariance structures, 9
- Ancestors, 166
- ANOVA (analysis of variance), 14, 17, 380
- Approximate fit, 60
- Approximate fit indexes
  - best practices in reporting on, 464–465
  - overview and description, 266–268
  - power estimation and, 292
  - recursive path model of illness example, 279
  - testing measurement invariance and, 400–401
- Arbitrary distribution estimators, 256–257
- Arbitrary distribution function, 256–257
- Arbitrary generalized least squares (AGLS)
  - estimation, 330
- Archival samples, 459
- Arcs, 165
- Arcsine square root transformations, 78
- Arrow (→), 165
- Asymptotically distribution free (ADF) estimator, 272, 279
- Asymptotic covariance matrix, 325–326
- Attenuation bias, 33–34
- Autocorrelated errors, 138
- Autoregressive integrative moving average (ARIMA) models, 392
- Autoregressive latent trajectory (ALT) models, 391
- Autoregressive paths, 139
- Autoregressive structures, 391–392
- Auxiliary variables, 84
- Available case methods, 85–86
- Average variance extracted (AVE), 313
- Back-door criterion, 177–179
- Back-door paths, 167–168
- Backward elimination, 38
- Badness-of-fit statistics, 266, 273–278
- Ballad of the Casual Modeler* (Rogosa), 9
- Baron–Kenny method, 435, 436
- Basic change latent growth model, 376–384
- Basis set
  - with causal directed graphs, 176
  - discussion of, 173–174, 175t
- Batch mode processing, 98, 99
- Bayesian networks, 164
- Bayesian statistics, 23
- Bayes Information Criterion (BIC), 287–289
- Belief and Decision Network Tool, 112, 171

- Bentler Comparative Fit Index (CFI), 269, 276–277, 292, 400, 401  
 Bentler–Raykov corrected  $R^2$ , 365–366  
 Bentler–Weeks representational system, 104, 122  
 Berkson's paradox, 169  
 Best indicator, 217  
 Beta weights, 30–32, 429–430  
 Bias  
     attenuation bias, 33–34  
     collider bias, 170  
     confirmation bias, 22, 55, 296, 466  
     corrections for multiple regressions, 32–33  
     endogenous selection bias, 170  
     overcontrol bias, 170  
 Bifactor models, 319–321  
 Biserial correlation, 42  
 Bivariate regression, 25–29  
 Blocked-error- $R^2$ , 365–366  
 Block recursive models, 152, 153  
 Bollen–Stine bootstrap, 239  
 Bollen's two-stage least squares (2SLS) method, 442–443  
 Bootstrapping, 60–62, 239  
 Bow-free patterns, 136  
 Bow patterns, 136  
 Box-and-whisker plots, 74–76  
 Box–Cox transformations, 79  
 Box plots, 74–76
- CALIS (Covariance Analysis of Linear Structural Equations), 102*t*, 109  
 Canonical variate analysis (canonical correlation), 17  
 Categorical latent variables, 18  
 Categorical outcomes, options for analyzing, 237–238  
 Categorization, of predictors or outcomes, 43–44  
 Causal directed graphs  
     acyclic; *see* Directed acyclic graphs  
     description, 174–176  
     testable implications, 176–177  
 Causal effects  
     assumptions regarding in regression analysis, 34  
     conditional process modeling, 432–434  
     model specification and, 456  
     in path analysis, 232–233  
     representation in diagrams, 121  
 Causal hypotheses, representation in graph theory, 165  
 Causal indicators, 197, 352–355, 356–360  
 Causal inference  
     discussion of, 122–126  
     with latent variables, 212  
     myths about the role of SEM in, 20  
     SCM and, 164–165  
 Causal inference frameworks, 18–20  
 Causality. *See also* Causal effects; Causal inference  
     causal assumptions in path models, 131–134  
     deterministic, 11  
     probabilistic, 11–12  
     reciprocal causation, estimating in nonrecursive models, 137–138  
     simultaneous causation, 123*n*
- Causally heterogeneous effects, 133  
 Causally homogenous assumption, 123–124  
 Causal mediation analysis, 181–184, 185, 435–437, 450  
 Censored regression, 43  
 Censored variables, 43  
 Center for Epidemiologic Studies Depression (CES-D) scale, 328–330, 414  
 Centering  
     in bivariate regression, 27  
     of predictors in moderated multiple regression, 425, 427–428  
     residual centering, 430, 443–444  
 Central F distributions, 58, 59  
 Central t distribution, 51–52  
 Centroid, 73  
 CFA. *See* Confirmatory factor analysis  
 CFI. *See* Bentler Comparative Fit Index  
 Chains, 166–167  
 Children, 166  
 Chi-square difference statistic  
     description, 281–283  
     modification indexes and, 283–284  
     recursive path model of illness example, 285–286  
     testing indicators in CFA models, 314  
     testing measurement invariance and, 400  
     Wald W statistic and, 284  
 Chi-square distributions, noncentral, 60  
 Chi-square statistics  
     model chi-square, 269, 270–273; *see also* Model chi-square  
     model test statistics, 265–266; *see also* Model test statistics  
     for other estimators, 272–273  
     recursive path model of illness example, 278–279  
 Classes, 18  
 Classical school of statistics, 262  
 Classical suppression, 37  
 Close fit, 60  
 Close-fit hypothesis, 274, 290  
 Close-yet-failing models, 274  
 Coding  
     best practices, 458–459; *see also* Input data effects coding method, 200  
     reverse, 197  
 Coefficient alpha, 91–92  
 Coefficient of determination, 29  
 Collapsed graphs, 173  
 Collider bias, 170  
 Colliders  
     in causal graphs, 167*f*, 168–169  
     d-separation criterion and, 170, 172–173  
 Collinearity  
     extreme collinearity, 71–72, 157  
     multicollinearity, 427  
 Combination rule, 277  
 Combined Frequency Index, 400, 401  
 Common method variance, 93  
 Common metric completely standardized solutions, 395  
 Common metric standardized solutions, 395

- Common variance, 190  
 Communality, 190  
 Compact symbolism, 431  
 Comparative fit indexes, 266. *See also* Bentler  
     Comparative Fit Index; Incremental fit indexes  
 Complex indicators, 195–196  
 Complex models  
     identification, 158–159, 457  
     specification, 456  
 Complex sampling design, 444–447, 457  
 Composite indicators, 353, 355, 359  
 Composite reliability (CR), 313–314  
 Computerized air traffic controller task (research example)  
     analysis of the latent growth model, 375–387  
     latent growth model compared with a polynomial growth model, 387–390  
 Computer languages. *See also* R programming language  
     SIMPLIS, 105, 357, 411  
     for symbolic processing, 149  
     tips for SEM programming, 100–101  
 Computer tools  
     advantages and drawbacks, 97–98  
     analysis of JWK models, 23  
     best practices in estimation, 461  
     bootstrap methods, 62  
     constrained estimation, 254  
     estimation of CFA models, 326–332  
     graphical editors, 98, 99–100  
     human-computer interactions, 98–99  
     maximum likelihood estimation, 236–238  
     recursive path model of illness example, 247–253, 254*t*, 255*f*  
     variations, 238–239  
 power analysis, 290–291  
 programs and procedures; *see also individual programs and procedures*  
     Amos, 102–103  
     EQS, 103–104  
     lavaan for R, 107–109  
     LISREL, 104–105  
     MATLAB, 111  
     Mplus, 105–106  
     Ωnyx, 106–109  
     overview, 9, 101–102  
     SAS/STAT, 109  
     Stata, 110  
     STATISTICA, 110–111  
     SYSTAT, 111  
 standardized solutions for SR models, 341  
 for the structural causal model, 112–113  
 summary, 113  
 symbolic processing, 149  
 tips for SEM programming, 100–101  
 Conditional independences  
     basis set, 173–174, 175*t*  
     concept and discussion of, 166–169  
     d-separation criterion and, 170–173  
     locating in directed cyclic graphs, 186  
     recursive path model of illness example, 240–241  
 Conditional indirect effects, 434  
 Conditional instruments, 181  
 Conditional process modeling, 432–434, 450  
 Confidence bands, 428  
 Confidence intervals  
     discussion of, 57–60  
     nonparametric bootstrapped, 61–62  
 Configural invariance  
     description, 396–397, 421  
     testing for, 399–400, 402, 406, 413, 415–416  
 Confirmation bias, 22, 55, 296, 466  
 Confirmatory factor analysis (CFA). *See also* Multiple-samples confirmatory factor analysis  
     analysis of measurement models in, 101  
     exploratory factor analysis and, 187, 188, 190–191, 192*f*, 198  
     item response theory as an alternative to, 332–333  
     overview and kinds of factor analysis, 189–191  
     research example, 206–207, 208*f*  
     summary, 207  
 Confirmatory factor analysis models  
     analyzing Likert-scale items as indicators, 323–332  
     constraint interaction in, 336–337  
     empirical underidentification, 206  
     equivalent models, 315–319  
     estimation  
         detailed example, 304–309, 310*t*, 311  
         empirical checks for identification, 303–304  
         interpretation of the estimates, 301–302  
         problems in, 302–303  
         types of standardized solutions, 302  
     identification  
         empirical checks for, 303–304  
         overview, 198  
         rules for nonstandard CFA models, 202–206  
         rules for standard CFA models, 201, 202*f*  
         scaling factors, 198–200  
     in the identification of SR models, 217, 218–219  
     latent variables in, 188–189  
     LISREL notation for, 210–211  
     in the modeling of SR models, 338–339, 340  
     multilevel, 449–450  
     respecification, 309–312  
     special models  
         bifactor models, 319–321  
         hierarchical models, 319, 320*f*  
         for multitrait–multimethod data, 321–323  
     special topics and tests, 312–315  
     specification  
         dimensionality, 195–196  
         directionality, 196–197  
         fallacies about factor labels, 300–301  
         indicator selection, 195  
         latent variables, 188–189  
         model characteristics, 193–195  
         understanding “exploratory” and “confirmatory,” 197  
         using CFA after EFA, 198  
     start value suggestions for measurement models, 335  
     structural invariance, 420  
     summary, 207

- Confirmatory tetrad analysis (CTA), 315  
 Confounding bias, 170  
 Congeneric indicators, 314  
 Consistent mediation, 247  
 Constrained baseline approach, 400  
 Constrained estimation, 253–254  
 Constrained optimization, 253–254  
 Constrained parameters, 129  
 Constraint interaction  
   CFA models, 312, 314, 336–337  
   importance of checking for, 462  
   SR models, 363  
 Constraints, model specification and, 456  
 Construct validity, 93  
 Content validity, 94  
 Contextual effects, estimation, 445  
 Continuous/categorical variable methodology, 324  
 Continuous indicators, 404–411  
 Continuous outcomes  
   nonnormal distribution, estimation methods, 256–257  
   normal distribution, estimation methods, 256  
 Continuous variables  
   defined, 42  
   interactive effects, 424–430, 431–432, 450  
   mean structures and, 372  
 Contracted chains, 166  
 Control group, 123  
 Controlled direct effect (CDE), 183, 184, 187, 435–437  
 Conventional medical model of recovery after cardiac surgery (research example), 287–289  
 Convergence criterion, 81  
 Convergent validity, 93  
 Corrected model test statistic, 238  
 Corrected normal theory method, 238  
 Correction for attenuation, 92–93  
 Correlated residuals, design-driven, specification, 455  
 Correlated trait–correlated method (CTCM) model, 321–323  
 Correlated-uniqueness (CU) model, 322f, 323  
 Correlation matrices  
   discussion of, 65–67, 95–96  
   fitting models to, 253–254  
 Correlation residuals  
   defined, 240  
   discussion of, 252–253, 254t, 255  
   tips for inspecting in fit testing, 278  
 Correlations  
   canonical, 17  
   disattenuating, 127  
   observed versus estimated, 41–44  
   partial and part, 39–41  
   Pearson correlation, 41–42, 43  
   predicted, 250–253  
   representation in diagrams, 121  
 Correlation size, impact on the model chi-square, 271  
 Counterfactuals, 18–19, 123  
 Counting rules, 145–146, 152–153  
 Count variables, 79  
 Covariance Analysis of Linear Structural Equations (CALIS), 102t, 109  
 Covariance equivalence, 293  
 Covariance matrices, 65–67, 95–96  
 Covariance residuals, 252–253. *See also* Correlation residuals  
 Covariances  
   analysis in SEM, 13–14  
   predicted, 250–253  
   representation in diagrams, 121–122  
 Covariance structure, 14  
 Covariance structure analysis, 9  
 Covariance structure modeling, 9  
 Covariates  
   back-door criterion, 177–179  
   minimally sufficient set, 178  
   in SR models, 355  
   sufficient set, 177  
 Criterion-related validity, 93  
 Critical ratios, 51–52  
 Cronbach's alpha, 91–92, 313  
 Cross-domain change, 391  
 Cross-factor equality constraint, 312, 314  
 Cross-group equality constraint, 129, 421  
 Cross-lagged paths, 139  
 Cross-level interactions, 445, 447  
 Cross-sectional designs, 134–135, 455  
 cspower macro, 290  
 Curve-of-factors model, 390  
 Curvilinear effects, 20, 432  
 Curvilinear growth, 375  
 DAGitty, 112  
 DAG (directed acyclic graph) Program, 112, 179  
 dagR package, 112–113  
 Data. *See also* Hierarchical data; Input data;  
   Longitudinal data; Missing data; Raw data  
   analyzed in SEM, 13–14  
   best practices, 458–461  
   longitudinal, path models for, 138–141  
   simulated, 459  
   time structured, 375  
 Data-based methods for missing data, 87–88  
 Data loss mechanisms, 83–84  
 Data matrices  
   best practices, 460–461  
   determinants, 68  
   ill-scaled covariance matrix, 81–82  
   out of bounds elements, 67–68  
   positive definite and nonpositive definite, 67–71  
   singular and nonsingular, 69  
   summaries of raw data, 65–67, 95–96  
 d-Connected variables, 171  
 Degrees of freedom  
   for continuous variables, 25  
   model degrees of freedom, 128, 145–148  
 Delta scaling, 326–327, 329  
 Demographic variables, 217  
 Dependent variables, 119. *See also* Endogenous variables  
 Depression, single-factor CFA model with ordinal indicators (research example), 328–330, 414–420

- Descendants, 166
- Determinants, of data matrices, 68
- Deterministic causality, 11
- Diagonally weighted least squares, 258
- Differential additive (acquiescence) response style, 398
- Differential item functioning (DIF), 398, 420–421
- Differential test functioning (DTF), 420–421
- DIGitty, 179
- Dimensional invariance, 397n2
- Directed acyclic graphs (DAG). *See also* Structural causal models
- basis set, 173–174, 175t
  - defined, 165
  - d-separation criterion, 170–173
  - elementary structures and conditional independences, 166–169
  - graphical identification criteria, 177–180
- Directed cyclic graphs (DCG), 165, 173, 186. *See also* Structural causal models
- Directed graphs
- acyclic (*see* Directed acyclic graphs)
  - causal directed graphs, 174–177
  - cyclic, 165, 173, 186
  - elementary structures and conditional independences, 166–169
  - vocabulary, 165–166
- Directed path, 166
- Direct effect and first-stage moderation, 434
- Direct effect and second-stage moderation, 434
- Direct effects
- causal mediation analysis, 435–437
  - conditional process modeling, 432–434
  - counterfactual definitions of, 187
  - in EFA models, 191
  - identification in SCM
    - graphical identification criteria, 177–180
    - instrumental variables, 180–181  - natural or controlled, 183, 184, 187
  - in path analysis, 232–233
  - in path models, 131–133
  - representation in diagrams, 121
  - representation in graph theory, 165
- Direct feedback loops, 135, 150–153
- Disattenuating correlations, 127
- Disconfirmatory procedures, 21
- Discriminant validity, 93–94
- Disturbance correlations
- bow patterns and bow-free patterns, 136
  - identification of models with, 150–153
  - in nonrecursive models, 136, 138
- Disturbance covariances
- in nonrecursive models, 136
  - start value suggestions, 261
- Disturbances
- defined, 130
  - distinct from regression residuals, 131
  - in graph theory, 166
  - in nonrecursive models, 138
  - representation in model diagrams, 130–131
  - scaling, 148
- Disturbance variances
- estimation, 130
  - scaling, 148
  - start value suggestions, 261
- Domain sampling model, 196
- Drawing editors. *See* Graphical editors
- d-Separation
- with causal directed graphs, 175–176
  - graphical identification criteria in SCM and, 177
  - instruments and, 180
  - local independence and, 188–189
- d-Separation criterion
- basis set and, 173–174, 175t
  - discussion of, 170–173
- d-Separation equivalence, 293
- Edges, 165
- EFA. *See* Exploratory factor analysis
- Effect decomposition, 364, 464
- Effect indicators, 196, 352, 353f, 354–355, 359
- Effects coding method, 200, 405–406
- Efficient estimators, 232
- Eigendecomposition, 67
- Eigenvalues, 67–68
- Eigenvectors, 67
- Elliptical distribution estimators, 256
- Elliptical distribution theory, 256
- Empirical growth record, 375
- Empirical sampling distribution, 61
- Empirical underidentification, 157–158, 206, 463
- Endogeneity, 132
- Endogenous instruments, 152
- Endogenous selection bias, 170
- Endogenous variables
- continuous, mean structures and, 372
  - disturbances, 130–131
  - order condition, 152–153
  - partitioning for evaluating the identification status of nonrecursive models, 154–155
  - in path models, 129–131
  - rank condition, 153, 161–163
  - reduced form, 365
  - representation in diagrams, 122
  - specification, 119
- Entanglement, 123n
- EQS (Equations) program
- analysis of CFA models, 330
  - constrained estimation, 254
  - identification of multivariate outliers, 73
  - incomplete data procedures, 88
  - independence model in, 266
  - modification indexes, 283
  - overview and description, 102t, 103–104
  - power analysis, 291
  - scale chi-square difference test, 282
  - using with SEM, 9
- Equal-fit hypothesis, chi-square difference test, 281, 283
- Equality constraint, 129, 156
- Equilibrium assumption, 124, 364

- Equivalent models, 22  
 discussion of, 292–297  
 equivalent CFA models, 315–319  
 equivalent SR models, 348–349  
 evaluating the favored model against, 120  
 importance of reporting on, 466
- Error correlations, in CFA models, 196, 202–206
- Error covariance structure, 138
- Error terms, 13
- Essential multicollinearity, 427
- Estimated correlations, versus observed correlations, 41–44
- Estimated power, 290
- Estimation  
 best practices, 461–463  
 causal effects in path analysis, 232–233  
 of CFA models, 301–309  
 constrained estimation, 253–254  
 of contextual effects, 445  
 a healthy perspective on, 258–259  
 of the interactive effects of latent variables  
   alternative methods, 442–444  
   Kenny–Judd method, 439–442  
 of Likert-scale items as indicators, 323–332  
 maximum likelihood estimation, 235–239 (*see also*  
   Maximum likelihood estimation)  
 of mean structures, 374  
 recursive path model of illness example  
   conditional independences, 240–241  
   estimation with maximum likelihood, 247–253,  
     254*t*, 255  
   overview, 239–240  
   single-equation estimation with multiple  
     regression, 241–247  
 in SEM, 118*f*, 120  
 single-equation methods, 231, 233–234, 241–247  
 summary, 259  
 types of estimators, 231–232
- Estimators  
 alternative, 255–258  
 efficient, 232  
 types of, 231–232
- Exact fit, 60
- Exact-fit hypothesis  
 model chi-square, 270–273  
 model test statistics, 265–266
- Exogeneity, 129
- Exogenous instruments, 152
- Exogenous variables  
 adding in the respecification of nonrecursive  
 models, 157  
 continuous, mean structures and, 372  
 in path models, 129–130  
 representation in diagrams, 121–122  
 specification, 119
- Expectation–maximization (EM) algorithm, 88, 443
- Expected parameter change, 284
- Experimental designs  
 establishing causal inference, 122–125, 126  
 random, 123
- Explaining away effect, 169
- Exploratory factor analysis (EFA)  
 as an alternative to invariance testing, 421  
 analysis of Likert-like items, 332  
 CFA and, 187, 188, 190–191, 192*f*, 198  
 not using CFA as a follow-up analysis, 198  
 origin of, 23  
 understanding “confirmatory” and “exploratory,”  
   197
- Exploratory factor analysis models  
 characteristics of, 191–193  
 in four-step modeling of SR models, 339–340
- Exploratory structural equation modeling (ESEM),  
 219–220
- Exploratory tetrad analysis (EFA), 315
- Extreme collinearity, 71–72, 157
- Extreme response style, 397
- Factor analysis. *See also* Confirmatory factor analysis;  
 Exploratory factor analysis  
 description, 189–190  
 historical overview, 23  
 kinds of, 190–191, 192*f*
- Factor indeterminacy, 189
- Factor labels, fallacies about, 300–301
- Factor loadings, 191
- Factor-of-curves model, 391
- Factor rho coefficient, 313
- Factors  
 in CFA models  
   analysis of and measurement reliability, 312–315  
   respecification, 310  
   scaling, 198–200  
   defined, 13  
   first-order, 319, 320, 321  
   second-order, 319
- Factor score indeterminacy, 193
- Faithfulness assumption, 170
- F distributions, 58–59
- Feedback loops  
 direct, 135, 150–153  
 indirect, 135, 150, 151  
 nonrecursive models with, identification, 150–153
- First-and-second stage moderation, 434
- First-order factors, 319, 320, 321
- First-order part correlation, 39–40
- First-order partial correlation, 39
- First-stage moderation, 433*f*, 434
- Fisher information matrix, 304
- Fit function, 235
- Fit statistic tunnel vision, 264
- Fitted covariances, 250–253
- Fitted residuals, 252–253. *See also* Correlation residuals
- Fit testing. *See also* Global fit testing; Local fit testing  
 discussion of and research about, 262–263  
 person-level fit, 264–265
- Fixed parameters, 128
- Fixed-weights composite, 355
- Forks, 167–168
- Formative measurement, 197, 456
- Formative measurement models, analyzing in SEM,  
 352–361

- Forward inclusion, 38  
 Four-step modeling, of fully latent SR models, 339–340  
 Free baseline approach, 399–400  
 Free parameters, 128  
 Frequency weights, computer tools and, 102  
 Front-door path, 166  
 Full information maximum likelihood (FIML), 87, 88, 331  
 Full-information methods, 231–232. *See also* Simultaneous estimation methods  
 Fully latent structural regression models  
     best practices in estimation, 462  
     equivalent, 348  
     four-step modeling, 339–340  
     identification, 217–219  
     overview and description, 213, 214f, 223  
     research example, job satisfaction factors, 220–222  
     two-step modeling, 338–339, 341–347  
 Fully weighted least squares (WLS), 256  
 Generalized appended product indicator (GAPI)  
     method, 443  
 Generalized least squares (GLS), 256  
 Generalized linear latent and mixed models  
     (GLAMM), 18  
 General linear model (GLM), 17–18  
 General-specific models, 319–321  
 Global fit testing  
     categories of  
         approximate fit indexes, 266–268  
         model test statistics, 265–266  
     difficulties and limitations, 263–265  
     discussion of fit testing, 262–263  
     recommend approach to fit evaluation  
         Bentler Comparative Fit Index, 276–277  
         model chi-square, 270–273  
         overview, 268–269  
     Standardized Root Mean Square Residual, 277–278  
     Steiger–Lind Root Mean Square Error of Approximation, 273–276  
     tips for inspecting residuals, 278  
     recursive path model of illness example, 278–280  
     summary, 297  
 Goodness-of-fit statistics, 266, 276–277  
 Graphical editors, 98, 99–100  
 Graphical identification criteria, in SCM, 177–180  
 Graphical rules, for identifying nonrecursive models, 153–155, 156f  
 Graph theory. *See also* Structural causal models  
     basis set, 173–174, 175t  
     causal directed graphs, 174–177  
     elementary directed graphs and conditional independences, 166–169  
     graphical identification criteria, 177–180  
     implications for regression analysis, 170–173  
     introduction to, 164–166  
     model identification and, 149  
     summary, 184–185  
     vocabulary, 165–166  
 Group means, comparing on observed variables, 462–463  
 Group-mean substitution, 86  
 gsem command, 102t, 110  
 Half longitudinal design, 140  
 Heteroscedasticity, 80–81  
 Heywood cases, 237–238  
 Hierarchical (nested) data  
     best practices, 461  
     multilevel modeling, 444–450  
 Hierarchical linear modeling (HLM), 374, 375. *See also* Multilevel modeling  
 Hierarchical models  
     chi-square difference test, 281–283  
     defined, 280  
     empirical versus theoretical respecification, 283–284  
     model building, 280–281, 285–286  
     model trimming, 280–281  
     specification scales, 284–285  
 Hierarchical regression, 38  
 Homoscedasticity, 80–81  
 Human-computer interactions, 98–99  
 Hypothesis testing, of hierarchical models, 285–296  
 Hypothetical constructs, 12–13. *See also* Latent variables  
 Identification  
     best practices, 457  
     of CFA models, 198–206, 303–304  
     empirical underidentification problems, 157–158  
     general requirements, 145–148  
     graph theory and, 149  
     a healthy perspective on, 157  
     managing problems, 158–159  
     of mean structures, 373–374  
     of nonrecursive models  
         with feedback loops and all possible disturbance correlations, 150–153  
         graphical rules for other types, 153–155, 156f  
         overview, 138, 150, 159–160  
         respecification of models not identified, 155–157  
     path analysis resistance example, 159  
     of recursive path models, 149–150  
     in SEM, 118f, 119–120  
     of SR models, 217–219, 225  
     of structural causal models, 119–120  
     summary, 159–160  
     undecidable problem, 149  
     unique estimates of parameters, 148–149  
 Identification heuristics  
     for CFA models  
         nonstandard, 202–206  
         standard, 201, 202f  
     for nonrecursive models, 150–153  
     overview, 149  
         for recursive path models, 149–150  
 Identity link, 44  
 Ignorable data loss mechanisms, 84  
 Ill-scaled covariance matrix, 81–82  
 Inadmissible solutions, in ML estimation, 237–238  
 Incomplete data. *See* Missing data  
 Inconsistent mediation, 247  
 Incremental fit indexes, 266, 276–277

- Independence model, 266  
 Independent variables, 119, 123. *See also* Endogenous variables  
 Indicant product approach, 437–439  
 Indicator labels, fallacies about, 300–301  
 Indicators  
   analyzing Likert-scale items as, 323–332  
   best practices in specification, 454–456  
   causal indicators, 352–355, 356–360  
   in CFA models  
     dimensionality, 195–196  
     directionality, 196–197  
     respecification, 310  
     rules for nonstandard CFA models, 202–206  
     rules for standard CFA models, 201, 202f  
     selection, 195  
     testing, 314  
     vanishing tetrads, 315  
   composite indicators, 353f, 355, 359  
   defined, 13  
   effect indicators, 352, 353f, 354–355, 359  
   measurement invariance testing example  
     with continuous indicators, 404–411  
     with ordinal indicators, 411–420  
   in SR models, 213–217  
 Indirect effects  
   best practices in interpreting, 465  
   best practices in reporting, 464  
   causal mediation analysis, 435–437  
   conditional process modeling, 432–434  
   counterfactual definitions of, 187  
   mediation and, 142  
   versus moderation, 134–135  
   natural, 183–184, 187  
   in path analysis, 232–233  
   in path models, 133–135  
   Sobel test, 245  
 Indirect feedback loops, 135, 150, 151  
 Inequality constraint, 129  
 Initial latent growth factors  
   in basic change models, 377–384  
   in polynomial growth models, 388–390  
   in prediction models, 384–387  
 Input data  
   best practices, 458–459  
   extreme collinearity, 71–72  
   forms of, 64–67  
   linearity and homoscedasticity, 80–81  
   missing data, 82–88; *see also* Missing data  
   normality, 74–77  
   outliers, 72–73  
   positive definiteness, 67–71  
   relative variances, 81–82  
   summary, 95–96  
   transformations, 77–81  
 Inputs, to SEM, 9–10  
 Instrumental variables, 150–152, 180–181  
 Instrumental variables regression, 234  
 Instruments  
   adding in the respecification of nonrecursive models, 157  
   conditional, 181  
   defined, 180–181  
   exogenous or endogenous, 152  
   overview and discussion of, 150–152  
 Interactive effects  
   curvilinear effects and, 432  
   importance of explaining, 466  
   of latent variables  
     alternative estimation methods, 442–444  
     estimation in the Kenny–Judd method, 439–442  
     indicant product approach, 437–439  
     model specification and, 456  
     of observed variables, 424–430  
     in path models, 431–432  
     SEM and, 20  
     summary, 450  
 Intercepts, in regression analysis of means, 369–370, 372–373  
 Intercepts-as-outcomes model, 447  
 Internal consistency reliability, 91–92  
 Interpretational confounding, 339  
 Interpretation–use arguments, 93  
 Interrater reliability, 92  
 Interval estimation, 57–60  
 Inverse function transformations, 78  
 Inverse probability fallacy, 56  
 Inverted forks, 167f, 168  
 Isolation, of the independent variable, 123  
 Item characteristic curve (ICC), 94–95  
 Item response theory (IRT)  
   as an alternative to invariance testing, 420–421  
   as an alternative to item-level CFA analysis, 332–333  
   computer tools and, 18  
   discussion of, 94–95  
 Iterative estimation, 101, 236–237, 259  
 Jangle fallacy, 301, 458  
 Jingle fallacy, 301, 458  
 Job satisfaction factors study (research example), 220–222, 341–347  
 Just-identified (just-determined) models  
   with all possible disturbance correlations, 150, 151f  
   defined and described, 146, 147  
   structural equation models, 147  
   structural regression models, 345–346, 348–349  
 JWK models, 23  
 Kant, Immanuel, 123n  
 Kaufman Assessment Battery for Children (KABC-I),  
   CFA model for, 206–207, 208f  
   analysis of, 304–309, 310t, 311t  
   respecification of, 310–312  
 Kenny–Judd method, 439–442  
 Kurtosis, 74, 75f, 76–77  
 Kurtosis index, 76–77  
 Lagrange Multiplier (LM), 283  
 Latent class analysis, 18  
 Latent class regression, 18

- Latent curve models, 374–375. *See also* Latent growth models
- Latent growth factors. *See also* Initial latent growth factors
- in basic change models, 377–384
  - linear, 387–390
  - in prediction models, 384–387
  - summary, 392
- Latent growth models (LGM)
- best practices in estimation, 462
  - defined and described, 374–375
  - detailed example
    - comparison with a polynomial growth model, 387–390
    - modeling change, 376–384
    - overview, 375–376
    - predicting change, 384–387
  - extensions of, 390–392
  - summary, 392
- Latent moderated structural equations (LMS) method, 443, 444
- Latent response variables
- scaling, 326–327
  - in WLS estimation, 324–327
- Latent transition model, 18
- Latent variable models, 18, 454
- Latent variables
- categorical, 18
  - causal inference with, 212
  - in CFA, 188–189
  - disturbances as, 130
  - extreme collinearity, 72
  - interactive effects, 437–444, 450
  - means of, estimation, 14
  - overview and description, 12–13
  - representation in diagrams, 121
  - scaling, model specification and, 457
- lavaan package
- analysis of a nonrecursive model, 349–352
  - description, 102t, 107–108
  - robust WLS estimation, 326–327
- lava package, 102t, 108
- Least squares criterion, 26
- Lee–Hershberger replacing rules, 293–296
- Left-out variables error, 35–36
- Leptokurtic distributions, 74, 75f
- Likelihood ratio chi-square, 270
- Likert scale, 257
- Likert-scale items. *See also* Ordinal indicators
- analyzing as indicators, 323–332
- Limited-information models, 231. *See also* Single-equation estimation methods
- Linearity, 80–81
- Linear latent growth factor, 387–390
- Link function, 44
- Links, 165
- LISREL III program, 23
- LISREL program
- all-Y notation, 227–228
  - analysis of a model of risk as a latent variable with causal indicators, 357–360
  - analysis of JWK models, 23
- analysis of latent variable models, 18
- CFA models
- analysis of the Kaufman Assessment Battery for Children example, 306
  - notation for, 210–211, 300
  - standardized solution, 302
- characteristics of, 102t
- incomplete data procedures, 85, 87, 88
- independence model in, 266
- measurement invariance testing, 411
- model chi-squares, 272–273, 299
- multiple-samples analyses, 395
- notation for path models, 143–144
- notation for SR models, 226–228
- overview and description, 104–105
- power analysis, 291
- PRELIS and, 104 (*see also* PRELIS program)
- recursive path model of illness example
- global fit statistics, 278–280
  - maximum likelihood estimation, 248–250, 253, 254t, 255f
  - ridge adjustment, 70–71
  - robust WLS estimation, 328, 330
  - scale chi-square difference test, 282
  - standardized solution of SR models, 341
  - using with SEM, 9
- Listwise deletion, 85–86
- L → M block, 352, 353f
- Loadings, 191
- Local fit testing
- best practices, 462
  - fitting models to correlation matrices, 253–254
  - recursive path model of illness example
    - conditional independences, 240–241
    - estimation with maximum likelihood, 247–253, 254t, 255f
    - overview, 239–240
    - single-equation estimation with multiple regression, 241–247
    - summary, 259
- Local independence, 188–189
- Local Type I error fallacy, 56
- Logarithmic transformations, 78
- Logistic function, 44–46
- Logistic regression, 44–46, 47
- Logit, 44, 45, 46
- Longitudinal data
- latent growth models and, 374 (*see also* Latent growth models)
  - path models, 138–141
- Longitudinal independence model, 277
- Longitudinal measurement invariance, 396
- Lower diagonal matrices, 65, 66t
- MacCallum–Browne–Sugawara power analysis, 290
- Mahalanbois distance, 73
- Manipulation-of-mediation model, 182–183
- MANOVA (multivariate ANOVA), 17, 380
- Marginal association, 167
- Marker variable, 194
- Marker variable method, 404–405
- Markov assumption, 167

- Markov Chain Monte Carlo (MCMC), 88  
 Markov chains, 166–167  
 Marlowe–Crowne Social Desirability Scale, 323  
 Masking, 72  
 Match-pairs indicators, 442  
 MATLAB, 111  
 Matrices. *See* Data matrices  
 Maximum likelihood, defined, 235  
 Maximum likelihood (ML) estimation  
     analysis of categorical outcomes and, 257–258  
     assumption of multivariate normality, 74  
     full-information version, 331  
     inadmissible solutions and Heywood cases, 237–238  
     for incomplete data, 65  
     iterative estimation and start values, 236–237  
     OLS estimation and, 236  
     other requirements, 238  
     overview and description, 235–236  
     recursive path model of illness example, 247–253,  
     254*t*, 255  
     scale freeness and scale invariance, 238  
     unstandardized variables assumption, 253  
     variance estimates, 236  
     variations, 238–239  
 McDardle–McDonald reticular action model (RAM), 121  
 $M \rightarrow C$  block, 353*f*, 355  
 McDonald's noncentrality index (NCI), 400, 401  
 Mean-adjusted least squares (WLSM)  
     description, 327–328  
     example, 328–330  
 Mean- and variance-adjusted weighted least squares  
     (WLSMV), 327–328  
 Mean residuals, 374  
 Means  
     of latent variables, estimation, 14  
     logic for analyzing, 369–373, 392  
     moments about the mean, 76  
 Mean structures  
     defined, 14  
     estimation of, 374  
     identification of, 373–374  
         in multiple-samples CFA, 404–406  
     logic of, 369–373  
     model specification and, 456  
     summary, 392  
 Mean substitution, 86  
 Measurement  
     best practices, 458  
     formative, 197  
     reflective, 196–197  
     SEM and, 8  
     unidimensional or multidimensional, 195  
 Measurement error  
     in path models, 131  
     propagation of, 33–34  
     SEM and, 13  
 Measurement invariance  
     alternative statistical techniques, 420–421  
     examples of testing  
         with continuous indicators, 403–411  
         with ordinal indicators, 411–420  
     overview, 463  
     failure of research to evaluate all aspects of, 399  
     overview and description, 396  
     summary, 421  
     testing strategy and related issues, 399–402  
     types of, 396–399  
 Measurement models  
     formative, analyzing in SEM, 352–361  
     reflective measurement models, 352, 353*f*, 455–456  
     restricted measurement models, 191, 192*f*, 193–195  
     start value suggestions for, 335  
     unrestricted measurement models, 191–193  
 Measurement-of-mediation model, 182, 183  
 Measures  
     best practices, 458  
     checklist for evaluation, 89  
     reporting the psychometric characteristics of, 88,  
     90  
     selection, 88–90, 127  
 Median absolute deviation (MAD), 72–73  
 Mediated moderation, 432, 433*f*  
 Mediation  
     causal mediation analysis, 181–184  
     consensus on the analysis of, 141–142  
     consistent or inconsistent, 247  
     estimating with full longitudinal designs, 140–141  
     estimating with half longitudinal designs, 140  
     indirect effects and, 142  
     mediation package, 437  
 Mediators  
     causal mediation analysis, 181–184  
     defined, 134  
     in full longitudinal designs, 140, 141  
     in half longitudinal designs, 140  
     presumed, 135  
*Mental Measurements Yearbook* (Carlson, Geisinger, & Jonson), 88  
 Metric invariance. *See* Weak invariance  
 MIMIC factors. *See* Multiple-indicators and multiple-causes factors  
 Minimally sufficient set, 178  
 Minimum fit function chi-square, 270  
 Minimum sample size, estimation in power analysis,  
     290, 291–292  
 Missing at random (MAR), 83–84, 85, 87  
 Missing completely at random (MCAR), 83, 84, 85,  
     87  
 Missing data  
     best practices in handling, 458, 460  
     classical methods for handling, 85–87  
     data loss mechanisms, 83–84  
     diagnosing, 84–85  
     overview, 82–93, 95  
 Missing not at random (MNAR), 84, 85, 87  
 Misspecification, global fit statistics and, 264  
 Mixture models, 18  
 $M \rightarrow L$  block, 352, 353*f*  
 ML estimation. *See* Maximum Likelihood estimation;  
     Maximum likelihood estimation  
 Model-based analysis of missing data, 87  
 Model building  
     description, 280–281  
     recursive path model of illness example, 285–286

- Model chi-square  
 best practices in reporting on, 464  
 modification indexes and, 283  
 overview and description, 269, 270–273  
 printed by LISREL, 299  
 recursive path model of illness example, 278–279, 285  
 Wald W statistic and, 284
- Model degrees of freedom, 128, 145–148
- Model diagrams  
 overview of symbols, 121–122  
 representation of disturbances, 130–131
- Model estimation. *See* Estimation
- Model generation, 11
- Model identification. *See* Identification
- Modeling school of statistics, 262
- Model respecification. *See* Respecification
- Models  
 complexity, 127–128  
 probabilistic causality and, 11–12  
 SEM and protection against equivalent models, 22  
 SEM as a disconfirmatory procedure, 21  
 testing of, SEM and, 11
- Model selection uncertainty, 289
- Model specification. *See* Specification
- Model test statistics  
 approximate fit indexes and, 267  
 overview and description, 265–266  
 recursive path model of illness example, 278–280
- Model trimming  
 description, 280–281  
 Wald W statistic, 284
- Moderated mediation, 433f, 434
- Moderated multiple regression, 424–430
- Moderated path analysis, 431–432, 450
- Moderation, 133, 134–135
- Modification indexes  
 description, 283–284  
 recursive path model of illness example, 285–286  
 respecification of CFA models and, 310–312
- Modified weighted least squares, 258
- Modularity assumption, 124
- Moments about the mean, 76
- Monotonic transformation, 77
- Monte Carlo methods, 291
- Mplus program  
 analysis of latent variable models, 18, 357–360  
 causal mediation analysis, 437  
 CFA models  
   analysis of the Kaufman Assessment Battery for Children example, 304–309, 311t  
   standardized solution, 302  
 characteristics of, 102t  
 compact symbolism for interactive effects in path models, 431  
 constrained estimation, 254  
 “difftest” option, 272, 276, 282  
 incomplete data procedure, 87  
 independence model in, 266  
 Kenny–Judd method for estimating structural equation models, 441–442
- measurement invariance testing  
 with continuous indicators, 406–411  
 with ordinal indicators, 414–420  
 overview and description, 105–106  
 path model estimation, 248  
 power analysis, 291  
 robust maximum likelihood estimation, 239  
 robust WLS estimation, 327–330  
 scale chi-square difference test, 282  
 standardized solution of SR models, 341  
 two-level regression analysis, 445, 446f  
 using with SEM, 9  
 WLS estimation, 326–327
- Multi-agent estimation algorithm, 9
- Multicollinearity, essential or nonessential, 427
- Multidimensional measurement, 195
- Multilevel CFA models, 449–450
- Multilevel modeling (MLM)  
 convergence with SEM, 447–450  
 limitations, 447  
 overview and description, 444–447
- Multilevel path models, 448–449
- Multilevel structural equation modeling (ML-SEM), 448–450
- Multiple factor CFA models, 315–317. *See also* Two-factor CFA models
- Multiple imputation, 87–88
- Multiple-indicator measurement, 127, 454
- Multiple-indicators and multiple-causes (MIMIC) factors, 318–319, 354
- Multiple optima, 9
- Multiple regression  
 assumptions, 33–35  
 corrections for bias, 32–33  
 description, 30–32  
 logic for analyzing means, 369–373  
 moderated, 424–430  
 single-equation estimation method  
   overview and description, 233–234  
   recursive path model of illness example, 241–247
- Multiple-samples confirmatory factor analysis  
 goal of, 396  
 measurement invariance, 396–399 (*see also* Measurement invariance)  
 methods to scale the factor and identify the mean structure, 404–406  
 model specification and, 457  
 summary of, 421
- Multiple-samples SEM analysis, 394–396, 421, 457, 462
- Multitrait–multimethod (MTMM) studies, 321–323
- Multivariate ANOVA (MANOVA), 17, 380
- Multivariate non-normality, 74–77, 271
- Multivariate normal distributions, 256
- Multivariate normality (multinormality), 74
- Multivariate outliers, 73
- MxModel objects, 108–109
- Mx package, 108
- Naming fallacy, 300, 466
- Natural direct effect (NDE), 183, 184, 187, 435–437
- Natural indirect effect (NIE), 183–184, 187, 435–437

- Near-equivalent models, 120, 297, 466  
 Negative kurtosis, 74, 75f  
 Negative skew, 74, 75f, 78  
 Negative suppression, 37  
 Nested data. *See* Hierarchical data  
 Nested-factor models, 319–321  
 Nested models, 280–286. *See also* Hierarchical models  
 Neyman–Rubin model, 18–19  
 Nil hypothesis, 53–54  
 Nodes, 165  
 Nonadjacent vertices, 165  
 Noncentral Distributional Calculator (NDC), 59–60  
 Noncentral distributions, 58–60  
   chi-square, 290  
   F, 58–59  
 Noncentrality index (NCI), 400, 401  
 Noncentrality parameters, 58  
 Noncontinuous outcomes, 20  
 Nondeterministic function of observed variables, 189  
 Nonessential multicollinearity, 427  
 Nonexperimental designs, 124–125, 126  
 Nonhierarchical models, 286–289, 297  
 Nonignorable data loss mechanisms, 84  
 Nonlinear constraints, 129  
 Nonlinear curve fitting, 377–384, 392  
 Non-nil hypothesis, 54  
 Non-normal distributions, 256–257  
 Non-normed fit index, 276–277  
 Nonparametric bootstrapped confidence intervals, 61–62  
 Nonparametric bootstrapping, 60–62, 239  
 Nonpositive definite data matrix, 67, 68, 69–71  
 Nonrecursive panel models, 139  
 Nonrecursive path models  
   complications of, 137–138  
   directed cyclic graphs and, 165  
   overview and description, 135–137  
 Nonrecursive structural models  
   corrected proportions of explained variance for, 365–366  
   effect decomposition and the equilibrium assumption, 364  
   identification  
     with feedback loops and all possible disturbance correlations, 150–153  
     graphical rules for other types, 153–155, 156f  
     overview, 150, 159–160  
   respecification, 155–157  
   single indicators in a model of organizational and occupational turnover intention, 222–223, 224f, 349–352  
 Nonsingular data matrix, 69  
 Nonstandard CFA models, 202–206, 207  
 Normal deviates, 25  
 Normality, 74–77  
 Normality assumption, 51  
 Normalized residuals, 252–253  
 Normalizing transformations, 77–79  
 Normal ogive model, 46  
 Normal probability plots, 76  
 Normal theory methods, 77, 256  
 Normed chi-square, 272  
 Not-close-fit hypothesis, 275, 290  
 N:*q* rule, 16  
 Null hypotheses, 52–54  
 Null model, 266  
 Numerical integration, 258  
 Oblique rotation, 193  
 Observational equivalence, 293  
 Observations  
   minimum degrees of freedom and model identification, 145–148  
   model complexity and, 127–128  
   rule for counting, 373  
 Observed correlations, versus estimated correlations, 41–44  
 Observed variable analyses, 91, 92, 312  
 Observed variable models. *See* Path models  
 Observed variables  
   comparing group means on, 462–463  
   interactive effects  
     estimation of, 424–428  
     extensions and challenges, 430  
     interpretation of, 428–430  
     summary, 450  
   nondeterministic function of, 189  
   overview and description, 12, 13  
   in reflective measurement models, 196  
   representation in diagrams, 121  
 Occupational turnover intention (research example), 222–223, 224f  
 Odd-powered polynomial transformations, 78  
 Odd-root function transformations, 78  
 Odds  
   against chance fallacy, 55–56  
   in logistic regressions, 44–45  
 Odds ratio, 44–45  
 OLS. *See* Ordinary least squares estimation  
 One-sided null hypothesis, 274  
 One-step modeling, of fully latent SR models, 338–339  
 QNyx program, 9, 102t, 106–109  
 OpenMX package, 102t, 108–109  
 Optima, multiple, 9  
 Order condition, 152–153  
 Ordered-categorical indicators, analyzing Likert-scale items as, 323–332  
 Ordered-categorical outcomes, 257  
 Ordinal data, best practices in estimation, 461  
 Ordinal indicators. *See also* Likert-scale items  
   measurement invariance testing example, 411–420  
   robust WLS estimation example, 328–330  
 Ordinary least squares (OLS) estimation  
   in bivariate regression, 26  
   ML estimation and, 236  
   overview and description, 233–234  
   recursive path model of illness example, 241–247  
   variance estimates, 236  
 Organizational and occupational turnover intention (research example), 222–223, 224f, 349–352  
 Orthogonality, test for, 314  
 Orthogonal rotation, 192–193  
 Outcome variables, 119. *See also* Endogenous variables  
   best practices in reporting on, 464  
   noncontinuous, SEM and, 20

- Outliers, 72–73  
 Out of bounds matrix elements, 67–68  
 Outputs, from SEM, 10  
 Overcontrol bias, 170  
 Overidentified (overdetermined) models  
     with all possible disturbance correlations, 150, 151f  
     defined and described, 147, 148  
     structural equation models, 147  
 Overidentifying restrictions, 148
- Pairwise deletion, 86  
 Panel models, 138–139  
 Parallel-forms reliability, 92  
 Parallel growth process, 391  
 Parallel indicators, 314  
 Parameter estimates  
     best practices in interpreting, 465  
     best practices in tabulation, 464  
     interpretation in CFA models, 301–302  
     interpretation in SR models, 340–341  
     unique estimates and model identification, 148–149  
 Parameterization, 326–327  
 Parameters  
     estimation (*see* Estimation)  
     expected parameter change, 284  
     free, fixed, or constrained, 128–129  
     minimum degrees of freedom and model  
         identification, 145–148  
     model complexity and, 127–128  
     representation in diagrams, 122  
 Parametric bootstrapping, 62  
 Parceling, 331–332  
 Parcels, 332, 458  
 Parent–child conflict study (research example), 403–411  
 Parents, 166  
 Parsimony-adjusted indexes, 266–267  
 Parsimony principle, 128, 456  
 Parsimony ratio, 267  
 Part correlation, 39–41  
 Partial correlation, 39–41  
 Partial-information models, 231. *See also* Single-equation estimation methods  
 Partial least squares path modeling (PLS-PM), 360–361  
 Partially latent SR models  
     identification, 219  
     overview and description, 213–214, 215f, 223  
 Partially recursive models, 136  
 Partial measurement invariance, 401–402, 417  
 Path, defined, 166  
 PATH1 programming language, 110  
 Path analysis  
     causal effects in, 232–233  
     elemental models and assumptions, 131–134  
     identification example, 159  
     JWK model and, 23  
     overview and basics of, 129–131  
 Path coefficients, 132, 232, 261  
 Path models. *See also* Nonrecursive path models  
     identification example, 159  
     interactive effects, 431–432  
     with a mean structure, 371–373  
     multilevel, 448–449  
     overidentified, 148  
     single-equation estimators, 233–234  
     slopes-and-intercepts-as-outcomes, 446f  
     specification  
         elemental models and assumptions, 131–134  
         LISREL notation, 143–144  
         for longitudinal data, 138–141  
         overview and basics of, 129–131  
         recursive and nonrecursive models, 135–138  
         summary, 141–142  
 Pattern coefficients  
     in CFA models, 194  
     identification rules for nonstandard models, 202, 203, 204–205, 206  
     interpretation, 301–302  
     in EFA models, 191  
 Pattern invariance. *See* Weak invariance  
 Pattern matching, 87  
 Pearson correlation, 41–42, 43, 302  
 Percentage or proportion of maximum scoring (POMS) transformation, 79–80, 101  
 Perfect fit, 60  
 Person-level fit, 264–265  
 Phi coefficient, 42  
 Piecewise latent trajectory model, 390–391  
 Platykurtic distributions, 74, 75f  
 PLS–Graph program, 361  
 Point-biserial correlation, 42  
 Poisson distributions, 79  
 Poisson regression, 79  
 Polychoric correlations, 43, 325–326  
 Polynomial growth model, 387–390, 392  
 Polyserial correlation, 42–43  
 POM (potential outcomes model), 18–19  
 Poor-fit hypothesis, 275  
 Positive definite data matrix, 67–71  
 Positive kurtosis, 74, 75f  
 Positive skew, 74, 75f, 78  
 Potential outcomes model (POM), 18–19  
 Power, of null hypotheses, 52–53  
 Power analysis, 290–292  
 Power terms, 44  
 Predicted correlations, 250–253  
 Predicted covariances, 250–253  
 Prediction latent growth models, 384–387  
 Predictive fit indexes, 267, 286–289, 466  
 Predictors  
     assumption of no causal effects, 34  
     assumption of no measurement error, 33–34  
     assumption of no specification error, 35  
     categorization and pseudo-groups, 43–44  
     centering in moderated multiple regression, 425, 427–428  
     left-out variables error, 35–36  
     partial and part correlation, 39–41  
     in prediction latent growth models, 384, 385f  
     selection and entry, 37–39  
     suppression, 36–37  
     time-varying or time-invariant, 390

- PRELIS program  
 censored variables, 43  
 incomplete data procedures, 85, 87  
 LISREL and, 104 (*see also* LISREL program)  
 polyserial or polychoric correlations, 43  
 power analysis, 291  
 robust WLS estimation example, 330
- Principal components method, 191n
- Probabilistic causality, 11–12
- Probabilistic graph models, 164
- Probability weights, 102
- Probit function, 46
- Probit regression, 46–47
- Product estimators, 134
- Product indicators, 439–442
- Product terms  
 interactive effects in path analysis, 431, 432  
 moderated multiple regression, 424–430
- Prolog computer language, 149
- Propagation of measurement error, 33–34
- Propagation of specification error, 235
- Proportionality constraint, 129, 156
- Pseudo-groups, of predictors or outcomes, 43–44
- Pseudo-isolation, 27
- Psychometrics  
 item response theory and item characteristic curves, 94–95  
 reporting practices, 88, 90  
 score reliability, 90–93  
 score validity, 93–94  
 selection of measures, 88–90
- Psychosomatic model of recovery after cardiac surgery (research example), 287–289
- p* values. *See also* Significance testing  
 best practices in interpreting, 465  
 criticisms and controversies, 54–57
- Quadratic latent growth factor, 387–390
- Quasi-maximum likelihood (QML) estimation, 443
- RAMONA (Reticular Action Model or Near Approximation) procedure, 102t, 111, 254
- Random coefficient modeling. *See* Multilevel modeling
- Random experimental designs, establishing causal inference, 123–124
- Random hot-deck imputation, 87
- Rank condition, 153, 161–163
- Rasch model, 95
- Raw data  
 best practices, 459, 460  
 files, 65  
 matrix summaries, 65–67
- Reciprocal causation, estimating in nonrecursive models, 137–138
- Reciprocal suppression, 37
- Recursive path model of illness (research example)  
 global fit statistics, 278–280  
 model building, 285–286  
 near-equivalent models, 297  
 parameter estimation and local fit testing  
 conditional independences, 240–241
- estimation with maximum likelihood, 247–253, 254t, 255f  
 overview, 239–240  
 single-equation estimation with multiple regression, 241–247  
 power analysis, 291–292
- Recursive path models  
 assumptions of, 137  
 detailed example (*see* Recursive path model of illness)  
 directed acyclic graphs and, 165  
 features, 135  
 identification, 149–150, 159  
 single-equation estimator, 233–234
- Reduced form, 365
- Reduced-form  $R^2$ , 365
- Redundancy test, 314
- Reference group method, 404
- Reference variable, 194, 199
- Reflective indicators, 196, 352. *See also* Effect indicators
- Reflective measurement, 196–197
- Reflective measurement models, 352, 353f, 455–456
- Regions of significance, 428
- Regression analysis. *See also* individual types of regression  
 implications of the SCM for, 170–173  
 left-out variables error, 35–36  
 multiple (*see* Multiple regression)  
 observed versus estimated correlations, 41–44  
 partial and part correlation, 39–41  
 predictor selection and entry, 37–39  
 SEM and, 8, 21, 47  
 suppression, 36–37
- Regression coefficients  
 analyzing means and, 369–370  
 assumptions regarding, 33  
 SEM and, 8  
 standardized, 28, 29  
 standardized partial, 30–32  
 unstandardized, 25, 26–28, 29  
 unstandardized partial, 30–31
- Regression models, slopes-and-intercepts-as-outcomes, 445–447
- Regression rule, 27
- Regression substitution, 86–87
- Reification, 300–301
- Reject–support test, 265
- Relative fit indexes, 266. *See also* Incremental fit indexes
- Relative Noncentrality Index, 276
- Relative variances, 81–82
- Reliability. *See also* Score reliability  
 importance of reporting, 90
- Reliability coefficients, 90–91
- Reliability induction, 90
- Replacing rules. *See* Lee–Hershberger replacing rules
- Replicability fallacy, 56
- Resampling, 60
- Residual centering, 430, 443–444
- Residualized product term, 430

- Residuals  
 assumptions regarding in regression analysis, 34  
 best practices in reporting on, 464  
 correlation residuals, 252–253, 254 $t$ , 255 $f$   
 disturbances are distinct from, 131  
 heteroscedasticity and homoscedasticity, 80–81  
 normalized residuals, 252–253  
 respecification of CFA models and, 310–312  
 SEM and, 13  
 standardized residuals, 252, 253, 254 $t$ , 255 $f$   
 tips for inspecting in fit testing, 278
- Resources, for SEM, 452–454
- Respecification  
 best practices, 463  
 of CFA models, 309–312  
 of hierarchical models, empirical versus theoretical, 283–284  
 of nonrecursive models, 155–157  
 in SEM, 118 $f$ , 120  
 specification searches, 284–285
- Restricted measurement models, 191, 192 $f$ , 193–195.  
*See also* Confirmatory factor analysis
- Results  
 best practices in interpretation, 465–466  
 best practices in tabulation, 464–465  
 reporting, 120–121
- Reticular action model (RAM), 106
- Reverse coding, 197
- Reversed indicator rule, 317–318
- Reverse scoring, 197
- Ridge adjustment, 70–71
- RMSEA. *See* Steiger–Lind Root Mean Square Error of Approximation
- Robust diagonally weighted least squares (RDWLS), 328
- Robust maximum likelihood (MLR) estimation, 238–239, 272
- Robust standard errors, 238
- Robust weighted least squares (WLS) estimation, 258, 323–330
- Root mean square residual (RMR), 277
- Rotation indeterminacy, 192
- Rotations, of EFA models, 192–193
- R programming language  
 causal mediation analysis, 437  
 dagR package, 112–113  
 overview and description, 107–109  
 semPLS package, 361  
 semTools package and power analysis, 109, 290  
 WLS estimation, 326–327
- Sample median, 72–73
- Samples  
 archival, 459  
 best practices, 458–461
- Sample size  
 best practices, 459  
 minimum, estimation in power analysis, 290, 291–292  
 reporting on, 467  
 requirements for SEM, 14–16  
 unique, impact on the model chi-square, 271
- Sampling distribution, 49–51
- Sampling error, 49
- Sampling weights, 102
- SAS/STAT program  
 causal mediation analysis, 437  
 csmPower macro, 290  
 incomplete data procedures, 88  
 overview and description, 9, 102 $t$ , 109  
 power analysis, 290
- Satorra–Bentler adjusted chi-square, 272–273
- Satorra–Bentler scaled chi-square, 272, 276, 282, 283
- SBDIFF .EXE program, 282
- Scalar invariance. *See* Strong invariance
- Scaled chi-square difference test, 282, 283
- Scale free, 238
- Scale invariance, 238
- Scaling  
 of disturbances, 148  
 of factors in CFA models, 198–200
- Scaling constant, 130, 148
- Scaling correction factor, 272
- SCM. *See* Structural causal models
- Score reliability  
 discussion of, 90–93  
 with interactive effects of observed variables, 430  
 selection of measures and, 127
- Score reliability coefficients, 312
- Scores  
 product terms, 424  
 reporting the psychometric characteristics of, 90
- Score validity, 93–94
- Second-order CFA models, 319, 320, 321
- Second-order factors, 319
- Second-order partial correlation, 39
- Second-stage moderation, 433 $f$ , 434
- Seed, 62
- Seemingly uncorrelated regressions, 108
- sem command, 102 $t$ , 110, 236, 341–345
- Semipartial correlation, 39
- SEMNET, 9
- sem package, 102 $t$ , 107
- semPlot package, 109
- semPLS package, 361
- semTools package, 109, 290
- Sensitivity analysis, 85
- SEPATH (Structural Equation Modeling and Path Analysis) module  
 Bayes Information Criterion example, 287–289  
 constrained estimation, 254  
 equivalent CFA models, 317  
 overview and description, 102 $t$ , 110–111
- Sequential entry, of predictors, 38
- Shape latent growth factor  
 in basic change models, 377–384  
 in prediction models, 384–387
- Shrinkage-corrected estimate of  $\rho^2$ , 33
- Significance testing  
 assumptions regarding in regression analysis, 34  
 based on the Steiger–Lind Root Mean Square Error of Approximation, 274–275  
 cognitive errors in, 55–56

- critical ratios, 51–52
- criticisms and controversies, 8, 54–57, 62
- interval estimation as an alternative to, 57–60
- power and types of null hypotheses, 52–54
- in SEM, 8, 17
  - standard errors, 49–51
- Simple indicators, 195
- Simple intercept, 428
- Simple linear growth, 375
- Simple regressions, 428–429
- Simple slope, 428
- Simple structure, 192
- SIMPLIS computer language, 105, 357, 411
- simsem* package, 109
- SimStat program, 61–62
- Simulated data, 459
- Simultaneous causation, 123n
- Simultaneous entry, of predictors, 38
- Simultaneous estimation methods
  - overview and description, 231–232, 235
  - recommended approach to fit evaluation, 268–269
- Sine function transformations, 78
- Single-door criterion, 179–180
- Single-equation estimation methods
  - defined, 231
  - drawbacks of, 231
  - with multiple regression
    - description, 233
    - recursive path model of illness example, 241–247
  - simultaneous methods and, 235
  - two-stage least squares, 234
- Single-factor CFA models
  - analysis of the Kaufman Assessment Battery for Children, 304–306
  - of depression, with ordinal indicators
    - measurement invariance testing, 414–420
    - robust WLS estimation, 328–330
  - equivalent models, 317–319
  - of parent-child conflict, measurement invariance testing, 403–411
- Single-imputation methods, 85, 86
- Single-indicator measurement, 127
- Single indicators
  - best practices in specification, 454
  - in SR models, 213–217, 222–223, 224f, 349–352
- Singular data matrix, 69
- Skew, 74–75
- Skew index, 76, 77
- Slopes-and-intercepts-as-outcomes model, 445–447
- Slopes-as-outcomes model, 447
- SmartPLS program, 361
- Sobel test, 245
- Soft modeling, 360
- Spearman's rank order correlation, 42
- Spearman's rho, 42
- Specification
  - best practices, 454–457
  - of CFA models, 188–198, 300–301
  - model complexity, 127–128
  - parameter status, 128–129
  - of path models, 129–142 (*see also* Path models)
  - selection of measures, 127
- in SEM, 118t, 119
- of SR models, 212–217
- what variables to include, 126–127
- Specification error, 35, 235
- Specification searches, 284–285
- Specific variance, 190
- Split-half reliability, 91
- SPSS program
  - causal mediation analysis, 437
  - diagnosing missing data, 85
  - random hot-deck imputation, 87
  - regression analysis, 32
- Square root transformations, 78
- SRMR. *See* Standardized Root Mean Square Residual
- Stability index, 364
- Stable unit treatment value assumption, 123
- Standard CFA models
  - characteristics of, 193–195
  - identification rules, 201, 202f
  - overview, 207
- Standard deviations (SD), 25
- Standard errors
  - critical ratios and, 51–52
  - description, 49–51
  - robust, 238
  - for unstandardized estimates, 52
- Standardized mean residuals, 374
- Standardized partial regression coefficients, 30–32
- Standardized pattern coefficients, 301–302
- Standardized regression coefficients, 28, 29, 429–430
- Standardized residuals, 252, 253, 254t, 255f
- Standardized Root Mean Square Residual (SRMR), 269, 277–278
- Standardized scores, 25
- Standardized variance, 131
- Start values
  - iterative estimation and, 101, 236–237
  - suggestions for measurement models, 335
  - suggestions for structural models, 261
- Stata program
  - analysis of latent variable models, 18
  - example analysis of a fully latent SR model of job satisfaction, 341–345
- MEDIATION module for causal mediation analysis, 437
  - overview and description, 9, 102t, 110
  - sem command, 236
- Statgraphics Centurion, 44, 45, 46
- Stationarity assumption, 137
- STATISTICA Advanced program
  - analysis of the Kaufman Assessment Battery for Children example, 307
  - power analysis, 111, 290–291, 347
  - SEPATH module (*see* SEPATH module)
- Statistical beauty, 22, 466–467
- Statistical significance, 62, 465. *See also* Significance testing
  - testing
- STATISTICA program, 9, 102t, 110–111
- Statistics
  - classical and modeling schools in, 262
  - discussion of fit testing, 262–263
- Statistics reform, 54

- Steiger–Lind Root Mean Square Error of Approximation (RMSEA)  
 description, 273–276  
 measurement invariance testing and, 401  
 power analysis and, 290, 292
- Stem-and-leaf plots, 74, 75f
- Stepwise regression, 38
- Stochastic regression imputation, 87
- Strict invariance  
 description, 396, 399, 421  
 testing for, 399–400, 403, 413, 417
- Strictly confirmatory applications, 11
- Strong causal assumption, 131
- Strong invariance  
 description, 396, 398–399, 421  
 testing for, 399–400, 408, 413, 416–417
- Structural causal models (SCM)  
 basis set, 173–174, 175t  
 causal directed graphs, 174–177  
 causal inference and, 164–165  
 causal mediation and causal mediation analysis, 181–184  
 computer tools for, 112–113  
 description and summary, 18, 19–20, 184–185  
 elementary directed graphs and conditional independences, 166–169  
 graphical identification criteria, 177–180  
 graph vocabulary, 165–166  
 implications for regression analysis, 170–173  
 instrumental variables, 180–181  
 model identification, 119–120  
 specification and, 454
- Structural equation modeling (SEM)  
 Bayesian statistics and, 23  
 best practices  
   avoiding confirmation bias, 466  
   estimation, 461–463  
   identification, 457  
   interpretation of results, 465–466  
   measures, 458  
   resources, 452–454  
   respecification, 463  
   sample and data, 458–461  
   specification, 454–457  
   tabulation of results, 464–465  
   using SEM as a tool in science, 466–467  
 causal inference, 122–126  
 convergence with multilevel modeling, 447–450  
 data analyzed in, 13–14  
 definition of, 9–10  
 as a disconfirmatory procedure, 21  
 forms of input data, 64–67 (*see also Data; Input data*)  
 goals of, 14, 22  
 hierarchical linear modeling and, 375  
 inputs and outputs, 9–10  
 model diagram symbols, 121–122  
 model testing and model generation, 11  
 myths about, 20–21  
 observed variables and latent variables, 12–13  
 origins and history of, 23–24  
 other causal inference frameworks and, 18–20  
 other statistical techniques and, 17–18
- parametric bootstrapping, 62
- Pearson correlations, 43
- popularity of and problems with applying correctly, 21–22
- preparing to learn, 7–9
- probabilistic causality and, 11–12
- regression analysis and, 8, 21, 47
- sample size requirements, 14–16
- significance testing in, 17
- specification concepts, 126–129
- steps of  
   basic steps, 117–121  
   optional steps, 121  
   theory testing and, 10–11
- Structural invariance, 420
- Structural models  
 Lee–Hersberger replacing rules, 293–296  
 of observed variables (*see Path models*)  
 start value suggestions, 261
- Structural regression (SR) models. *See also Fully latent structural regression models;*  
*Nonrecursive structural models*  
 analysis, summary of, 361–362  
 analyzing formative measurement models in SEM, 352–361  
 constraint interaction in, 363  
 with continuous variables, mean structures and, 372  
 equivalent models, 348–349  
 exploratory structural equation modeling, 219–220  
 four-step modeling, 339–340  
 identification, 217–219, 225  
 interpretation of parameter estimates and problems, 340–341  
 LISREL notation for, 226–228  
 with multiple-indicators and multiple-causes factors, 318–319  
 research examples  
   fully latent model of job satisfaction factors, 220–222  
   single indicators in a nonrecursive model of organizational and occupational turnover intention, 222–223, 224t, 349–352
- specification  
   causal inference with latent variables, 212  
   model types, 213–214, 215f  
   standardized solutions of computer tools, 341  
   structural invariance, 420  
   summary, 223, 225  
   testing across multiple samples, 395  
   two-step modeling, 338–339
- Structure coefficient, 194, 302
- Sufficient set, 177
- Sum of squared deviations (SS), 25
- Suppression, 36–37
- Supriousness, 39
- Symbolic processing, 149
- SYSTAT program, 9, 102t, 111
- systemfit package, 102t, 108
- System matrix, 161–163
- Tailored tests, 94, 333
- Tau-equivalent indicators, 314

- TCALIS, 109  
 Templates ("wizards"), 98, 99  
 Temporal precedence, 123, 125–126  
 Test for orthogonality, 314  
 Test for redundancy, 314  
 Test-retest reliability, 92  
 Tetrachoric correlation, 43  
 TETRAD V program, 315  
 Theory testing, 10–11  
 Theta scaling, 327  
 Three-indicator rule, 201, 202f  
 Three-parameter IRT model, 94–95  
 Three-stage least squares (3SLS) estimation, 234  
 Threshold residuals, 328  
 Thresholds, in WLS estimation, 324–325  
 Threshold structure, 326  
 Time-invariant predictors, 390  
 Time precedence, 296–297  
 Time structured data, 375  
 Time-varying predictors, 390  
 Tolerance, 71  
 Total effect moderation, 434  
 Total effects  
     causal mediation analysis, 435–437  
     defined, 134  
 Total indirect effects, 232  
 Tracing rules, 250–253  
 Trained incapacity, 54  
 Transformations, 77–81  
 Triangle inequality, 68  
 Tucker–Lewis index, 276–277  
 Two-factor CFA models  
     equivalent models, 315–317  
     Kaufman Assessment Battery for Children, 305–306, 307–309, 310t, 311t  
 Two-indicator rule, 201, 202f  
 Two-level regression analysis, 445, 446f  
 Two-parameter IRT model, 94, 95f  
 2+ Emitted paths rule, 354  
 Two-stage least squares (2SLS) estimation, 234, 235, 442–443  
 Two-step identification rule, 217–219  
 Two-step modeling, of fully latent SR models, 338–339, 341–347  
 Type I errors, 56, 57  
 Type II errors, 56–57  
 Unconstrained approach to estimation, 443  
 Undecidable identification problem, 149  
 Underidentification, empirical, 157–158  
 Underidentified (underdetermined) models, 146, 147  
 Undirected path, 166  
 Unidimensional measurement, 195  
 Unique variance, 190, 271  
 Unit loading identification (ULI) constraint, 148, 199  
 Unit variance identification (UVI) constraint, 199  
 Univariate outliers, 72–73  
 Unknown weights composite, 355  
 Unrestricted measurement models, 191–193. *See also*  
     Exploratory factor analysis  
 Unstandardized bivariate regression, 25–28  
 Unstandardized partial regression coefficients, 30–31  
 Unstandardized regression coefficients, 25, 26–28, 29  
 Unstandardized residual path coefficients, 130  
 Unweighted least squares (ULS) estimation, 256, 330–331  
 Validity. *See* Score validity  
 Valid research hypothesis fallacy, 56  
 Valid tracings, 250–251  
 Vanishing tetrads, 315  
 Variables. *See also* Continuous variables; Endogenous variables; Exogenous variables; Instrumental variables; Latent response variables; Latent variables; Outcome variables  
     auxiliary, 84  
     censored, 43  
     count, 79  
     d-connected, 171  
     demographic, 217  
     extreme collinearity, 71–72  
     in graph theory, 165–166  
     instrumental, 150–152, 180–181  
     marker, 194  
     reference, 194, 199  
     specification, 126–127  
     transformations, 77–81  
 Variance inflation factor (VIF), 71  
 Variances  
     corrected proportions of explained variance for nonrecursive structural models, 365–366  
     estimation in the maximum likelihood method, 236  
     in factor analysis, 190  
     relative, 81–82  
     specific, 190  
     standardized variance for a continuos endogenous variable in a path model, 131  
     unique, 190, 271  
 Vertices, 165  
 Wald W statistic, 284  
 Weak causal assumption, 131  
 Weak invariance  
     description, 396, 397, 421  
     testing for, 399–400, 406, 408, 409, 413, 416  
 Weighted least squares (WLS) estimation, 256, 258, 259. *See also* Mean- and variance-adjusted weighted least squares; Robust weighted least squares estimation  
 Weight matrix, 256  
 Welch–James test, 410, 423  
 Welch–Satterthwaite equation, 423  
 Wherry equation, 33  
 Within-groups completely standardized solutions, 395  
 Within-groups standardized solutions, 395  
 "Wizards" (templates), 98, 99  
 WLS. *See* Weighted least squares  
 Wright's tracing rules, 250–253  
 Zero-order associations, 37–38  
 z Scores, 72

## About the Author

**Rex B. Kline, PhD**, is Professor of Psychology at Concordia University in Montréal. Since earning a doctorate in clinical psychology, his areas of research and writing have included the psychometric evaluation of cognitive abilities, behavioral and scholastic assessment of children, structural equation modeling, training of researchers, statistics reform in the behavioral sciences, and usability engineering in computer science. Dr. Kline has published seven books and 10 chapters in these areas. His website is <http://tinyurl.com/rexkline>.