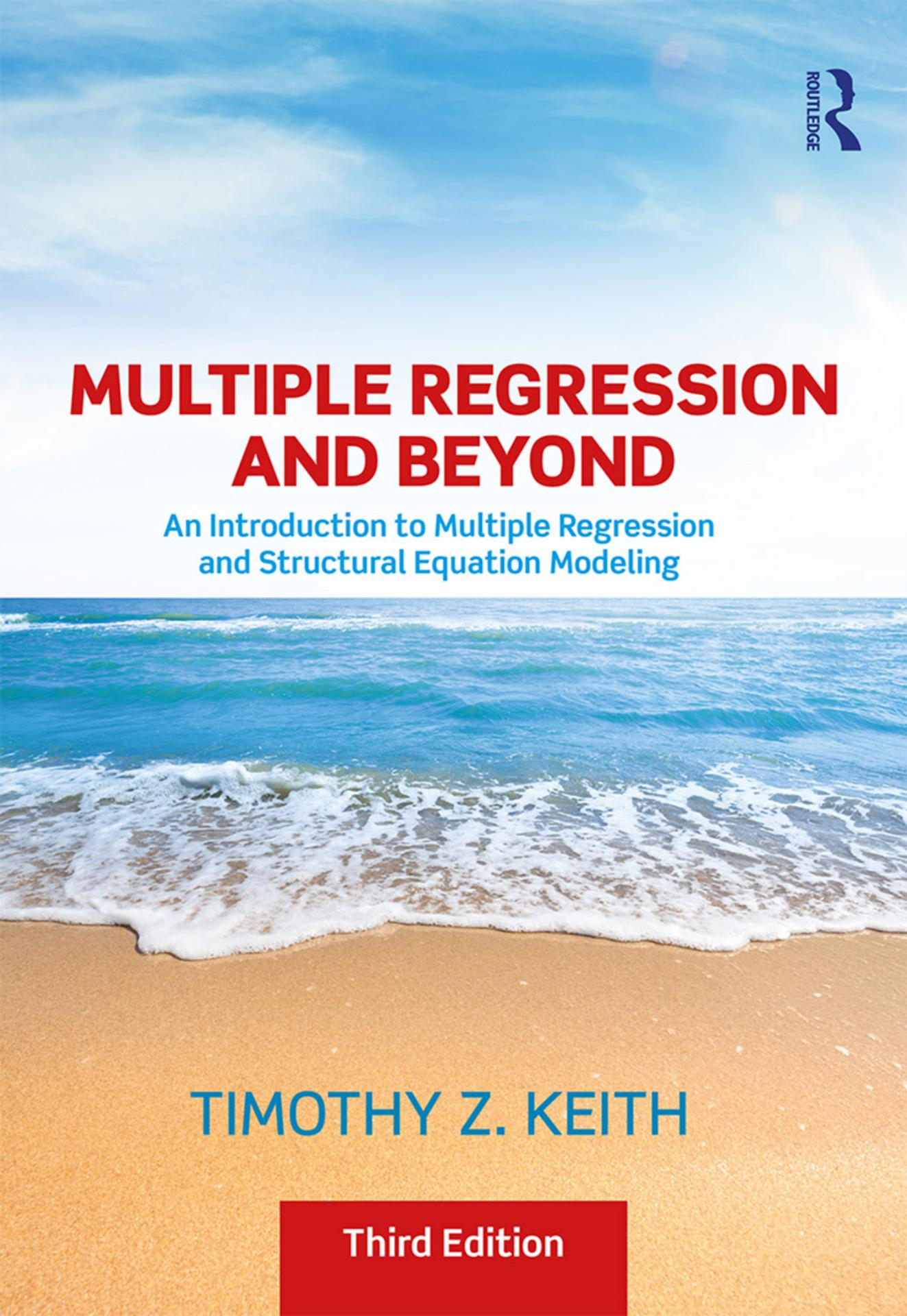




MULTIPLE REGRESSION AND BEYOND

An Introduction to Multiple Regression
and Structural Equation Modeling

A photograph of a sandy beach meeting the ocean waves. The water is a vibrant blue, and the sand is a light tan color. The waves are crashing onto the shore, creating white foam. The sky above is a clear, pale blue.

TIMOTHY Z. KEITH

Third Edition

Multiple Regression and Beyond

Multiple Regression and Beyond offers a conceptually-oriented introduction to multiple regression (MR) analysis and structural equation modeling (SEM), along with analyses that flow naturally from those methods. By focusing on the concepts and purposes of MR and related methods, rather than the derivation and calculation of formulae, this book introduces material to students more clearly, and in a less threatening way. In addition to illuminating content necessary for coursework, the accessibility of this approach means students are more likely to be able to conduct research using MR or SEM—and more likely to use the methods wisely.

This book:

- Covers both MR and SEM, while explaining their relevance to one another
- Includes path analysis, confirmatory factor analysis, and latent growth modeling
- Makes extensive use of real-world research examples in the chapters and in the end-of-chapter exercises
- Extensive use of figures and tables providing examples and illustrating key concepts and techniques

New to this edition:

- New chapter on mediation, moderation, and common cause
- New chapter on the analysis of interactions with latent variables and multilevel SEM
- Expanded coverage of advanced SEM techniques in chapters 18 through 22
- International case studies and examples
- Updated instructor and student online resources

Timothy Z. Keith is Professor of Educational Psychology at the University of Texas, Austin. His research is focused on the nature and measurement of intelligence, including the validity of tests of intelligence and the theories from which they are drawn. His research has been recognized with awards from the three major journals in school psychology, and he was awarded the senior scientist distinction by the School Psychology division of APA.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Multiple Regression and Beyond

*An Introduction to
Multiple Regression
and Structural
Equation Modeling*

Third Edition

Timothy Z. Keith

Third edition published 2019
by Routledge
52 Vanderbilt Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

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

© 2019 Taylor & Francis

The right of Timothy Z. Keith to be identified as author of this work has been asserted by him in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

First edition published by Pearson Education 2006
Second edition published by Routledge 2014

Library of Congress Cataloging-in-Publication Data
Names: Keith, Timothy Z., author.

Title: Multiple regression and beyond : an introduction to multiple regression and structural equation modeling / Timothy Z. Keith.

Description: Third Edition. | New York : Routledge, 2019. | Revised edition of the author's Multiple regression and beyond, 2015.

Identifiers: LCCN 2018041116 | ISBN 9781138061422 (hardback) | ISBN 9781138061446 (pbk.) | ISBN 9781315162348 (ebook)

Subjects: LCSH: Regression analysis.

Classification: LCC HA31.3 .K45 2019 | DDC 519.5/36—dc23

LC record available at <https://lccn.loc.gov/2018041116>

ISBN: 978-1-138-06142-2 (hbk)

ISBN: 978-1-138-06144-6 (pbk)

ISBN: 978-1-315-16234-8 (ebk)

Typeset in Minion
by Apex CoVantage, LLC

Visit the companion website: www.tzkeith.com

Contents

Preface	vii
Acknowledgments	xiii
Part I Multiple Regression	1
1 Simple Bivariate Regression	3
2 Multiple Regression: Introduction	26
3 Multiple Regression: More Depth	44
4 Three and More Independent Variables and Related Issues	57
5 Three Types of Multiple Regression	77
6 Analysis of Categorical Variables	108
7 Regression With Categorical and Continuous Variables	129
8 Testing for Interactions and Curves With Continuous Variables	161
9 Mediation, Moderation, and Common Cause	177
10 Multiple Regression: Summary, Assumptions, Diagnostics, Power, and Problems	195
11 Related Methods: Logistic Regression and Multilevel Modeling	226

Part II	Beyond Multiple Regression: Structural Equation Modeling	255
12	Path Modeling: Structural Equation Modeling With Measured Variables	257
13	Path Analysis: Assumption and Dangers	281
14	Analyzing Path Models Using SEM Programs	296
15	Error: The Scourge of Research	334
16	Confirmatory Factor Analysis I	348
17	Putting It All Together: Introduction to Latent Variable SEM	389
18	Latent Variable Models II: Multigroup Models, Panel Models, Dangers and Assumptions	409
19	Latent Means in SEM	444
20	Confirmatory Factor Analysis II: Invariance and Latent Means	475
21	Latent Growth Models	513
22	Latent Variable Interactions and Multilevel Modelling in SEM	534
23	Summary: Path Analysis, CFA, SEM, Mean Structures, and Latent Growth Models	561
Appendices		
	Appendix A: Data Files	585
	Appendix B: Review of Basic Statistics Concepts	587
	Appendix C: Partial and Semipartial Correlation	605
	Appendix D: Symbols Used in This Book	613
	Appendix E: Useful Formulae	615
	References	617
	Author Index	629
	Subject Index	633

Preface

Multiple Regression and Beyond is designed to provide a conceptually oriented introduction to multiple regression along with more complex methods that flow naturally from multiple regression: path analysis, confirmatory factor analysis, and structural equation modeling. Multiple regression (MR) and related methods have become indispensable tools for modern social science researchers. MR closely implements the general linear model and thus subsumes methods, such as analysis of variance (ANOVA), that have traditionally been more commonplace in psychological and educational research. Regression is especially appropriate for the analysis of nonexperimental research, and with the use of dummy variables and modern computer packages, it is often more appropriate or easier to use MR to analyze the results of complex quasi-experimental or even experimental research. Extensions of multiple regression—particularly structural equation modeling (SEM)—partially obviate threats due to the unreliability of the variables used in research and allow the modeling of complex relations among variables. A quick perusal of the full range of social science journals demonstrates the wide applicability of the methods.

Despite its importance, MR-based analyses are too often poorly conducted and poorly reported. I believe one reason for this incongruity is inconsistency between how material is presented and how most students best learn.

Anyone who teaches (or has ever taken) courses in statistics and research methodology knows that many students, even those who may become gifted researchers, do not always gain conceptual understanding only through numerical presentation. Although many who teach statistics understand the processes underlying a sequence of formulas and gain conceptual understanding through these formulas, many students do not. Instead, such students often need a thorough conceptual explanation to gain such understanding, after which a numerical presentation may make more sense. Unfortunately, many multiple regression textbooks assume that students will understand multiple regression best by learning matrix algebra, wading through formulas, and focusing on details.

At the same time, methods such as structural equation modeling (SEM) and confirmatory factor analysis (CFA) are easily taught as extensions of multiple regression. If structured properly, multiple regression flows naturally into these more complex topics, with nearly complete carry-over of concepts. Path models (simple SEMs) illustrate and help deal with some of the problems of MR, CFA does the same for path analysis, and latent variable SEM combines all the previous topics into a powerful, flexible methodology.

I have taught courses including these topics at four universities (the University of Iowa, Virginia Polytechnic Institute & State University, Alfred University, and the University of Texas). These courses included students and faculty in architecture, engineering, educational psychology, educational research and statistics, kinesiology, management, political science, psychology, social work, and sociology, among others. This experience leads me to believe that it is possible to teach these methods by focusing on the concepts and purposes of MR and related methods, rather than the derivation and calculation of formulas. Non-quantitatively-oriented students generally find such an approach clearer, more conceptual, and less threatening than other approaches. As a result of this conceptual approach, students become interested in conducting research using MR, CFA, or SEM and are more likely to use the methods wisely.

THE ORIENTATION OF THIS BOOK

My overriding bias in this book is that these complex methods can be presented and learned in a conceptual, yet rigorous, manner. I recognize that not all topics are covered in the depth or detail presented in other texts, but I will direct you to other sources for topics for which you may want additional detail. My style is also fairly informal; I've written this book as if I were teaching a class.

Data

I also believe that one learns these methods best by doing, and the more interesting and relevant that “doing,” the better. For this reason, there are numerous example analyses throughout this book that I encourage you to reproduce as you read. To make this task easier, the Web site that accompanies the book (www.tzkeith.com) includes the data in a form that can be used in common statistical analysis programs. Many of the examples are taken from actual research in the social sciences, and I've tried to sample from research from a variety of areas. In most cases simulated data are provided that mimic the actual data used in the research. You can reproduce the analyses of the original researchers and, perhaps, improve on them.

And the data feast doesn't end there! The Web site also includes data from a major federal data set: 1000 cases from the National Education Longitudinal Study (NELS) from the National Center for Education Statistics. NELS was a nationally representative sample of 8th-grade students first surveyed in 1988 and resurveyed in 10th and 12th grades and then twice after leaving high school. The students' parents, teachers, and school administrators were also surveyed. The Web site includes student and parent data from the base year (8th grade) and student data from the first follow-up (10th grade). Don't be led astray by the word Education in NELS; the students were asked an incredible variety of questions, from drug use to psychological well-being to plans for the future. Anyone with an interest in youth will find something interesting in these data. Appendix A includes more information about the data at www.tzkeith.com.

Computer Analysis

Finally, I firmly believe that any book on statistics or research methods should be closely related to statistical analysis software. Why plug numbers into formulas and churn out the answers on a calculator—when a statistical program can do the calculations more quickly and accurately with, for most people, no loss of understanding? Freed from the need for hand calculations, you can then concentrate on asking and answering important

research questions, rather than on the intricacies of calculating statistics. This bias toward computer calculations is especially important for the methods covered in this book, which quickly become unmanageable by hand. Use a statistical analysis program as you read this book; do the examples with me and the problems at the end of the chapters, using that program.

Which program? I use SPSS as my general statistical analysis program, and you can get the program for a reasonable price as a student in a university (approximately \$100 per year for the “Grad Pack” as this is written). But you need not use SPSS; any of the common packages will do (e.g., SAS or Stata or R). The output in the text has a generic look to it, which should be easily translatable to any major statistical package output. In addition, the website (www.tzkeith.com) includes sample multiple regression and SEM output from various statistical packages.

For the second half of the book, you will need access to a structural equation modeling program. Fortunately, student versions of many such programs are available online. Student pricing for the program used extensively in this book, Amos, is available, at this writing, for approximately \$50 per year as an SPSS for Windows add-on. Although programs (and pricing) change, one current limitation of Amos is that there is no Mac OS version of Amos. If you want to use Amos, you need to be able to run Windows. Amos is, in my opinion, the easiest SEM program to use (and it produces really nifty pictures). The other SEM program that I will frequently reference is Mplus. We’ll talk more about SEM in Part 2 of this book. The website for this text has many examples of SEM input and output using Amos and Mplus.

Overview of the Book

This book is divided into two parts. Part 1 focuses on multiple regression analysis. We begin by focusing on simple, bivariate regression and then expand that focus into multiple regression with two, three, and four independent variables. We will concentrate on the analysis and interpretation of multiple regression as a way of answering interesting and important research questions. Along the way, we will also deal with the analytic details of multiple regression so that you understand what is going on when we do a multiple regression analysis. We will focus on three different types, or flavors, of multiple regression that you will encounter in the research literature, their strengths and weaknesses, and their proper interpretation. Our next step will be to add categorical independent variables to our multiple regression analyses, at which point the relation of multiple regression and ANOVA will become clearer. We will learn how to test for interactions and curves in the regression line and to apply these methods to interesting research questions.

The penultimate chapter for Part 1 is a review chapter that summarizes and integrates what we have learned about multiple regression. Besides serving as a review for those who have gone through Part 1, it also serves as a useful introduction for those who are interested primarily in the material in Part 2. In addition, this chapter introduces several important topics not covered completely in previous chapters. The final chapter in Part 1 presents two related methods, logistic regression and multilevel modeling, in a conceptual fashion using what we have learned about multiple regression.

Part 2 focuses on structural equation modeling—the “Beyond” portion of the book’s title. We begin by discussing path analysis, or structural equation modeling with measured variables. Simple path analyses are easily estimated via multiple regression analysis, and many of our questions about the proper use and interpretation of multiple regression will be answered with this heuristic aid. We will deal in some depth with the problem of valid versus invalid inferences of causality in these chapters. The problem of error (“the scourge of

research") serves as our jumping off place for the transition from path analysis to methods that incorporate latent variables (confirmatory factor analysis and latent variable structural equation modeling). Confirmatory factor analysis (CFA) approaches more closely the constructs of primary interest in our research by separating measurement error from variation due to these constructs. Latent variable structural equation modeling (SEM) incorporates the advantages of path analysis with those of confirmatory factor analysis into a powerful and flexible analytic system that partially obviates many of the problems we discuss as the book progresses. As we progress to more advanced SEM topics we will learn how to test for interactions in SEM models, and for differences in means of latent constructs. SEM allows powerful analysis of change over time via methods such as latent growth models. Even when we discuss fairly sophisticated SEMs, we reiterate one more time the possible dangers of nonexperimental research in general and SEM in particular.

CHANGES TO THE SECOND EDITION

If you are coming to the second edition from the first, thank you! There are changes throughout the book, including quite a few new topics, especially in Part 2. Briefly, these include:

Changes to Part 1

All chapters have been updated to add, I hope, additional clarity. In some chapters the examples used to illustrate particular points have been replaced with new ones. In most chapters I have added additional exercises and have tried to sample these from a variety of disciplines.

New to Part 1 is a chapter on Logistic Regression and Multilevel Modeling (Chapter 11). This brief introduction is not intended as an introduction to these important topics but instead as a bridge to assist students who are interested in pursuing these topics in more depth in subsequent coursework. When I teach MR classes I consistently get questions about these methods, how to think about them, and where to go for more information. The chapter focuses on using what students have learned so far in MR, especially categorical variables and interactions, to bridge the gap between a MR class and ones that focus in more detail on LR and MLM.

Changes to Part 2

What is considered introductory material in SEM has expanded a great deal since I wrote the first edition to Multiple Regression and Beyond, and thus new chapters have been added to address these additional topics.

A chapter on Latent Means in SEM (Chapter 19) introduces the topic of mean structures in SEM, which is required for understanding the next three chapters and which has increasingly become a part of introductory classes in SEM. The chapter uses a research example to illustrate two methods of incorporating mean structures in SEM: MIMIC-type models and multi-group mean and covariance structure models.

A second chapter on Confirmatory Factor Analysis has been added (Chapter 20). Now that latent means have been introduced, this chapter revisits CFA, with the addition of latent means. The topic of invariance testing across groups, hinted at in previous chapters, is covered in more depth.

Chapter 21 focuses on Latent Growth Models. Longitudinal models and data have been covered in several places in the text. Here latent growth models are introduced as a method of more directly studying the process of change.

Along with these additions, Chapter 18 (Latent Variable Models II: Multigroup Models, Panel Models, Dangers and Assumptions) and the final SEM summary chapter (Chapter 23) have been extensively modified as well.

Changes to the Appendices

Appendix A, which focused on the data sets used for the text, is considerably shortened, with the majority of the material transferred to the web (www.tzkeith.com). Likewise, the information previously contained in appendices illustrating output from statistics programs and SEM programs has been transferred to the web, so that I can update it regularly. There are still appendices focused on a review of basic statistics (Appendix B) and on understanding partial and semipartial correlations (Appendix C). The tables showing the symbols used in the book and useful formulae are now included in appendices as well.

ACKNOWLEDGMENTS, SECOND EDITION

This project could not have been completed without the help of many people. I was amazed by the number of people who wrote to me about the first edition with questions, compliments, and suggestions (and corrections!). Thank you! I am very grateful to the students who have taken my classes on these topics over the years. Your questions and comments have helped me understand what aspects of the previous edition of the book worked well and which needed improvement or additional explanation. I owe a huge debt to the former and current students who “test drove” the new chapters in various forms.

I am grateful to the colleagues and students who graciously read and commented on various new sections of the book: Jacqueline Caemmerer, Craig Enders, Larry Greil, and Keenan Pituch. I am especially grateful to Matthew Reynolds, who read and commented on every one of the new chapters and who is a wonderful source of new ideas for how to explain difficult concepts.

I thank my hard-working editor, Rebecca Novack, and her assistants at Routledge for all of their assistance. Rebecca’s zest and humor, and her commitment to this project, were key to its success. None of these individuals is responsible for any remaining deficiencies of the book, however.

Finally, a special thank you to my wife and to my sons and their families. Davis, Scotty, and Willie, you are a constant source of joy and a great source of research ideas! Trisia provided advice, more loving encouragement than I deserve, and the occasional nudge, all as needed. Thank you, my love, I really could not have done this without you!

CHANGES TO THE THIRD EDITION

In addition to the normal updating of material, references, and examples that takes place with any revision, I have added two new chapters for the third edition. In Part 1, Chapter 9 expands on the topics of mediation, moderation, and common cause. The topics of mediation and moderation were introduced in previous chapters, and I hope these timely topics are well-consolidated here. Also included are examples of various methods of testing for mediation in MR. I believe the topic of common causes is extremely important, and discuss it throughout the text. This chapter consolidates some of that discussion and differentiates this topic from the sometimes-confused topics of mediation and moderation.

In Part 2, Chapter 22 covers two fairly advanced SEM techniques: the analysis of interactions for continuous latent variables, and the analysis of multilevel structural equation models. Both topics are briefly introduced and illustrated with an example. Of course, the other chapters are renumbered, with new material, and things moved here and there to, I hope, improve the flow.

Acknowledgments

I am very grateful to everyone who used the first and second editions of this textbook to learn about multiple regression and structural equation modelling. Thank you! And thank you to those who took time to write to me about aspects of the book that you found valuable and for suggestions for this third edition and the website. Thank you also to those who wrote to alert me to errors and needed corrections!

I am once again very grateful to students who have taken my courses on these topics or read this book and then asked questions. Your questions and comments are incredibly helpful in helping me hone the material in this text. I thank the students in my research lab for coming up with new problems to study and for their zeal in applying the methods discussed here to those problems. My thanks to my former students and now colleagues Matt Reynolds and Jackie Caemmerer for your helpful suggestions on both the second edition and on the new material presented in this edition and for your continued discussion of MR, CFA, and SEM methodology.

I was once again blessed with a dedicated editor, Hannah Shakespeare and her attentive assistant, Matt Bickerton.

Finally, once again I cannot thank my wife Trisia enough. Your encouragement, support, excitement, and advice make all the difference in the world. Thank you, my love!



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Part I

Multiple Regression



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

1

Simple Bivariate Regression

Simple Bivariate Regression	4
<i>Example: Homework and Math Achievement</i>	4
Regression in Perspective	15
<i>Relation of Regression to Other Statistical Methods</i>	15
<i>Explaining Variance</i>	17
<i>Advantages of Multiple Regression</i>	18
Other Issues	19
<i>Prediction Versus Explanation</i>	19
<i>Causality</i>	19
Review of Some Basics	20
<i>Variance and Standard Deviation</i>	20
<i>Correlation and Covariance</i>	20
Working With Extant Data Sets	21
Summary	23
Exercises	24
<i>Notes</i>	24

This book is designed to provide a conceptually-oriented introduction to multiple regression along with more complex methods that flow naturally from multiple regression: path analysis, confirmatory factor analysis, and structural equation modeling. In this introductory chapter, we begin with a discussion and example of simple, or bivariate, regression. For many readers, this will be a review, but, even then, the example and computer output should provide a transition to subsequent chapters and to multiple regression. The chapter also reviews several other related concepts, and introduces several issues (prediction and explanation, causality) that we will return to repeatedly in this book. Finally, the chapter relates regression to other approaches with which you may be more familiar, such as analysis of variance (ANOVA). I will demonstrate that ANOVA and regression are fundamentally the same process and that, in fact, regression subsumes ANOVA.

As I suggested in the Preface, we start this journey by jumping right into an example and explaining it as we go. In this introduction, I have assumed that you are familiar with the topics of correlation and statistical significance testing and that you have some familiarity with statistical procedures such as the *t* test for comparing means and analysis of variance. If these concepts are not familiar to you a quick review is provided in Appendix B. This appendix

reviews basic statistics, distributions, standard errors and confidence intervals, correlations, *t* tests, and ANOVA.

SIMPLE BIVARIATE REGRESSION

Let's start our adventure into the wonderful world of multiple regression with a review of simple, or bivariate, regression; that is, regression with only one influence (independent variable) and one outcome (dependent variable).¹ Pretend that you are the parent of an adolescent. As a parent, you are interested in the influences on adolescents' school performance: what's important and what's not? Homework is of particular interest because you see your daughter Lisa struggle with it nightly and hear her complain about it daily. A quick search of the Internet reveals conflicting evidence. You may find books (Kohn, 2006) and articles (Wallis, 2006) critical of homework and homework policies. On the other hand, you may find links to research suggesting homework improves learning and achievement (Cooper, Robinson, & Patall, 2006). So you wonder if homework is just busywork or is it a worthwhile learning experience?

Example: Homework and Math Achievement

The Data

Fortunately for you, your good friend is an 8th-grade math teacher and you are a researcher; you have the means, motive, and opportunity to find the answer to your question. Without going into the levels of permission you'd need to collect such data, pretend that you devise a quick survey that you give to all 8th-graders. The key question on this survey is:

Think about your math homework over the last month. Approximately how much time did you spend, per week, doing your math homework? Approximately ____ (fill in the blank) hours per week.

A month later, standardized achievement tests are administered; when they are available, you record the math achievement test score for each student. You now have a report of average amount of time spent on math homework and math achievement test scores for 100 8th-graders.

A portion of the data is shown in Figure 1.1. The complete data are on the website that accompanies this book, www.tzkeith.com, under Chapter 1, in several formats: as an SPSS System file (homework & ach.sav), as a Microsoft Excel file (homework & ach.xls), and as an ASCII, or plain text, file (homework & ach.txt). The values for time spent on Math Homework are in hours, ranging from zero for those who do no math homework to some upper value limited by the number of free hours in a week. The Math Achievement test scores have a national mean of 50 and a standard deviation of 10 (these are known as *T* scores, which have nothing to do with *t* tests).²

Let's turn to the analysis. Fortunately, you have good data analytic habits: you check basic descriptive data prior to doing the main regression analysis. Here's my rule: *Always, always, always, always, always, always check your data* prior to conducting analyses! The frequencies and descriptive statistics for the Math Homework variable are shown in Figure 1.2. Reported Math Homework ranged from no time, or zero hours, reported by 19 students, to 10 hours per week. The range of values looks reasonable, with no excessively high or impossible values. For example, if someone had reported spending 40 hours per week on Math Homework, you might be a little suspicious and would check your original data to make sure you entered the data correctly (e.g., you may have entered a "4" as a "40"; see Chapter 10 for more information about spotting data problems). You might be a little surprised that the average amount of time spent on Math Homework per week is only 2.2 hours, but this value is certainly plausible. (As noted in the Preface, the regression and other results shown are portions of

Math Homework	Math Achievement
2	54
0	53
4	53
0	56
2	59
0	30
1	49
0	54
3	37
0	49
4	55
7	50
3	45
1	44
1	60
0	36
3	53
0	22
1	56

(Data Continue.....)

Figure 1.1 Portion of the Math Homework and Achievement data. The complete data are on the website under Chapter 1.

MATHHOME Time Spent on Math Homework per Week

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	19	19.0	19.0
	1.00	19	19.0	38.0
	2.00	25	25.0	63.0
	3.00	16	16.0	79.0
	4.00	11	11.0	90.0
	5.00	6	6.0	96.0
	6.00	2	2.0	98.0
	7.00	1	1.0	99.0
	10.00	1	1.0	100.0
Total	100	100.0	100.0	

Statistics

MATHHOME Time Spent on Math Homework per Week

N	Valid	100
	Missing	0
Mean		2.2000
Median		2.0000
Mode		2.00
Std. Deviation		1.8146
Variance		3.2929
Minimum		.00
Maximum		10.00
Sum		220.00

Figure 1.2 Frequencies and descriptive statistics for Math Homework.

6 • MULTIPLE REGRESSION

an SPSS printout, but the information displayed is easily generalizable to that produced by other statistical programs.)

Next, turn to the descriptive statistics for the Math Achievement test (Figure 1.3). Again, given that the national mean for this test is 50, the 8th-grade school mean of 51.41 is reasonable, as is the range of scores from 22 to 75. In contrast, if the descriptive statistics had shown a high of, for example, 90 (four standard deviations above the mean), further investigation would be called for. The data appear to be in good shape.

The Regression Analysis

Next, we conduct regression: we regress Math Achievement scores on time spent on Homework (notice the structure of this statement: we regress the outcome on the influence or influences). Figure 1.4 shows the means, standard deviations, and correlation between the two variables.

Descriptive Statistics								
	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation	Variance
MATHACH Math Achievement Test Score	100	53.00	22.00	75.00	5141.00	51.4100	11.2861	127.376
Valid N (listwise)	100							

Figure 1.3 Descriptive statistics for Math Achievement test scores.

Descriptive Statistics			
	Mean	Std. Deviation	N
MATHACH Math Achievement Test Score	51.4100	11.2861	100
MATHHOME Time Spent on Math Homework per Week	2.2000	1.8146	100

		MATHACH Math Achievement Test Score	MATHHOME Time Spent on Math Homework per Week
Pearson Correlation	MATHACH Math Achievement Test Score MATHHOME Time Spent on Math Homework per Week	1.000 .320	.320 1.000
Sig. (1-tailed)	MATHACH Math Achievement Test Score MATHHOME Time Spent on Math Homework per Week	. .001	.001 .001
N	MATHACH Math Achievement Test Score MATHHOME Time Spent on Math Homework per Week	100 100	100 100

Figure 1.4 Results of the regression of Math Achievement on Math Homework: descriptive statistics and correlation coefficients.

The descriptive statistics match those presented earlier, without the detail. The correlation between the two variables is .320, not large, but certainly statistically significant ($p < .01$) with this sample of 100 students. As you read articles that use multiple regression, you may see this ordinary correlation coefficient referred to as a zero-order correlation (which distinguishes it from first-, second-, or multiple-order partial correlations, topics discussed in Appendix C).

Next, we turn to the regression itself; although we have conducted a simple regression, the computer output is in the form of multiple regression to allow a smooth transition. First, look at the model summary in Figure 1.5. It lists the R , which normally is used to designate the multiple correlation coefficient, but which, with one predictor, is the same as the simple Pearson correlation (.320).³ Next is the R^2 , which denotes the variance explained in the outcome variable by the predictor variables. Homework time explains, accounts for, or predicts .102 (proportion) or 10.2% of the variance in Math test scores. As you run this regression yourself, your output will probably show some additional statistics (e.g., the adjusted R^2); we will ignore these for the time being.

Is the regression, that is, the multiple R and R^2 , statistically significant? We know it is, because we already noted the statistical significance of the zero-order correlation, and this “multiple” regression is actually a simple regression with only one predictor. But, again, we’ll check the output for consistency with subsequent examples. Interestingly, we use an F test, as in ANOVA, to test the statistical significance of the regression equation:

$$F = \frac{ss_{\text{regression}} / df_{\text{regression}}}{ss_{\text{residual}} / df_{\text{residual}}}$$

The term $ss_{\text{regression}}$ stands for sums of squares regression and is a measure of the variation in the dependent variable that is explained by the independent variable(s); the ss_{residual} is the variance unexplained by the regression. If you are interested in knowing how to calculate these values by hand, turn to Note 4 at the end of this chapter; here, we will use the values from the statistical output in Figure 1.5.⁴ The sums of squares for the regression versus the

Model Summary

Model	R	R Square
1	.320 ^a	.102

- a. Predictors: (Constant), MATHHOME Time Spent on Math Homework per Week

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1291.231	1	1291.231	11.180	^{.001^a}
	Residual	11318.959	98	115.500		
	Total	12610.190	99			

- a. Predictors: (Constant), MATHHOME Time Spent on Math Homework per Week
 b. Dependent Variable: MATHACH Math Achievement Test Score

Figure 1.5 Results of the regression of Math Achievement on Math Homework: statistical significance of the regression.

8 • MULTIPLE REGRESSION

residual are shown in the ANOVA table. In regression, the degrees of freedom (df) for the regression are equal to the number of independent variables (k), and the df for the residual, or error, are equal to the sample size minus the number of independent variables in the equation minus 1 ($N - k - 1$); the df are also shown in the ANOVA table. We'll double-check the numbers:

$$\begin{aligned} F &= \frac{1291.231/1}{11318.959/98} \\ &= \frac{1291.231}{115.500} \\ &= 11.179 \end{aligned}$$

which is the same value shown in the table, within errors of rounding. What is the probability of obtaining a value of F as large as 11.179 if these two variables were in fact unrelated in the population? According to the table (in the column labeled "Sig."), such an occurrence would occur only 1 time in 1,000 ($p = .001$); it would seem logical that these two variables are indeed related. We can double-check this probability by referring to an F table under 1 and 98 df ; is the value 11.179 greater than the tabled value? Instead, however, I suggest that you use a computer program to calculate these probabilities. Excel, for example, will find the probability for values of all the distributions discussed in this text. Simply put the calculated value of F (11.179) in one cell, the degrees of freedom for the regression (1) in the next, and the df for the residual in the next (98). Go to the next cell, then click on Insert, Function, and select the category of Statistical and scroll down until you find FDIST.RT, for F distribution (the older FDIST function in Excel also provides this information).

Click on it and point to the cells containing the required information. Alternatively, you could go directly to Function and FDIST.RT and simply type in these numbers, as was done in Figure 1.6. Excel returns a value of .001172809, or .001, as shown in the Figure. Although I present this method of determining probabilities as a way of double-checking the computer output at this point, at times your computer program will not display the probabilities you are interested in, and this method will be useful.

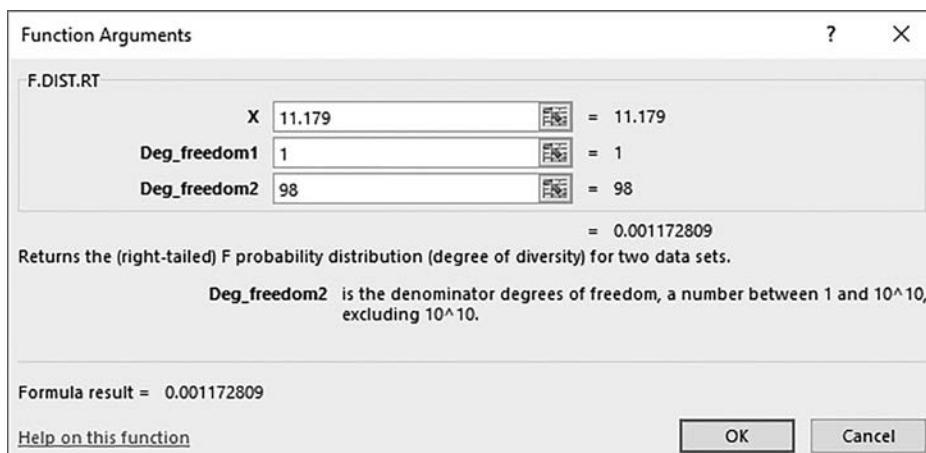


Figure 1.6 Using Excel to calculate probability: statistical significance of an F (1,98) of 11.179.

There is another formula you can use to calculate F , an extension of which will come in handy later:

$$F = \frac{R^2 / k}{(1 - R^2) / (N - k - 1)}$$

This formula compares the proportion of variance explained by the regression (R^2) with the proportion of variance left unexplained by the regression ($1 - R^2$). This formula may seem quite different from the one presented previously until you remember that (1) k is equal to the df for the regression, and $N - k - 1$ is equal to the df for the residual, and (2) the sums of squares from the previous formula are also estimates of variance. Try this formula to make sure you get the same results (within rounding error).

I noted that the $ss_{regression}$ is a measure of the variance explained in the dependent variable by the independent variables, and also that R^2 denotes the variance explained. Given these descriptions, you may expect that the two concepts should be related. They are, and we can calculate the R^2 from the $ss_{regression}$: $R^2 = \frac{ss_{regression}}{ss_{total}}$. We can put this formula into words: There is a certain amount of variance in the dependent variable (total variance), and the independent variables can explain a portion of this variance (variance due to the regression). The R^2 is a proportion of the total variance in the dependent variable that is explained by the independent variables. For the current example, the total variance in the dependent variable, Math Achievement (ss_{total}), was 12610.190 (Figure 1.5), and Math Homework explained 1291.231 of this variance. Thus,

$$\begin{aligned} R^2 &= \frac{ss_{regression}}{ss_{total}} \\ &= \frac{1291.231}{12610.190} \\ &= .102 \end{aligned}$$

and Homework explains .102 or 10.2% of the variance in Math Achievement. Obviously, R^2 can vary between 0 (no variance explained) and 1 (100% explained).

The Regression Equation

Next, let's take a look at the coefficients for the regression equation, the notable parts of which are shown in Figure 1.7. The general formula for a regression equation is $Y = a + bX + e$, which, translated into English, says that a person's score on the dependent variable (in this case, Math Achievement) is a result of a constant (a), plus a coefficient (b) times his or her value on the independent variable (Math Homework), plus error. Values for both a and b are shown in the second column of the table in Figure 1.7 (Unstandardized Coefficients, B ; SPSS uses the uppercase B rather than the lower case b). a is a constant, called the intercept, and its value is 47.032 for this homework–achievement example. *The intercept is the predicted score on the dependent variable for someone with a score of zero on the independent variable.* b , the unstandardized regression coefficient, is 1.990. Because we don't have a direct estimate of the error, we'll focus on a different form of the regression equation: $Y' = a + bX$, in which Y' is the predicted value of Y . The completed equation is $Y' = 47.032 + 1.990X$, meaning that to predict a person's Math Achievement score we can multiply his or her report of time spent on Math Homework by 1.990 and add 47.032. Thus, the predicted score for a student who does no homework would be 47.032. The predicted score for an 8th-grader who does 1 hour of homework is 49.022

Model	Coefficients ^a							
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	Intercept (Constant)	47.032	1.694		.27.763	.000	43.670	50.393
	MATHHOME Time							
	Spent on Math Homework per Week	1.990	.595	.320	3.344	.001	.809	3.171

a. Dependent Variable: MATHACH Math Achievement Test Score

Figure 1.7 Results of the regression of Math Achievement on Math Homework: Regression Coefficients.

($1 \times 1.990 + 47.032$), the predicted score for a student who does 2 hours of homework is 51.012 ($2 \times 1.990 + 47.032$), and so on.

Several questions may spring to mind after these last statements. Why, for example, would we want to predict a student's Achievement score (Y') when we already know the student's real Achievement score? The answer is that we want to use this formula to summarize the relation between homework and achievement for all students at the same time. We may also be able to use the formula for other purposes: to predict scores for another group of students or, to return to the original purpose, to predict Lisa's likely future math achievement, given her time spent on math homework. Or we may want to know what would likely happen if a student or group of students were to increase or decrease the time they spent on math homework.

Interpretation

But to get back to our original question, we now have some very useful information for Lisa, contained within the regression coefficient ($b = 1.99$), because this coefficient tells us the amount we can expect the outcome variable (Math Achievement) to change for each 1-unit change in the independent variable (Math Homework). Because the Homework variable is in hours spent per week, we can make this statement: "For each additional hour students spend on Mathematics Homework every week, they can expect to see close to a 2-point increase in Math Achievement test scores." Now, Achievement test scores are not that easy to change; it is much easier, for example, to improve grades than test scores (Keith, Diamond-Hallam, & Fine, 2004), so this represents an important effect. Given the standard deviation of the test scores (10 points), a student should be able to improve his or her scores by a standard deviation by studying a little more than 5 extra hours a week; this could mean moving from average-level to high-average-level achievement. Of course, this proposition might be more interesting to a student who is currently spending very little time studying than to one who is already spending a lot of time working on math homework.

The Regression Line

The regression equation may be used to graph the relation between Math Homework and Achievement, and this graph can also illustrate nicely the predictions made in the previous paragraph. The intercept (a) is the value on the Y (Achievement) axis for a value of zero for X (Homework); in other words, the intercept is the value on the Achievement test we would expect for someone who does no homework. We can use the intercept as one data point for drawing the regression line ($X = 0, Y = 47.032$). The second data point is simply the point defined by the mean of X ($M_x = 2.200$) and the mean of Y ($M_y = 51.410$). The graph, with these two data points highlighted, is shown in Figure 1.8. We can use the graph and data to

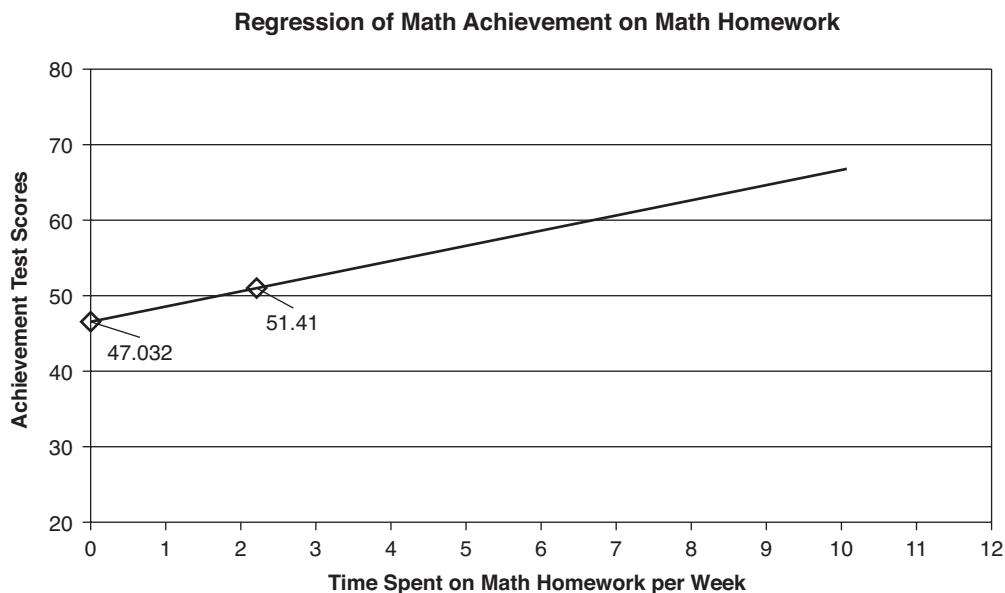


Figure 1.8 Regression line for Math Achievement on Math Homework. The line is drawn through the intercept and the joint means of X and Y .

check the calculation of the value of b , which is equal to the slope of the regression line. The slope is equal to the increase in Y for each unit increase in X (or the rise of the line divided by the run of the line); we can use the two data points plotted to calculate the slope:

$$\begin{aligned}
 b &= \frac{\text{rise}}{\text{run}} = \frac{M_y - a}{M_x - 0} \\
 &= \frac{51.410 - 47.032}{2.200} \\
 &= 1.990
 \end{aligned}$$

Let's consider for a few moments the graph and these formulas. The slope represents the predicted increase in Y for each unit increase in X . For this example, this means that for each unit—in this case, each hour—increase in Homework, Achievement scores increase, on average, 1.990 points. This, then, is the interpretation of an unstandardized coefficient: It is the predicted increase in Y expected for each unit increase in X . When the independent variable has a meaningful metric, like hours spent studying Mathematics every week, the interpretation of b is easy and straightforward. We can also generalize from this group-generated equation to individuals (to the extent that they are similar to the group that generated the regression equation). Thus the graph and b can be used to make predictions for others, such as Lisa. She can check her current level of homework time and see how much payoff she might expect for additional time (or how much she can expect to lose if she studies less). The intercept is also worth noting; it shows that the average Achievement test score for students who do no studying is 47.032, slightly below the national average.

Because we are using a modern statistical package, there is no need to draw the plot of the regression line ourselves; any such program will do it for us. Figure 1.9 shows the data points and regression line drawn using SPSS (a scatterplot was created using the graph feature; see www.tzkeith.com for examples). The small circles in this figure are the actual data points;

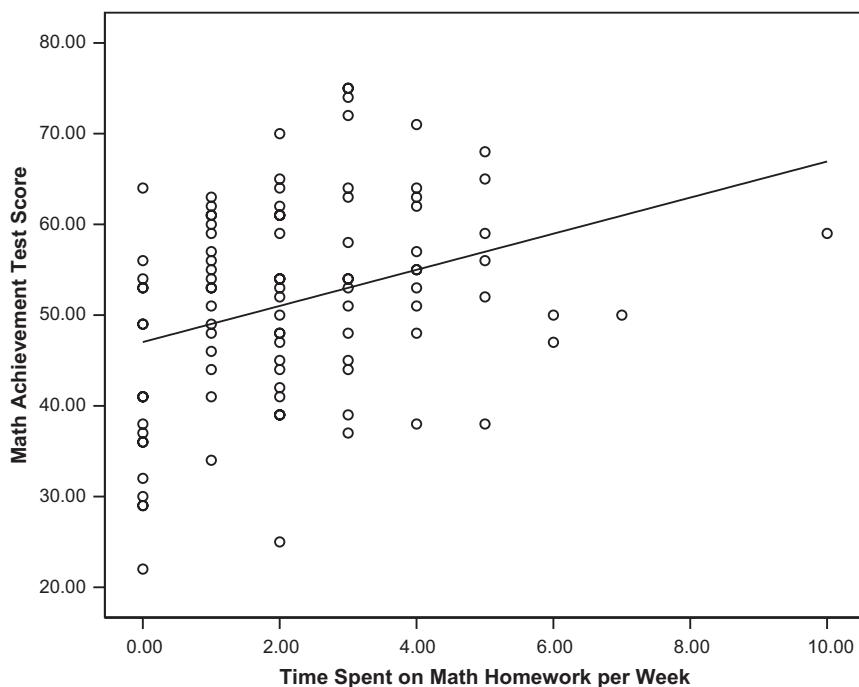


Figure 1.9 Regression line, with data points, as produced by the SPSS Scatter/Dot graph command.

notice how variable they are. If the R were larger, the data points would cluster more closely around the regression line. We will return to this topic in a subsequent chapter.

Statistical Significance of Regression Coefficients

There are a few more details to study for this regression analysis before stepping back and further considering the meaning of the results. With multiple regression, we will also be interested in whether each regression coefficient is statistically significant. Return to the table of regression coefficients (Figure 1.7), and note the columns labeled t and Sig. . The values corresponding to the regression coefficient are simply the results of a t test of the statistical significance of the regression coefficient (b). The formula for t is one of the most ubiquitous in statistics (Kerlinger, 1986):

$$t = \frac{\text{statistic}}{\text{standard error of the statistic}}, \text{ or, in this case,}$$

$$t = \frac{b}{SE_b} = \frac{1.990}{.595} = 3.345.$$

As shown in Figure 1.7, the value of t is 3.344, with $N - k - 1$ degrees of freedom (98). If we look up this value in Excel (using the function T.DIST.2T), we find the probability of obtaining such a t by chance is .001171 (a two-tailed test) rounded off to .001 (the value shown in the table). We can reject the null hypothesis that the slope of the regression line is zero. As a general rule of thumb, with a reasonable sample size (say 100 or more), a t of 2 or greater will be statistically significant with a probability level of .05 and a two-tailed (nondirectional) test.

This finding of the statistical significance of the regression coefficient for Homework does not tell us anything new with our simple regression; the results are the same as for the F test

of the overall regression. You probably recall from previous statistics classes that $t^2 = F$. Here t^2 indeed does equal F (as always, within errors of rounding). When we progress to multiple regression, however, the overall regression may be significant, but the regression coefficients for some of the independent variables may not be statistically significant, whereas others are significant.

Confidence Intervals

We calculated the t previously by dividing the regression coefficient by its standard error. The standard error and the t have other uses, however. In particular, we can use the standard error to estimate a confidence interval around the regression coefficient. Keep in mind that b is an estimate, but what we are really interested in is the likely true (population) value of the regression coefficient (or slope, or b) in the population. The use of confidence intervals makes this underlying thinking more obvious. The 95% confidence interval (CI) is also shown in Figure 1.7 (.809 to 3.171). A common (but less-than-accurate) interpretation of this range is “there is a 95% chance that the true (but unknown) regression coefficient is somewhere within the range .809 to 3.171.” This interpretation is inaccurate because the information we have is about this sample, and our CI is based on this sample, not the true value. A better interpretation, still based on the assumption that the CI we calculated provides “a range of plausible values” for the regression coefficient (Cumming & Finch, 2005, p. 174) is “we can be 95% confident that the CI of .809 to 3.171 includes the actual (population) value of b ”. Another alternative is “If we were to conduct this study and compute CIs repeatedly, around 95% of those CIs should include the true value of the regression coefficient” (adapted from Cumming & Finch, 2005, pp. 174–175; see also Cumming, Fidler, Kalinowski, & Lai, 2012).

The fact that this CI range does not include zero is equivalent to the finding that the b is statistically significant; if the range did include zero, our conclusion would be that we could not say with confidence that the coefficient was different from zero (see Cumming & Finch, 2005; Cumming, Fidler, Kalinowski, & Lai, 2012; or Thompson, 2006, for further information about confidence intervals). The interpretations shared here are based on those discussed in these two Cumming references.

Although the t tells us that the regression coefficient is statistically significantly different from zero, the confidence interval can be used to test whether the regression coefficient is different from any specified value. Suppose, for example, that previous research had shown a regression coefficient of 3.0 for the regression of Math Achievement on Math Homework for high school students, meaning that for each hour of Homework students completed, their Achievement increased by 3 points. We might reasonably ask whether our finding for 8th-graders is inconsistent; the fact that our 95% confidence interval includes the value of 3.0 means that our results are not statistically significantly different from the high school results.

We also can calculate intervals for any level of confidence. Suppose we are interested in the 99% confidence interval. Conceptually, we are forming a normal curve of possible b 's, with our calculated b as the mean. Envision the 99% confidence interval as including 99% of the area under the normal curve so that only the two very ends of the curve are not included. To calculate the 99% confidence interval, you will need to figure out the numbers associated with this area under the normal curve; we do so by using the standard error of b and the t table. Return to Excel (or a t table) and find the t associated with the 99% confidence interval. To do so, use the inverse of the usual t calculator, which will be shown when you select T.INV.2T as the function in Excel. This will allow us to type in the probability level in which we are interested (.01, or $1 - .99$) and the degrees of freedom (98). As shown in Figure 1.10, the t value associated with this probability is 2.627, which we multiply times the standard error ($.595 \times 2.627 = 1.563$). We then add and subtract this product from the b to find the 99% confidence interval: $1.990 \pm 1.563 = .427 - 3.553$. We can be 99% confident that the CI of .427 to 3.553 includes the actual (population) value of b . If we were to conduct this study

14 • MULTIPLE REGRESSION

100 times, and each time calculated a confidence interval, 99 times out of 100 the confidence intervals would include the true value of b . This range does not include a value of zero, so we know that the b is statistically significant at this level ($p < .01$) as well, and we can determine whether our calculated b is different from values other than zero, as well.

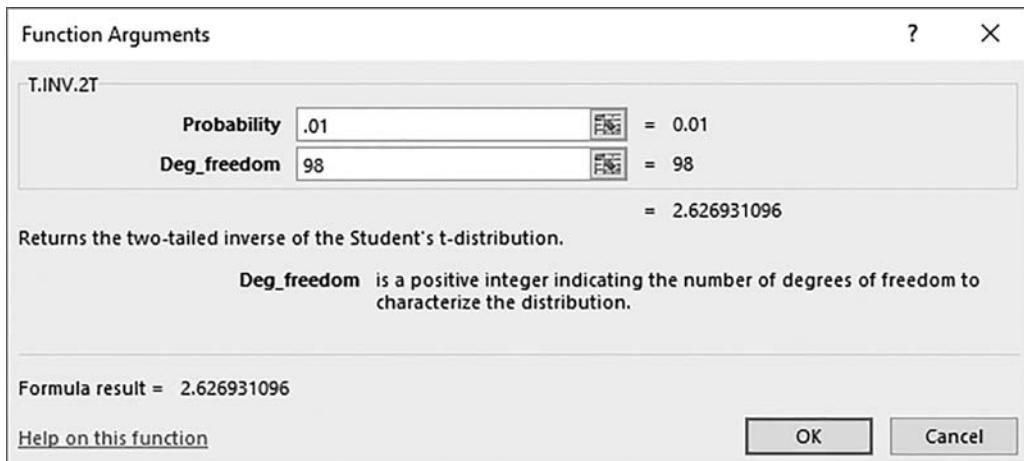


Figure 1.10 Using Excel to calculate a t value for a given probability level and degrees of freedom.

To review, we calculated the confidence intervals as follows:

1. Pick a level of confidence (e.g., 99%).
2. Convert to a probability (.99) and subtract that probability from 1 ($1 - .99 = .01$).
3. Look up this value with the proper degrees of freedom in the (inverse) t calculator or a t table. (Note that these directions are for a two-tailed test.) This is the value of t associated with the probability of interest.
4. Multiply this t value times the standard error of b , and add and subtract the product from the b . This is the confidence interval around the regression coefficient.

The Standardized Regression Coefficient

We skipped over one portion of the regression printout shown in Figure 1.7, the standardized regression coefficient, or Beta (β). Recall that the unstandardized coefficient is interpreted as the change in the outcome for each unit change in the influence. In the present example, the b of 1.990 means that for each 1-hour change in Homework, predicted Achievement goes up by 1.990 points. The β is interpreted in a similar fashion, but the interpretation is in standard deviation (SD) units. The β for the present example (.320) means that for each SD increase in Homework, Achievement will increase, on average, by .320 standard deviation, or about a third of a SD . The β is the same as the b would be if we standardized both the independent and dependent variables (converted them to z -scores).

It is simple to convert from b to β , or the reverse, by taking into account the SD s of each variable. The basic formula is:

$$\beta = b \frac{SD_x}{SD_y} \text{ or } b = \beta \frac{SD_y}{SD_x}.$$

$$\text{So, using the data from Figures 1.3 and 1.6, } \beta = 1.990 \frac{1.815}{11.286} \\ = .320$$

Note that the standardized regression coefficient is the same as the correlation coefficient. This is the case with simple regression, with only one predictor, but will not be the case when we have multiple predictors (it does, however, illustrate that a correlation coefficient is also a type of standardized coefficient).

With a choice of standardized or unstandardized coefficients, which should you interpret? This is, in fact, a point of debate (cf., Kenny, 1979, chap. 13; Pedhazur, 1997, chap. 2), but my position is simply that both are useful at different times. We will postpone until later a discussion of the advantages of each and the rules of thumb for when to interpret each. In the meantime, simply remember that it is easy to convert from one to the other.

REGRESSION IN PERSPECTIVE

Relation of Regression to Other Statistical Methods

How do the methods discussed previously and throughout this book fit with other methods with which you are familiar? Many users of this text will have a background in analytic methods, such as *t* tests and analysis of variance (ANOVA). It is tempting to think of these methods as doing something fundamentally different from regression. After all, ANOVA focuses on differences across groups, whereas regression focuses on the prediction of one variable from others. As you will learn here, however, the processes are fundamentally the same and, in fact, ANOVA and related methods are subsumed under multiple regression and can be considered special cases of multiple regression (Cohen, 1968). Thinking about multiple regression may indeed require a change in your thinking, but the actual statistical processes are the same.

Let's demonstrate that equivalence in two ways. First, most modern textbooks on ANOVA teach or at least discuss ANOVA as a part of the general linear model (Howell, 2013; Thompson, 2006). Remember formulas along the lines of $Y = \mu + \beta + e$, which may be stated verbally as any person's score on the dependent variable Y is the sum of the overall mean μ , plus variation due to the effect of the experimental treatment (β), plus (or minus) random variation due to the effect of error (e).

Now consider a simple regression equation: $Y = a + bX + e$, which may be verbalized as any person's score on the dependent variable is the sum of a constant that is the same for all individuals (a), plus the variation (b) due to the independent variable (X), plus (or minus) random variation due to the effect of error (e). As you can see, these are basically the same formulas with the same basic interpretation. The reason is that ANOVA is a part of the general linear model; multiple regression is virtually a direct implementation of the general linear model.

Second, consider several pieces of computer printout. The first printout, shown in Figure 1.11, shows the results of a *t* test examining whether boys or girls in the National Education Longitudinal Study (NELS) score higher on the 8th-grade Social Studies Test (Appendix A and the website www.tzkeith.com provide more information about the NELS data; the actual variables used were BYTxHStd and Sex_d). We will not delve into these data or these variables in depth right now; for the time being I simply want to demonstrate the consistency in findings across methods of analysis. For this analysis, Sex is the independent variable, and the Social Studies Test score is the dependent variable. The figure shows that 8th-grade girls score about a half a point higher on the test than do 8th-grade boys. The results suggest no statistically significant differences between boys and girls: the *t* value was .689, and the probability that this magnitude of difference would happen by chance (given no difference

16 • MULTIPLE REGRESSION

in the population) was .491, which means that this difference is not at all unusual. If we use a conventional cutoff that the probability must be less than .05 to be considered statistically significant, this value (.491) is obviously greater than .05 and thus would not be considered statistically significant. For now, focus on this value (the probability level, labeled Sig.) in the printout.

The next snippet of printout (Figure 1.12) shows the results of a one-way analysis of variance. Again, focus on the column labeled Sig. The value is the same as for the t test; the results are equivalent. You probably aren't surprised by this finding because you remember that with two groups a t test and an ANOVA will produce the same results and that, in fact, $F = t^2$. (Check the printouts; does $F = t^2$ within errors of rounding?)

Now, focus on the third snippet in Figure 1.13. This printout shows some of the results of a regression of the 8th-grade Social Studies Test score on student Sex. Or, stated differently, this printout shows the results of using Sex to predict 8th-grade Social Studies scores. Look at the Sig. column. The probability is the same as for the t test and the ANOVA: .491! (And check out the t associated with Sex.) All three analyses produce the same results and the same answers. The bottom line is this: the t test, ANOVA, and regression tell you the same thing.

Group Statistics					
		N	Mean	Std. Deviation	
Social Studies Standardized Scores	Male	499	51.14988	10.180993	
	Female	462	51.58123	9.155953	

Independent Samples Test					
	t-test for Equality of Means				
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Social Studies Standardized Score	.689	959	.491	-.431346	.626385

Figure 1.11 Results of a t test of the effects of sex on 8th-grade students' social studies achievement test scores.

			Sum of Squares	df	Mean Square	F	Sig.
Social Studies Standardized Score	Between Groups	(Combined)	44.634	1	44.634	.474	.491
	Within Groups		90265.31	959	94.124		
	Total		90309.95	960			

Figure 1.12 Analysis of variance results of the effects of sex on 8th-grade students' social studies achievement test scores.

Regression ^a						
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1 SEX	.431	.626	.022	.689	.491	

a. Dependent Variable: Social Studies Standardized Score

Figure 1.13 Results of the regression of 8th-grade students' social studies achievement test scores on sex.

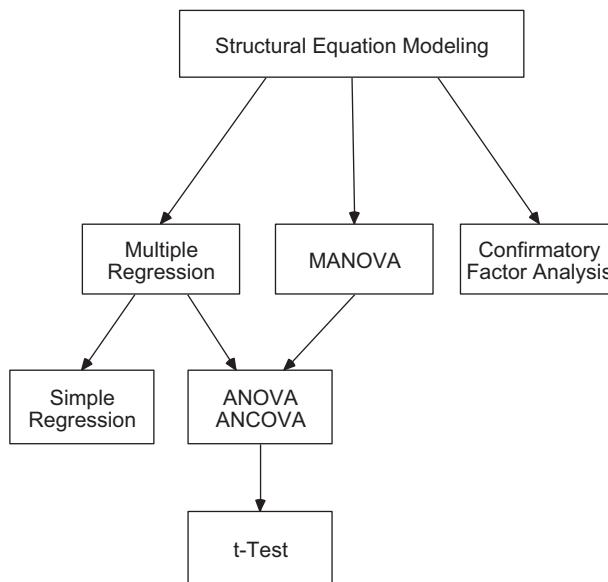


Figure 1.14 Relations among several statistical techniques. ANOVA may be considered a subset of multiple regression; multiple regression, in turn, may be considered a subset of structural equation modeling.

Another way of saying this is that multiple regression subsumes ANOVA, which subsumes a *t* test. And, in turn, multiple regression is subsumed under the method of structural equation modeling, the focus of the second half of this book. Or, if you prefer a pictorial representation, look at Figure 1.14. The figure could include other methods, and portions could be arranged differently, but for our present purposes the lesson is that these seemingly different methods are, in fact, all related.

In my experience, students schooled in ANOVA sometimes find it difficult to make the switch to multiple regression. And not just students; it is not uncommon to see academics conduct research in which ANOVA was used to perform an analysis that would have been better conducted through multiple regression. Given the example previously, this may seem reasonable; after all, they do the same thing, right? No. Regression subsumes ANOVA, is more general than ANOVA, and has certain advantages. We will discuss these advantages briefly, and we will return to them as this book progresses.

Explaining Variance

The primary task of science, simply put, is to explain phenomena. In the social sciences, we ask such questions as “Why do some children do well in school, while others do poorly?” or “Which aspects of psychological consultation produce positive change?” We wish to explain the phenomena of school performance or consultation outcome. At another level, however, we are talking about explaining variation: variation in school performance, such that some children perform well, while others do not, and variation in consultation outcome, with some consultees solving their presenting problem and learning a great deal versus those who make little progress. In medicine or nursing we may ask why some patients comply with postoperative instructions closely and some do not. Here, we wish to explain variation in patient compliance.

And how do we seek to explain this variation? Through variation in other variables! We may reason that children who are more motivated will perform better in school, whereas

those who are less motivated will not. In this case, we seek to explain variation in school performance through variation in motivation. In the consultation example, we may reason that consultants who go through the proper sequence of steps in the identification of the problem will be more successful in producing positive change than consultants who simply “wing it.” Here we have posited variation in consultation implementation as explaining variation in consultation outcome. In nursing, we may reason that a combination of visual and verbal instructions will produce better compliance than verbal instructions alone. In this example, we are assuming that variations in instructions will produce variations in postoperative compliance.

Advantages of Multiple Regression

Our statistical procedures analyze variation in one variable as a function of variation in another. In ANOVA, we seek to explain the variation in an outcome, or dependent, variable (e.g., consultation success) through variation in some treatment, or independent variable (e.g., training versus no training of consultants in problem identification). We do the same using regression; we may, for example, regress a measure of school performance (e.g., achievement test scores from high to low), our dependent variable, on a measure of academic motivation (with scores from high to low), our independent variable. One advantage of multiple regression over methods such as ANOVA is that we can use either categorical independent variables (as in the consultation example), or continuous variables (as in the motivation example), or both. ANOVA, of course, requires categorical independent variables. It is not unusual to see research in which a continuous variable has been turned into categories (e.g., a high-motivation group versus a low-motivation group) so that the researcher can use ANOVA in the analysis rather than regression. Such categorization is generally wasteful, however; it discards variance in the independent variable and leads to a weaker statistical test (Cohen, 1983).⁵

But why study only *one* possible influence on school performance? No doubt many plausible variables can help to explain variation in school performance, such as students’ aptitude, the quality of instruction they receive, or the amount of instruction they receive (Carroll, 1963; Walberg, 1981). What about variation in these variables? This is where the *multiple* in multiple regression (MR) comes in; with MR we can use multiple independent variables to explain variation in a dependent variable. In the language of MR, we can regress a dependent variable on multiple independent variables; we can regress school performance on measures of motivation, aptitude, quality of instruction, and quantity of instruction, all at the same time. Here is another advantage of MR: It easily incorporates these four independent variables; an ANOVA with four independent variables would tax even a gifted researcher’s interpretive abilities.

A final advantage of MR revolves around the nature of the research design. ANOVA is often more appropriate for experimental research, that is, research in which there is active manipulation of the independent variable and, preferably, random assignment of subjects to treatment groups. Multiple regression can be used for the analysis of such research (although ANOVA is often easier), but it can also be used for the analysis of nonexperimental research, in which the “independent” variables are not assigned at random or even manipulated in any way. Think about the motivation example again; could you assign students, at random, to different levels of motivation? No. Or perhaps you could try, but you would be deluding yourself by saying to normally unmotivated Johnny, “OK, Johnny, I want you to be highly motivated today.” In fact, in the original example, motivation was not manipulated at all; instead, we simply measured existing levels of motivation from high to low. This, then, was nonexperimental research. Multiple regression is almost always more appropriate for the analysis of nonexperimental research than is ANOVA.

We have touched on three advantages of multiple regression over ANOVA:

1. MR can use both categorical and continuous independent variables.
2. MR can easily incorporate multiple independent variables.
3. MR is appropriate for the analysis of experimental or nonexperimental research.

OTHER ISSUES

Prediction Versus Explanation

Observant readers will notice that I use the term “explanation” in connection with MR (e.g., explaining variation in achievement through variation in motivation), whereas much of your previous experience with MR may have used the term “prediction” (e.g., using motivation to predict achievement). What’s the difference?

Briefly, explanation subsumes prediction. If you can explain a phenomenon, you can predict it. On the other hand, prediction, although a worthy goal, does not necessitate explanation. As a general rule, in this book we will be more interested in explaining phenomena than in predicting them.

Causality

Observant readers may also be feeling queasy by now. After all, isn’t another name for non-experimental research *correlational* research?⁶ And when we make such statements as “motivation helps explain school performance,” isn’t this another way of saying that motivation is one possible cause of school performance? If so (and the answers to both questions are yes), how can I justify what I recommend, given the one lesson that everyone remembers from his or her first statistics class, the admonition “Don’t infer causality from correlations?” Aren’t I now implying that you should break that one cardinal rule of introductory statistics?

Before I answer, I’d like you to take a little “quiz.” It is mostly tongue-in-cheek, but designed to make an important point.

Are these statements true or false?

1. It is improper to infer causality from correlational data.
2. It is inappropriate to infer causality unless there has been active manipulation of the independent variable.

Despite the doubts I may have planted, you are probably tempted to answer these statements as true. Now try these:

3. Smoking increases the likelihood of lung cancer in humans.
4. Parental divorce affects children’s subsequent achievement and behavior.
5. Personality characteristics affect life success.
6. Gravity keeps the moon in orbit around the Earth.

I assume that you answered “true” or “probably true” for these statements. But if you did, your answers are inconsistent with answers of true to statements 1 and 2! Each of these is a causal statement. Another way of stating statement 5, for example, is “Personality characteristics partially cause life success.” And each of these statements is based on observational or correlational data! I, for one, am not aware of any experiments in which Earth’s gravity has been manipulated to see what happens to the orbit of the moon!⁷ And do you think you

20 • MULTIPLE REGRESSION

could randomly assign personality characteristics in an effort to examine subsequent life success?

Now, try this final statement:

7. Research in sociology, economics, and political science is intellectually bankrupt.

I am confident that you should and did answer “false” to this statement. But if you did, this answer is again inconsistent with an answer of true to statements 1 and 2. True experiments are relatively rare in these social sciences; nonexperimental research is far more common.

The bottom line of this little quiz is this: whether we realize it or not, whether we admit it or not, we often do make causal inferences from “correlational” (nonexperimental) data. Here is the important point: under certain conditions, we can make such inferences validly and with scientific respectability. In other cases, such inferences are invalid and misleading. What we need to understand, then, is when such causal inferences are valid and when they are invalid. We will return to this topic later; in the meantime, you should mull over the notion of causal inference. Why, for example, do we feel comfortable making a causal inference when a true experiment has been conducted, but may not feel so in nonexperimental research? These two issues—prediction versus explanation and causality—are ones that we will return to repeatedly in this text.

REVIEW OF SOME BASICS

Before turning to multiple regression in earnest, it is worth reviewing several fundamentals, things you probably know, but may need reminders about. The reason for this quick review may not be immediately obvious, but if you store these tidbits away, you’ll find that occasionally they will come in handy as you learn a new concept.

Variance and Standard Deviation

First is the relation between a variance and a standard deviation; the standard deviation is the square root of the variance: $SD = \sqrt{V}$ or $V = SD^2$. Why use both? Standard deviations are in the same units as the original variables; we thus often find it easier to use SDs . Variances, on the other hand, are often easier to use in formulas and, although I’ve already promised that this book will use a minimum of formulas, some will be necessary. If nothing else, you can use this tidbit for an alternative formula to convert from the unstandardized to the standardized regression coefficient: $\beta = b\sqrt{\frac{V_x}{V_y}}$

Correlation and Covariance

Next is a covariance. Conceptually, the variance is the degree to which one variable varies around its mean. A covariance involves two variables and gets at the degree to which the two variables vary together. When the two variables vary from the mean, do they tend to vary together or independently? A correlation coefficient is a special type of covariance; it is, in essence, a standardized covariance, and we can think of a covariance as an unstandardized correlation coefficient. As a formula, $r_{xy} = \frac{CoV_{xy}}{\sqrt{V_x V_y}} = \frac{CoV_{xy}}{SD_x SD_y}$. Just as with standardized and unstandardized regression coefficients, if we know the standard deviations (or variances) of the variables, we can easily convert from covariances (unstandardized) to correlations

(standardized) and back. Conceptually, you can think of a correlation as a covariance, but one in which the variance of X and Y are standardized. Suppose, for example, you were to convert X and Y to z -scores ($M = 0$, $SD = 1$) prior to calculating the covariance. Since a z score has a SD of 1, our formula for converting from a covariance to a correlation then becomes $r_{xy} = \frac{Cov_{xy}}{1 \times 1}$ when the variables are standardized.

In your reading about multiple regression, and especially about structural equation modeling, you are likely to encounter variance–covariance matrices and correlation matrices. Just remember that if you know the standard deviations (or variances) you can easily convert from one to another. Table 1.1 shows an example of a covariance matrix and the corresponding correlation matrix and standard deviations. As is common in such presentations, the diagonal in the covariance matrix includes the variances.

Table 1.1 Example of a Covariance Matrix and the Corresponding Correlation Matrix. For the Covariance Matrix, the Variances Are Shown in the Diagonal (thus it is a Variance-Covariance Matrix); the Standard Deviations Are Shown Below the Correlation Matrix.

Sample Covariances

	<i>Matrix</i>	<i>Block</i>	<i>Similarities</i>	<i>Vocabulary</i>
Matrix	118.71			
Block	73.41	114.39		
Similarities	68.75	62.92	114.39	
Vocabulary	73.74	64.08	93.75	123.10

Sample Correlations

	<i>Matrix</i>	<i>Block</i>	<i>Similarities</i>	<i>Vocabulary</i>
Matrix	1.00			
Block	0.63	1.00		
Similarities	0.59	0.55	1.00	
Vocabulary	0.61	0.54	0.79	1.00
SDs	10.90	10.70	10.70	11.10

WORKING WITH EXTANT DATA SETS

The data used for our initial regression example were not real but were simulated. The data were modeled after data from the National Education Longitudinal Study (NELS), a portion of which are on the website (www.tzkeith.com) that accompanies this book.

Already existing, or extant, data offer an amazing resource. For our simulated study, we pretended to have 100 cases from one school. With the NELS data included here, you have access to 1,000 cases from schools across the nation. With the full NELS data set, the sample size is over 24,000, and the data are nationally representative. The students who were first surveyed in 8th grade were followed up in 10th and 12th grades and then twice since high school. If the researchers or organization that collected the data asked the questions you are interested in, then why reinvent the wheel only to get a small, local sample? And see <https://nces.ed.gov/surveys/> for many more educational data sets.

The potential drawback, of course, is that the researchers who initially collected the data may not have asked the questions in which you are interested or did not ask them in the best possible manner. As a user of extant data, you have no control over the questions and how they were asked. On the other hand, if questions of interest were asked, you have no need to go collect additional data.

Another potential problem is less obvious. Each such data set is set up differently and may be set up in a way that seems strange to you. Extant data are of variable quality; although the NELS data are very clean, other data sets may be quite messy and using them can be a real challenge. At the beginning of this chapter I mentioned good data analysis habits; such habits are especially important when using existing data.

An example will illustrate. Figure 1.15 shows the frequency of one of the NELS variables dealing with Homework. It is a 10th-grade item (the F1 prefix to the variable stands for first follow-up; the S means the question was asked of students) concerning time spent on math homework. Superficially, it was similar to our pretend Homework variable. But note that the NELS variable is not in hour units but rather in blocks of hours. Thus, if we regress 10th-grade Achievement scores on this variable, we cannot interpret the resulting b as meaning “for each additional hour of Homework . . .” Instead, we can only say something about each additional unit of Homework, with “unit” only vaguely defined. More importantly, notice that one of the response options was “Not taking math class,” which was assigned a value of 8. If we analyze this variable without dealing with this value (e.g., recoding 8 to be a missing value), our interpretation will be incorrect. When working with extant data, you should always look at summary statistics prior to analysis: frequencies for variables that have a limited number of values (e.g., time on Homework) and descriptive statistics, including minimum and maximum, for those with many values (e.g., Achievement test scores). Look

F1S36B2 TIME SPENT ON MATH HOMEWORK OUT OF SCHL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 NONE	141	14.1	14.9	14.9
	1 1 HOUR OR LESS	451	45.1	47.7	62.6
	2 2-3 HOURS	191	19.1	20.2	82.8
	3 4-6 HOURS	97	9.7	10.3	93.0
	4 7-9 HOURS	16	1.6	1.7	94.7
	5 10-12 HOURS	8	.8	.8	95.6
	6 13-15 HOURS	2	.2	.2	95.8
	7 OVER 15 HOURS	6	.6	.6	96.4
	8 NOT TAKING MATH	34	3.4	3.6	100.0
	Total	946	94.6	100.0	
Missing	96 MULTIPLE RESPONSE	8	.8		
	98 MISSING	19	1.9		
	System	27	2.7		
	Total	54	5.4		
Total		1000	100.0		

Figure 1.15 Time spent on Math Homework from the first follow-up (10th grade) of the NELS data. Notice the value of 8 for the choice “Not taking math class.” This value would need to be classified as missing prior to statistical analysis.

for impossible or out of range values, for values that need to be flagged as missing, and for items that should be reversed. Make the necessary changes and recodings, and then look at the summary statistics for the new or recoded variables. Depending on the software you use, you may also need to change the value labels to be consistent with your recoding. Only after you are sure that the variables are in proper shape should you proceed to your analyses of interest.

Some of the variables in the NELS file on the accompanying website have already been cleaned up; if you examine the frequencies of the variable just discussed, for example, you find that the response “Not taking math class” has already been recoded as missing. But many other variables have not been similarly cleaned. The message remains: always check and make sure you understand your variables before analysis. Always, always, always, always, always check your data!

SUMMARY

Many newcomers to multiple regression are tempted to think that this approach does something fundamentally different from other techniques, such as analysis of variance. As we have shown in this chapter, the two methods are in fact both part of the general linear model. In fact, multiple regression is a close implementation of the general linear model and subsumes methods such as ANOVA and simple regression. Readers familiar with ANOVA may need to change their thinking to understand MR, but the methods are fundamentally the same.

Given this overlap, are the two methods interchangeable? No. Because MR subsumes ANOVA, MR may be used to analyze data appropriate for ANOVA, but ANOVA is not appropriate for analyzing all problems for which MR is appropriate. In fact, there are a number of advantages to multiple regression:

1. MR can use both categorical and continuous independent variables.
2. MR can easily incorporate multiple independent variables.
3. MR is appropriate for the analysis of experimental or nonexperimental research.

We will primarily be interested in using multiple regression for explanatory, rather than predictive, purposes. Thus, it will be necessary to make tentative causal inferences, often from nonexperimental data. These are two issues that we will revisit often in subsequent chapters, in order to distinguish between prediction and explanation and to ensure that we make such inferences validly.

This chapter reviewed simple regression with two variables as a prelude to multiple regression. Our example regressed Math Achievement on Math Homework using simulated data. Using portions of a printout from a common statistical package, we found that Math Homework explained approximately 10% of the variance in Math Achievement, which is statistically significant. The regression equation was $Achievement_{predicted} = 47.032 + 1.990 \text{ Homework}$, which suggests that, for each hour increase in time spent on Math Homework, Math Achievement should increase by close to 2 points. We can be 95% confident that the CI range of .809 to 3.171 includes the actual (population) value of the regression coefficient; such confidence intervals may be used to test both whether a regression coefficient differs significantly from zero (a standard test of statistical significance) and whether it differs from other values, such as those found in previous research.

Finally, we reviewed the relation between variances and standard deviations ($SD = \sqrt{V}$) and between correlations and covariances (correlations are standardized covariances). Since many of our examples will use an existing data set, NELS, a portion of which is included on the website for this book, we discussed the proper use of existing, or extant, data. I noted that

24 • MULTIPLE REGRESSION

good data analytic habits, such as always examining the variables we use prior to complex analysis, are especially important when using extant data.

EXERCISES

Think about the following questions. Answer them, however tentatively. As you progress in your reading of this book, revisit these questions on occasion; have your answers changed?

1. Why does MR subsume ANOVA? What does that mean?
2. What's the difference between explanation and prediction? Give a research example of each. Does explanation really subsume prediction?
3. Why do we have the admonition about inferring causality from correlations? What is wrong with making such inferences? Why do we feel comfortable making causal inferences from experimental data but not from nonexperimental data?
4. Conduct the regression analysis used as an example in this chapter (again, the data are found on the website under Chapter 1). Do your results match mine? Make sure you understand how to interpret each aspect of your printout.
5. Using the NELS data (see www.tzkeith.com), regress 8th-grade Math Achievement (ByTxMStd) on time spent on Math Homework (ByS79a). Be sure that you examine descriptive information before you conduct the regression. How do your results compare with those from the example used in this chapter? Which aspects of the results can be compared? Interpret your findings: what do they mean?

Notes

1. Although I here use the terms independent and dependent variables to provide a bridge between regression and other methods, the term independent variable is probably more appropriate for experimental research. Thus, throughout this book I will often use the term influence or predictor instead of independent variable. Likewise, I will often use the term outcome to carry the same meaning as dependent variable.
2. Throughout this text I will capitalize the names of variables, but will not capitalize the constructs that these variables are meant to represent. Thus, Achievement means the variable achievement, which we hope comes close to achievement, meaning the progress that students make in academic subjects in school.
3. With a single predictor, the value of R will equal that of r , with the exception that r can be negative, whereas R cannot. If r were $-.320$, for example, R would equal $.320$.
4. If you are interested, here is how to calculate $ss_{regression}$ and $ss_{residual}$ by hand (actually, with the help of Excel). Use the “homework & ach.xls” version of the data. Use the sum and power function tools in Excel to calculate

$$\sum x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N},$$

$$\sum y^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}, \text{ and}$$

$\sum xy^2 = (\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N})^2$, where the capital X and Y refer to the raw scores. $ss_{regression}$ is then: $ss_{regression} = \frac{\sum xy^2}{\sum x^2}$. And $ss_{residual}$ is $ss_{residual} = \sum y^2 - ss_{regression}$. You should calculate the same values as shown in the output in Figure 1.5. These and other methods of calculation are shown in more depth in Pedhazur (1997).

5. Fortunately, this categorization practice is, I think, becoming less common. Also note that you can analyze both categorical and continuous variables in analysis of covariance, a topic for a subsequent chapter.

- 6 I encourage you to use the term nonexperimental rather than correlational. The term correlational research confuses a statistical method (correlations) with a type of research (research in which there is no manipulation of the independent variable). Using correlational research to describe nonexperimental research would be like calling experimental research ANOVA research.
- 7 Likewise, researchers have not randomly assigned children to divorced versus intact families to see what happens to their subsequent achievement and behavior, nor has anyone assigned personality characteristics at random to see what happens as a result. The smoking example is a little trickier. Certainly, animals have been assigned to smoking versus non-smoking conditions, but I am confident that humans have not. These examples also illustrate that when we make such statements we do not mean that X is the one and only cause of Y . Smoking is not the only cause of lung cancer, nor is it the case that everyone who smokes will develop lung cancer. Thus, you should understand that causality has a probabilistic meaning. If you smoke, you will increase your probability of developing lung cancer.

2

Multiple Regression Introduction

A New Example: Regressing Grades on Homework and Parent Education	27
<i>The Data</i>	27
<i>The Regression</i>	27
<i>Interpretations</i>	33
<i>Figural Representation</i>	34
Questions	35
<i>Controlling for . . .</i>	35
<i>b versus β</i>	36
<i>Comparison Across Samples</i>	38
Direct Calculation of β and R^2	41
Summary	42
Exercises	42
<i>Notes</i>	43

Let's return to the example that was used in Chapter 1, in which we were curious about the effect on math achievement of time spent on math homework. Given our finding of a statistically significant effect, you might reasonably have a chat with your daughter about the influence of homework on achievement. You might say something like "Lisa, these data show that spending time on math homework is indeed important. In fact, they show that for each additional hour you spend on math homework every week, your achievement test scores should go up by approximately 2 points. And that's not just grades but *test scores*, which are more difficult to change. So, you say you are now spending approximately 2 hours a week on math homework. If you spent an additional 2 hours per week, your achievement test scores should increase by about 4 points; that's a pretty big improvement!"¹

Now, if Lisa is anything like my children, she will be thoroughly unimpressed with any argument you, her mere parent, might make, even when you have hard data to back you up. Or perhaps she's more sophisticated. Perhaps she'll point out potential flaws in your reasoning and analyses. She might say that she cares not one whit whether homework affects achievement test scores; she's only interested in grades. Or perhaps she'll point to other variables you should have taken into account. She might say, "What about the parents? Some of the kids in my school have very well-educated parents, and those are usually the kids who do well on tests. I'll bet they are also the kids who study more, because their parents think

it's important. You need to take the parents' education into account." Your daughter has in essence suggested that you have chosen the wrong outcome variable and have neglected what we will come to know as a "common cause" of your independent and dependent variables. You suspect she's right.

A NEW EXAMPLE: REGRESSING GRADES ON HOMEWORK AND PARENT EDUCATION

Back to the drawing board. Let's take this example a little further and pretend that you devise a new study to address your daughter's criticisms. This time you collect information on the following:

1. 8th-grade students' overall Grade-point average in all subjects (on a standard 100-point scale).
2. The level of Education of the students' parents, in years of schooling (i.e., a high school graduate would have a score of 12, a college graduate a score of 16). Although you collect data for both parents, you use the data for the parent with the higher level of education. For students who live with only one parent, you use the years of schooling for the parent the student lives with.
3. Average time spent on Homework per week, in hours, across all subjects.

The data are in three files on the Web site (www.tzkeith.com), under Chapter 2: chap2, hw grades.sav (SPSS file), chap2, hw grades.xls (Excel file), and chap2, hw grades data.txt (DOS text file). As in the previous chapter, the data are simulated.

The Data

Let's look at the data. The summary statistics and frequencies for the Parent Education variable are shown in Figure 2.1. The figure also shows the frequencies displayed graphically in a histogram (I'm a big fan of pictorial depictions of data). As shown, parents' highest level of education ranged from 10th grade to 20 years, suggesting a parent with a doctorate; the average level of education was approximately 2 years beyond high school (14.03 years). As shown in Figure 2.2, students reported spending, on average, about 5 hours (5.09 hours) on homework per week, with four students reporting spending 1 hour per week and one reporting 11 hours per week. Most students reported between 4 and 7 hours per week. The frequencies and summary statistics look reasonable. The summary statistics for students' GPAs are shown in Figure 2.3. The average GPA was 80.47, a B minus. GPAs ranged from 64 to 100; again, the values look reasonable.

The Regression

Next we regress students' GPA on Parent Education and Homework. Both of the explanatory variables (Homework and Parent Education) were entered into the regression equation at the same time, in what we will call a *simultaneous* regression. Figure 2.4 shows the inter-correlations among the three variables. Note that the correlation between Homework and Grades (.327) is only slightly higher than was the correlation between Math Homework and Achievement in Chapter 1. Parent Education, however, is correlated with both time spent on Homework (.277) and Grade-point average (.294). It will be interesting to see what the multiple regression looks like.

Statistics

pared Parents' Education (highest)

N	Valid	100
	Missing	0
Mean		14.0300
Median		14.0000
Mode		13.00
Std. Deviation		1.93038
Variance		3.726
Minimum		10.00
Maximum		20.00

pared Parents' Education (highest)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	10.00	4	4.0	4.0	4.0
	11.00	3	3.0	3.0	7.0
	12.00	13	13.0	13.0	20.0
	13.00	23	23.0	23.0	43.0
	14.00	19	19.0	19.0	62.0
	15.00	15	15.0	15.0	77.0
	16.00	12	12.0	12.0	89.0
	17.00	8	8.0	8.0	97.0
	18.00	2	2.0	2.0	99.0
	20.00	1	1.0	1.0	100.0
Total		100	100.0	100.0	

Histogram

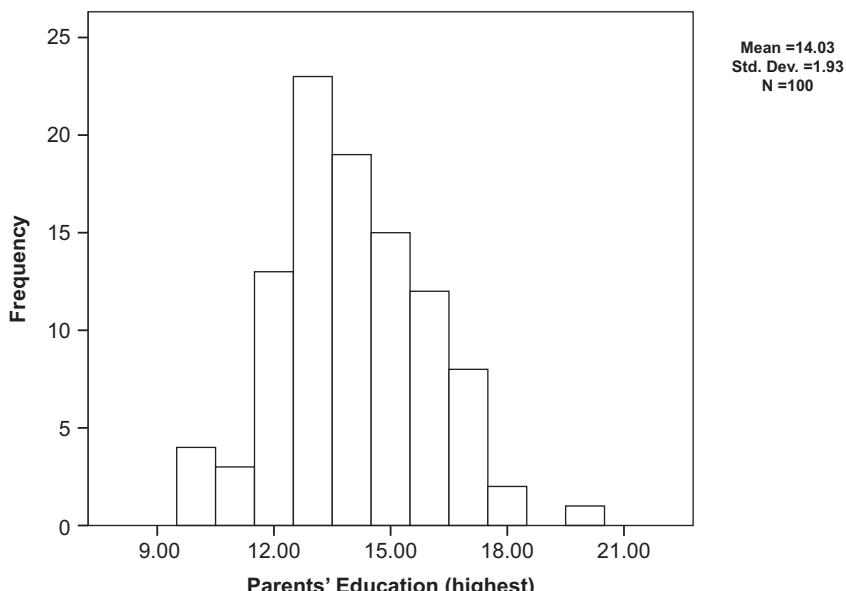


Figure 2.1 Descriptive statistics for Parent Education for 8th-graders.

Statistics

hwork Average Time Spent on Homework per Week

N	Valid	100
	Missing	0
Mean		5.0900
Median		5.0000
Mode		5.00
Std. Deviation		2.05527
Variance		4.224
Minimum		1.00
Maximum		11.00

hwork Average Time Spent on Homework per Week

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	4	4.0	4.0	4.0
	2.00	8	8.0	8.0	12.0
	3.00	8	8.0	8.0	20.0
	4.00	18	18.0	18.0	38.0
	5.00	24	24.0	24.0	62.0
	6.00	12	12.0	12.0	74.0
	7.00	14	14.0	14.0	88.0
	8.00	8	8.0	8.0	96.0
	9.00	2	2.0	2.0	98.0
	10.00	1	1.0	1.0	99.0
	11.00	1	1.0	1.0	100.0
Total		100	100.0	100.0	

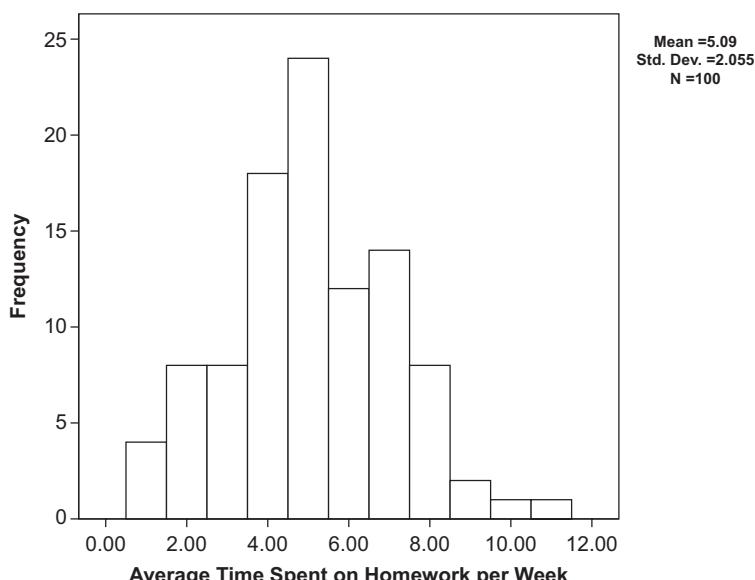


Figure 2.2 Descriptive statistics for Homework Time for 8th-graders.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
grades Grade Point Average	100	64.00	100.00	80.4700	7.62300	58.110
Valid N (listwise)	100					

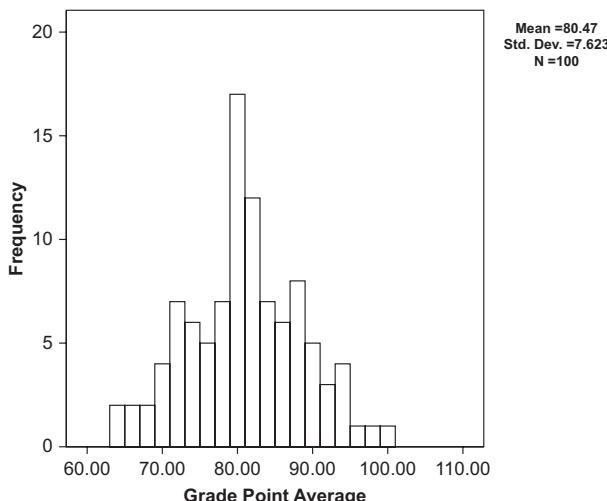


Figure 2.3 Descriptive statistics for the outcome variable, Grade Point Average, for 8th-graders.

Correlations

	GRADES Grade Point Average	GRADES Grade Point Average	PARED Parents' Education (highest)	HWORK Average Time Spent on Homework per Week
Pearson Correlation	GRADES Grade Point Average	1.000	.294	.327
	PARED Parents' Education (highest)	.294	1.000	.277
	HWORK Average Time Spent on Homework per Week	.327	.277	1.000
Sig. (1-tailed)	GRADES Grade Point Average	.	.001	.000
	PARED Parents' Education (highest)	.001	.	.003
	HWORK Average Time Spent on Homework per Week	.000	.003	.
N	GRADES Grade Point Average	100	100	100
	PARED Parents' Education (highest)	100	100	100
	HWORK Average Time Spent on Homework per Week	100	100	100

Figure 2.4 Correlations among Grades, Parent Education, and Homework time.

Multiple R

Figure 2.5 shows the multiple correlation coefficient (denoted as a capital R , a value of .390, and sometimes referred to as the “mult R ”). Also shown is the squared multiple correlation, R^2 , of .152, which shows that the two explanatory variables, Homework and Parent Education level, together explain 15.2% of the variance in students’ GPAs.

Are you surprised that the R is not larger? Perhaps you expected that R might equal the sum of the correlations of the two explanatory variables with GPA (i.e., .294 + .327)? You cannot add correlation coefficients in this way, but you can sometimes add variances, or r^2 s. But when you try adding variances, you find that $R^2 \neq r_{\text{ParEdGPA}}^2 + r_{\text{HWWorkGPA}}^2$; that is, $.152 \neq .294^2 + .327^2$. Why not? The short answer is that R^2 is not equal to the sum of the r^2 s because the two explanatory variables are also correlated with each other. Ponder why that might be while we look at the remainder of the regression results.

The ANOVA table, also shown in Figure 2.5, shows that the regression is statistically significant $F(2, 97) = 8.697, p < .001$. What does that mean? It means that taken together, in some optimally weighted combination, Homework and Parent Education level predict or explain students’ Grades to a statistically significant degree. (We will examine what is meant by an “optimally weighted combination” in the next chapter.)

Either of the two formulas from Chapter 1 for calculating F will work with multiple regression:

$$F = \frac{ss_{\text{regression}}/df_{\text{regression}}}{ss_{\text{residual}}/df_{\text{residual}}} \text{ or } F = \frac{R^2/k}{(1-R^2)/(N-K-1)}.$$

Recall that the df for the regression is equal to k , which is equal to the number of independent (predictor) variables, in this case 2. The df for the residual is equal to the total N , minus k , minus 1 (97). Try both of these formulas to make sure your answer is the same as that shown in the figure (within errors of rounding).

Model Summary

1	R	R Square	Adjusted R Square	Std. Error of the Estimate
Model	.390 ^a	.152	.135	7.0916

a. Predictors: (Constant), HWORK Average Time Spent on Homework per Week, PARED Parents' Education (Highest)

ANOVA^b

1		Sum of Squares	df	Mean Square	F	Sig.
Model	Regression	874.739	2	437.369	8.697	^a
	Residual	4878.171	97	50.290		
	Total	5752.910	99			

a. Predictors: (Constant), HWORK Average Time Spent on Homework per Week, PARED Parents' Education (Highest)

b. Dependent Variable: GRADES Grade Point Average

Figure 2.5 Model summary and test of statistical significance of the regression of Grades on Parent Education and Homework.

Regression Coefficients

Next we turn to the regression coefficients (Figure 2.6). With simple regression, there was only one b , and its probability was the same as that of the overall regression equation. The corresponding β was equal to the original correlation. All this changes with multiple independent variables. With multiple regression, each independent variable has its own regression coefficient; the b for Parent Education is .871, and the b for Time Spent on Homework is .988; the intercept is 63.227.

The regression equation is $Y = 63.227 + .871X_1 + .988X_2 + \text{error}$ or, for *predicted* Grades, $\text{Grades}_{(\text{predicted})} = 63.227 + .871\text{ParEd} + .988\text{HWork}$.

We could use this formula to predict any participant's GPA from his or her values on Homework and Parent Education. If a student spends 5 hours per week on homework and one of the parents completed college (16 years of education), his or her predicted GPA would be 82.103.

With multiple regression, we can test each independent variable separately for statistical significance. It is not unusual, especially when we have a half-dozen or so variables in the regression equation, to have a statistically significant R^2 but to have one or more independent variables that are not statistically significant (an example is shown in Chapter 4). For the present case, note that the t ($t = b/se_b$) associated with Parent Education is 2.266 ($p = .026$), and the 95% confidence interval for the b is .108 – 1.633. The fact that this range does not include zero tells us the same thing as the significance level of b : for a probability level of .05, the variable Parent Education is a statistically significant predictor of GPA. The regression coefficient (.871) suggests that, for each additional year of parental schooling, students' GPA will increase by .871, or close to one point on the 100-point GPA scale, once time spent on homework is taken into account.

Of greater interest is the regression coefficient for time spent on Homework, .988, which suggests that for each additional hour spent studying per week GPA should increase by close to 1 point (controlling for Parent Education). To increase GPA by 5 points, a student would need to spend a little more than 5 extra hours a week studying, or about an extra hour every night. As shown in the figure, this value is also statistically significant ($p = .007$).

You might wonder which of these two variables, Parent Education or Homework, has a stronger effect on Grades? You may be tempted to conclude that it is Homework, based on a comparison of the b 's. You would be correct, but for the wrong reason. The Parent Education and Homework variables have different scales, so it is difficult to compare them. The b for Parent Education pertains to years of schooling, whereas the b for Homework pertains to hours of homework. If we want to compare the relative influence of these two variables

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients Beta	<i>t</i>	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	63.227	5.240	12.067	.000	52.828	73.627
	PARED Parents' Education (Highest)	.871	.384	.220	2.266	.026	.108
	HWORK Average Time Spent on Homework per Week	.988	.361	.266	2.737	.007	.272
							1.704

a. Dependent Variable: GRADES Grade Point Average

Figure 2.6 Unstandardized and standardized regression coefficients for the regression of Grades on Parent Education and Homework.

we need to compare the β 's, the *standardized* regression coefficients. When we do, we see that Homework ($\beta = .266$) is indeed a slightly more powerful influence on GPA than is Parent Education ($\beta = .220$). Each standard deviation increase in Homework will lead to .266 of a *SD* increase in Grades, whereas a standard deviation increase in Parent Education will result in .220 of a *SD* in Grades. Let's postpone asking whether this difference is statistically significant.²

As an aside, think about which of these two findings is more interesting. I assume most of you will vote for the homework finding, for the simple reason that homework time is potentially manipulable, whereas parent education is unlikely to change for most students. Another way of saying this is that the homework finding has implications for intervention, or school or home rules. Still another way to make a similar point is to note that our original interest was in the effect of homework on GPA, and we included the variable Parent Education in the analysis as a background or "control" variable.

Interpretations

Formal

Let's consolidate the interpretation of these findings and then move on to discuss several other issues. Our first, formal interpretation might be something along these lines:

This research was designed to determine the influence of time spent on homework on 8th-grade students' Grade-point averages (GPAs) while controlling for parents' level of education. Students' 8th-grade GPAs were regressed on their average time spent on homework per week and the higher of their parents' levels of education. The overall multiple regression was statistically significant ($R^2 = .152$, $F[2, 97] = 8.697$, $p < .001$), and the two variables (Homework and Parent Education) accounted for 15% of the variance in Grades. Each of the two independent variables also had a statistically significant effect on Grades. The unstandardized regression coefficient (b) for Parent Education was .871 ($t[97] = 2.266$, $p = .026$), meaning that for each additional year of parents' schooling, students' Grades increase by .871 points, controlling for time spent on homework. Of more direct interest was the b associated with time spent on Homework ($b = .988$, $t[97] = 2.737$, $p = .007$). This finding suggests that, for each hour students spend on Homework per week, their Grade-point average will increase by .988 points, controlling for parent education.

Although I have written this interpretation as it might appear in a journal, an example this simple would not be accepted for publication. It is included to illustrate the interpretation of regression results, however. Note that my interpretation has focused on the unstandardized coefficients; that is because the metrics for all three variables in this example are meaningful (more on this later).

We should take this interpretation a step further and discuss in English what these findings mean.

These results suggest that homework is indeed an important influence on students' grades and that this effect holds even after students' family backgrounds (parent education) are taken into account. Students who want to improve their grades may do so by spending additional time on homework. These findings suggest that each additional hour spent per week should result in close to a 1-point increase in students' overall GPA.

Real World

I believe that it is important also to be able to provide a real-world (versus statistical) interpretation of these findings, in addition to the one that uses all the proper jargon. So, for example, here is how you might interpret these findings to a group of parents:

I conducted research to determine the influence on their grades of the time middle school students spend on homework. I also considered the students' parents' level of education as a background variable. As you might expect, the results indicated that parents' education indeed had an effect on students' grades. Parents with more education had students who earn higher grades. This may be related to the educational environment they provide or numerous other reasons. What is important, however, is that homework also had a strong and important effect on grades. In fact, it had a slightly stronger effect than did parent education levels (readers note that this interpretation is based on the β s). What this means is that students—no matter what their background—can perform at a higher level in school through the simple act of spending additional time on homework. The findings suggest that, on average, each additional hour per week spent on homework will result in a close to 1-point increase in overall grade-point average. So, for example, suppose your daughter generally spends 5 hours per week on homework and has an 80 average. If she spent an additional 5 hours per week on homework—or an additional 1 hour per weekday evening—her average should increase to close to 85. Please note that these are averages, and the effect of homework will vary for individual students.

Since our initial reason for completing this study was because of concerns about your daughter, you should develop an interpretation for her, as well. You might say something like:

You were right, Lisa, about parent education being important. Our new research shows that parents with higher education do indeed have children who earn higher grades in school. But homework is still important, even when you take parents' education into account. And homework is important for your grades in addition to test scores. Our new research shows that for each additional hour you spend per week on homework, your GPA should increase, on average, by close to 1 point. That may seem like a lot of work, but think about it: if you spend 2 hours on homework every night instead of 1 hour, your GPA should increase by close to 5 points. And that's your overall GPA, for the entire grading period, not just one test. It might be worth a try.

Figural Representation

As I mentioned earlier in the chapter, I am a big fan of pictorial representations of data and of analyses. Figure 2.7 shows one such method for displaying regression results pictorially. This path diagram, or path model, shows the variables in the multiple regression in rectangles. Arrows, or paths, are used to signify regression coefficients (in this case, the β s), and the curved, double-headed arrow between the two predictor variables represents the correlation between them. As noted, this model shows the standardized coefficients; it would also be possible to use the unstandardized regression coefficients (in which case we would include the covariance, rather than the correlation, between the Parent Education and Homework Time). We will develop such models in much more depth in Part 2, but they are introduced here because they will prove useful for understanding aspects of multiple regression.

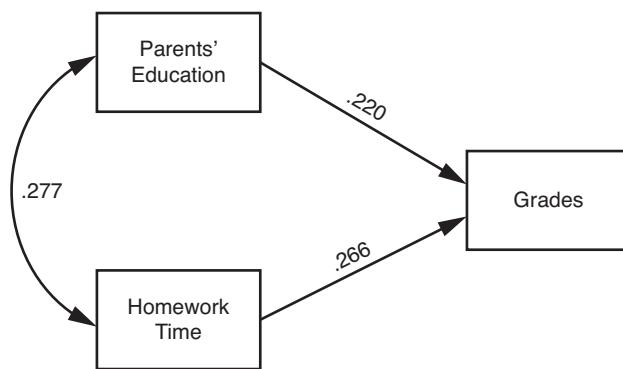


Figure 2.7 Multiple regression results displayed as a path model.

QUESTIONS

Controlling for ...

For many of the interpretations listed previously, you will notice remarks like “their Grade-point average will increase by .988 points, *controlling for Parent Education*” or “once variable X is taken into account.” What do these statements mean? At the most basic level, we add these clarifications to indicate that we have taken into account variables other than the single predictor and single outcome that are being interpreted and thus differentiate this interpretation from one focused on zero-order correlations or simple, bivariate regression coefficients. The two (simple regression coefficient and multiple regression coefficient) are rarely the same, and the MR coefficients will often be smaller.

Another variation of these statements is “Grade-point average will increase by .988 points, *within levels of parent education*.” Consider if we were to regress Grades on Homework for students with parents with 10th-grade educations, and then for those whose parents completed the 11th grade, then for those whose parents completed high school, and so on, through students whose parents completed doctoral degrees. The .988 we calculated for the regression coefficient is conceptually equivalent to the average of the regression coefficients we would get if we were to conduct all these separate regressions.

Of course “control” in this nonexperimental research is not the same as control in the case of experimental research, where we may assign people who have a college education to one versus the other treatment, thus actually controlling which treatment they receive. Instead, we are talking about *statistical* control. With statistical control, we essentially take into account the variation explained by the other variables in the model. We take into account the variation explained by Parent Education when examining the effect of Homework on Grades, and we take into account the variation due to Homework when examining the effect of Parent Education on Grades.

I confess that I have mixed feelings about appending “controlling for . . .” to such interpretive statements. On the one hand, these qualifications are technically correct and provide a sense of the other variables taken into account. On the other hand, if we are correct in our interpretation, that is, discussing the *effect* of homework on GPA, then effects are effects, regardless of what else is “controlled.” Said differently, if we have controlled for the proper variables, then this is indeed a valid estimate of the effect of homework on GPA. If we have not controlled for the proper variables, then it is not a valid estimate. Figuring out the *proper variables* is an issue that we will return to repeatedly in this book and will finally resolve in the beginning chapters of Part 2. At any rate, perhaps these kinds of qualifications

(“controlling for . . .”) are more appropriate for a formal interpretation of results and less so for the English and real-world interpretations. If so, however, then it should be understood that for any interpretation that does not include a qualification like “controlling for x ,” we are also saying, perhaps under our breath, “assuming that I have included the correct variables in my regression equation.” Again, what we mean by effects and when we are correct in such interpretations are topics that we will return to repeatedly in this book.

This discussion makes obvious the chief advantage of multiple, over simple, regression: it allows us to control for other relevant variables. When you conduct nonexperimental (or even experimental) research and try to tease out the effect of one variable on another, you will often be asked questions along the lines of “OK, but did you take into account (control for) variable x ?” We opened this chapter with such a question from Lisa, who argued that we needed to take parent education into account. Multiple regression allows you to take these other variables into account, to control for them statistically. The hard part is figuring out which variables need to be controlled! The other big advantage of multiple regression over simple regression is that, by controlling for additional variables, we increase the variance we are able to explain in the dependent variable; we are able to explain the phenomenon of interest more completely. This advantage was also illustrated with the current example.

Partial and Semipartial Correlations

The preceding discussion has focused on the *effect* of one variable on another while taking a third variable into account. It is also possible to control for other variables without making assumptions about one variable influencing, affecting, or predicting another. That is, it is possible to calculate correlations between two variables, with other variables controlled. Such correlations are termed *partial correlations* and can be thought of as the correlation between two variables with the effects of another variable controlled, or removed, or “partialed” out. We could, for example, calculate the partial correlation between homework and grades, with the effects of parent education removed from homework and grades. We could have several such control variables, calculating, for example, the partial correlation of homework and grades while controlling for both parent education and previous achievement.

It is also possible to remove the effects of the control variable from only *one* of the two variables being correlated. For example, we could examine the correlation of homework (with parent education controlled) with grades. In this example, the effects of parent education are removed only from the homework variable, not the grades variable. This variation of a correlation coefficient is called a *semipartial correlation*. It is also referred to as a *part correlation*.

Although I will mention partial and semipartial correlations at several points in this text, they are not discussed in detail in the text itself but rather in Appendix C. There are several reasons for this decision. First, the topic is somewhat of a detour from the primary topic of Part 1, multiple regression. Second, in my experience, different instructors like to fit this topic in at different places in their lectures. Putting the material in Appendix C makes such placement more flexible. Third, although the topic fits better conceptually in Part 1, I think that partial and semipartial correlations are much easier to explain and understand with reference to the figural, or path, models that are used throughout the text but that are explained in depth in the beginning chapters of Part 2. Feel free to turn to Appendix C at any point that you want to learn more about part and partial correlations, however.

***b* versus β**

Believe it or not, the choice of interpreting the unstandardized versus the standardized regression coefficient can be controversial. It need not be. Briefly, b and β are both useful, but

for different aspects of interpretation. As our examples have already illustrated, b can be very useful when the variables have a meaningful scale. In the present example, Homework time is measured in hours per week, and everyone is familiar with a standard 100-point grade scale. Thus, it makes a great deal of sense to make interpretations like “each hour increase in Homework per week should result in a .988-point increase in overall Grade-point average.” Very often, however, the scales of our independent or dependent variables, or both, are not particularly meaningful. The test score metric used in Chapter 1 is probably not that familiar to most readers, except possibly measurement specialists who encounter T scores often. And the scale of the Parent Education variable used in the present example, although logical, is not very common; a much more common scale might be something along the lines of 1 = did not graduate from high school; 2 = high school graduate; 3 = some college; 4 = college graduate; and so on. This scale may be better from a measurement standpoint in dealing with cases such as someone who attends college for 6 years but never completes a degree, but it is not a readily interpretable metric. We will encounter many other variables without a meaningful metric. In these cases, it makes little sense to interpret b : “each 1-point increase in X should result in a 4-point increase in Y .” What does a 1-point increase in X mean? What does a 4-point increase in Y mean? When the variables of interest do not have a meaningful metric, it makes more sense to interpret β : “each standard deviation increase in X should result in a .25-standard deviation increase in Y .”

As we have already seen, β is generally our interpretive choice when we want to compare the relative importance of several variables in a single regression equation. Different variables in a regression equation generally have different metrics, so it makes no sense to compare unstandardized regression coefficients; it is like comparing apples to oranges. The standardized coefficients place all variables on the same metric (standard deviation units) and thus may be compared in a qualitative manner. Of course, if the independent variables in a regression equation used the same metric, the b 's for each could be compared, but this situation (variables sharing the same metric) is not very common.³

Often we are interested in the policy implications of our regression analyses. Using the present example, you want to give Lisa advice about the likely impact of completing more homework. More broadly, you may want to urge the local school board to encourage teachers to increase homework demands. When you are interested in making predictions about what will happen (“if you spend 5 more hours a week on homework . . .”) or are interested in changing or intervening in a system, or are interested in developing policy based on the findings of a regression analysis, then b is probably a better choice for interpretation if the variables have a meaningful metric.

Finally, we may want to compare our regression results with those from previous research. We may, for example, want to compare the effect of Homework in this example with the apparent effect of Homework in a published study. To compare across samples or populations, b is more appropriate. The reason for this rule of thumb is that different samples likely have different distributions for the same variables. If you measure Homework time in 8th grade and 4th grade, it is likely that the means and standard deviations for Homework will differ in the two grades. These differences in distributions—notably the standard deviations—affect the β 's, but not the b 's. To get an intuitive understanding of this point, look at the regression line shown in Figure 2.8. Assume that the b , which is the slope of the regression line, is .80 and that the β is also .80. (How could this be? The SD 's of the independent and dependent variables are equal.) Now assume that we remove the data in the areas that are shaded. With this new sample, the b could remain the same; the regression line remains the same, just shorter, so its slope remains the same. But what about the β ? Obviously, the SD of the independent variable has decreased because we discarded all data from the shaded area. The SD of the dependent variable will also decrease but not as much as the SD of the independent variable.

Suppose the new SD's are 7 for X and 9 for Y . Now recall from Chapter 1 how we can convert b to β : $\beta = b \frac{SD_x}{SD_y}$; the b remained the same, but the β changed. To return to the original point: to compare regression results across two different samples or studies, b is more appropriate (given that the variables are measured on the same scale for the two samples). These rules of thumb are summarized in Table 2.1.

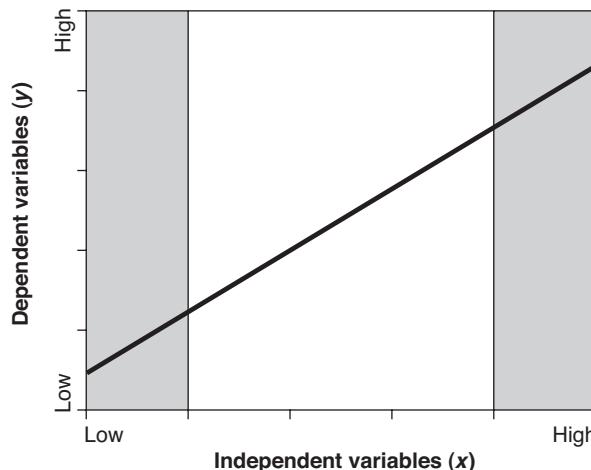


Figure 2.8 Effect of a change in variability on the regression coefficients. The figure shows the regression line from the regression of a hypothetical dependent variable on a hypothetical independent variable.

Table 2.1 Rules of Thumb for When to Interpret b versus β

INTERPRET b :

- When the variables are measured in a meaningful metric
- To develop intervention or policy implications
- To compare effects across samples or studies

INTERPRET β :

- When the variables are not measured in a meaningful metric
 - To compare the relative effects of different predictors in the same sample
-

Again, you may read or hear strong defenses for the routine interpretation of b versus β , or vice versa. Just remember that with knowledge of the SD's of the variables you can easily convert from one to another. Both are useful; they simply are useful for different purposes.

Comparison Across Samples

I have mentioned comparison of regression coefficients across samples or studies. As an example, we might ask whether the effect of Homework on Grades in this chapter is consistent with the estimate we calculated in Chapter 1 examining the effect of Homework on achievement. Unfortunately, these two analyses used different dependent variables (Math Achievement test scores versus overall GPA), making such comparisons difficult. Instead, let's pretend that we redo the research, asking the same questions, on a sample of high school students. The descriptive statistics for this sample are shown in Figure 2.9, and the results of the multiple regression for this sample are shown in Figure 2.10.

Descriptive Statistics

	Mean	Std. Deviation	N
GRADES	81.5348	7.46992	100
PARED	13.8300	2.04028	100
HWORK	6.9800	2.14608	100

Correlations

	GRADES	PARED	HWORK
Pearson Correlation	GRADES	1.000	.191
	PARED	.191	1.000
	HWORK	.354	.368
Sig. (1-tailed)	GRADES	.	.028
	PARED	.028	.
	HWORK	.000	.000
N	GRADES	100	100
	PARED	100	100
	HWORK	100	100

Figure 2.9 Descriptive statistics and correlations among variables for high school students.**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.360 ^a	.130	.112	7.03925

a. Predictors: (Constant), HWORK, PARED

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	717.713	2	358.856	7.242	.001 ^a
	Residual	4806.452	97	49.551		
	Total	5524.165	99			

a. Predictors: (Constant), HWORK, PARED

b. Dependent Variable: GRADES

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	69.984	4.881	14.338	.000	60.297	79.671
	PARED	.258	.373	.692	.490	-.482	.998
	HWORK	1.143	.355	.328	.002	.440	1.847

a. Dependent Variable: GRADES

Figure 2.10 Regression output for the Homework example for high school students.

Note the b associated with Homework in this new regression: 1.143, representing an estimate of the effect of time spent on homework on grades for high school students. The two estimates—1.143 for high school students versus .988 for 8th-graders earlier in this chapter—are obviously different, but are the differences statistically significant? There are several ways we might make this comparison. The easiest is to use the confidence intervals.

Phrase the question this way: Is the present estimate of 1.143 (the value for the high school sample) statistically significantly different from our earlier estimate of .988? Look at the 95% confidence interval for the regression coefficient for the 8th-grade sample (Figure 2.6): .272 to 1.704. The value of 1.143 falls within this range, so we can confidently say that our current value is *not* statistically different from our previous estimate.

Another way of making this determination is the good old *t* test. To ask whether our current value is different from some value other than zero, we make a minor change in the formula: $t = \frac{b - \text{value}}{SE_b}$, where *value* represents the other value to which we wish to compare *b* (Darlington & Hayes, 2017 chap. 4). For the present example, the formula would be

$$\begin{aligned} t &= \frac{.988 - 1.143}{.361} \\ &= -.429 \end{aligned}$$

Using our rule of thumb (*t*'s of 2 or greater are significant; ignore whether *t* is positive or negative), we again see that the high school value is not statistically significantly different at the .05 level from the value estimated for 8th-graders. Or, using the T.DIST.2T function in Excel, we see that this *t* would happen commonly by chance alone ($p = .669$, two-tailed, with 97 *df*; just use the value .429, without the negative sign, in Excel).

Note that this test compares our 8th-grade estimate, with confidence intervals, to a specific value. It is also possible to compare the two regression estimates, considering the standard errors of both. The formula

$$z = \frac{b_1 - b_2}{\sqrt{SE_{b_1}^2 + SE_{b_2}^2}}$$

can be used to compare regression coefficients from two *separate* (independent) regression equations (Cohen & Cohen, 1983, p. 111). It doesn't matter which *b* goes first; it's easiest to make the larger one *b*₁. For the current example,

$$\begin{aligned} z &= \frac{1.143 - .988}{\sqrt{.355^2 + .361^2}} \\ &= \frac{.155}{\sqrt{.256}} \\ &= .306 \end{aligned}$$

You can look this *z* value up in Excel (using the function NORM.S.DIST, for standard normal distribution). You will need to subtract the value returned (.620) from 1.0, for a probability of .38. The two regression coefficients are not statistically significantly different. Once more, note the difference in orientation between these comparisons. The first compared a coefficient to a specific number, taking that number as a given. It asked if the current estimate of the regression coefficient is different from a specific value. The second asks whether two regression coefficients are statistically significantly different.

Cautions

Having gone through this illustration, we might be tempted to compare our results from this chapter with those from Chapter 1 using simple regression. I would not make this comparison because the two analyses used different dependent variables. In Chapter 1, our

conclusion was that each additional hour of (math) Homework led to a 2- point increase in math achievement test scores. In this chapter, our conclusion was that each additional hour of Homework led to a 1-point increase in Grades. Although Grades and test scores are certainly related, they are not the same thing; a 1-point increase in Grades is not the same as a 1-point increase in test scores.

In this example, I would be tempted instead to make a qualitative, rather than statistical, interpretation based on the standardized coefficients (β 's), despite our rules of thumb. On the one hand, the two values are from separate regressions with different samples. On the other hand, at least with standardized coefficients, we have a chance of interpreting the same scale (standard deviation units). The β for Homework from Chapter 1 was .320; here it is .266. These values do not seem that different, so maybe the results are consistent after all.

DIRECT CALCULATION OF β AND R^2

So far we have shown how to convert b to β and the reverse, but how could you calculate these values directly? We will focus on the direct calculation of β because it is instructive, and because it will be useful later in the book. It is fairly easy to calculate β with only two independent variables:

$$\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \text{ and } \beta_2 = \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}$$

Let's apply this formula to the 8th-grade Homework example:

$$\beta_{\text{hwork}} = \frac{r_{\text{grades-hwork}} - r_{\text{grades-pared}}r_{\text{pared-hwork}}}{1 - r_{\text{pared-hwork}}^2}$$

Note that the β of Homework on Grades depends, in part, on the simple correlation between Homework and Grades. But it also depends on the correlation between Parent Education and Grades and the correlation between Homework and Parent Education. Calculate the β :

$$\begin{aligned}\beta_{\text{hwork}} &= \frac{.327 - .294 \times .277}{1 - .277^2} \\ &= \frac{.246}{.923} \\ &= .267\end{aligned}$$

which is, within errors of rounding, the same as the value calculated by SPSS (.266). From β we can calculate b :

$$\begin{aligned}b &= \beta \frac{SD_y}{SD_x} \\ &= .266 \frac{7.623}{2.055} \\ &= .987\end{aligned}$$

which again is equivalent to the value from SPSS. The important thing to keep in mind is that the value for each β (and b) depends not only on the correlation between the independent and dependent variable but also on all the *other correlations* among the variables in the model. This

is why one way to interpret the regression coefficients is a statement like this: Homework had a strong effect on Grades, even when parents' level of education was controlled. The regression coefficients take the other variables into account; thus, don't be tempted to interpret them as if they were correlations. At the same time, note that it would be computationally challenging to calculate regression coefficients in this manner with a half-dozen or so variables!

Likewise, it is worth noting various formulas for calculating R^2 . To calculate R^2 from the sums of squares,

$$R^2 = \text{ss}_{\text{regression}} / \text{ss}_{\text{total}}$$

$$\text{To calculate } R^2 \text{ using } \beta\text{'s, } R_{y12}^2 = \beta_1 r_{y1} + \beta_2 r_{y2}.$$

To calculate R^2 from the correlations,

$$R_{Y12}^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}.$$

Note, as we discovered at the beginning of this chapter, R^2 is not equal to the sum of the two r^2 's; instead, it is reduced by a certain extent. Simply note for now that this reduction is related to the correlation between the two independent variables, r_{12} .

SUMMARY

This chapter introduced *multiple* regression, with two independent variables and one dependent variable. We conducted a regression designed to determine the effect of time spent on homework on grade-point average, controlling for parents' level of education. The regression equation was statistically significant. Unlike simple regression, with multiple regression it is possible for the overall regression to be statistically significant but to have some independent variables be nonsignificant. Here, however, the regression coefficients showed that each variable—Parent Education and time spent on Homework—had an effect on students' GPAs. We interpreted the findings from a variety of orientations. Because all the variables in the equation used a meaningful scale, we focused our interpretation primarily on the unstandardized regression coefficients.

We examined how to calculate many of the important statistics in multiple regression for this simple example: β , b , and R^2 . We discussed the pros and cons of interpreting standardized versus unstandardized regression coefficients. Both standardized and unstandardized coefficients are useful, but they serve different purposes. Unstandardized coefficients (b) are most useful when the variables are measured in a meaningful metric (e.g., hours of homework), when we wish to compare effects across studies, and when we are interested in developing policy or intervention implications from our research. Standardized coefficients (β) are more useful when the variables are not measured in a meaningful metric or when we are interested in comparing the relative importance of different predictors in the same regression equation. Rules of thumb for the use of regression coefficients are shown in Table 2.1.

Make sure you understand completely the topics presented in this chapter, because they form the foundation for much of the rest of the book. In the next chapter, we will delve deeper into this fairly simple multiple regression example.

EXERCISES

1. Conduct the Homework analysis from this chapter yourself.
2. Conduct a similar analysis using the NELS data set. Try regressing FFUGrad (GPA in 10th Grade) on BYParEd (Parents' Highest Level of Education) and F1S36A2 (Time Spent on Homework out of School). Be sure to check descriptive statistics. Notice the scales for the independent variables. The dependent variable is the average of respondents'

Grades in English, math, science, and social studies; for each of these subjects, the scale ranges from 1 = mostly below D to 8 = mostly A's.

3. Interpret the results of the regression in Exercise 2. Should you interpret the b 's or the β 's? Why would it be inappropriate to compare these results statistically with those presented in this chapter? Qualitatively, are the results similar to those presented with our simulated data?
4. The examples in this chapter suggest that students' home environments may affect their school performance. You may wonder, however, whether it is the educational environment of the home that is important or if it is the financial resources of the home that are important. The file "exercise 4, grades, ed, income.sav" has simulated data that will allow a test of this question. (The data are on the Web site [www.tzkeith.com] under Chapter 2. Also included are Excel and plain text versions of the data.) Included are measures of grade-point average (Grades, a standard 100-point scale), parents' highest level of education (ParEd, in years), and family income (Income, in thousands of dollars). Regress Grades on Parent Education and Family Income. Be sure to also check the summary statistics. Is the overall regression statistically significant? Are both variables—Parent Education and Family Income—statistically significant predictors of students' Grades? Interpret the results of this regression. Interpret both the unstandardized and the standardized regression coefficients. Which interpretation is more meaningful, the b 's or the β 's? Why? Which home variable appears to be more important for students' school performance?

Notes

- 1 A reader felt uncomfortable with this individual interpretation of regression results, especially given the smallish R^2 . Yet I think that translation of research results to the individual level is often among the most useful things we can do with them. The results of this hypothetical research were both meaningful and statistically significant and therefore (in my opinion) ripe for interpretation. Keep in mind, however, that not everyone feels comfortable on this point.
- 2 As you progress through the chapter, you may be tempted to use the b 's and their standard errors for such a comparison. These are good instincts, but it will not work, because the b 's are in different metrics. Some programs will produce standard errors of the β 's that could be useful.
- 3 Although β is the most common metric for comparing the relative influence of the variables in a regression equation, it is not the only possible metric, nor is it without its problems (especially when the independent variables are highly correlated). Darlington and Hayes (2017), for example, argued for the use of the semipartial correlations, rather than β , as measures of the relative importance of the independent variables. Yet β works well for most analyses, and it fulfills this role better than do other statistics that are commonly produced by statistics programs. We will continue to focus on β as providing information about the relative influence of different variables, at least for the time being. As already noted, partial and semipartial correlations are discussed in Appendix C.

3

Multiple Regression

More Depth

Why $R^2 \neq r^2 + r^2$	44
Predicted Scores and Residuals	47
Regression Line	50
Least Squares	52
Regression Equation = Creating a Composite?	54
Assumptions of Regression and Regression Diagnostics	54
Summary	55
Exercises	55
Note	56

In this chapter we will delve into a little more depth about multiple regression and explain some concepts a little more fully. This chapter probably includes more formulas than most, but I will try to explain concepts several different ways to ensure that at least one explanation makes sense to every reader. The chapter is short, but if math and statistics do not come easily to you, you may need to read this chapter more than once. Your perseverance will pay off with understanding!

WHY $R^2 \neq r^2 + r^2$

I noted in the last chapter that, as a general rule, R^2 is not equal to $r^2 + r^2$ (in a two-variable multiple regression) and briefly mentioned that this was due to the correlation between the independent variables. Let's explore this phenomenon in more detail. To review, in the example used in the beginning of Chapter 2, $r^2_{HWork\cdot Grades} = .327^2 = .107$, and $r^2_{ParEd\cdot Grades} = .294^2 = .086$.

The R^2 from the regression of GPA on Homework and Parent Education was .152. Obviously, $.152 \neq .107 + .086$. Why not? We'll approach this question several different ways. First, recall one of the formulas for R^2 :

$$R^2_{y_{12}} = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}$$

Note that the squared multiple correlation depends not only on the correlation between each independent variable and the dependent variable but also on the correlation between the two independent variables, r_{12} , or, in this case $r_{HWork\cdot ParEd}$, a value of .277.

Next, look at Figure 3.1. The circles in the figure represent the variance of each variable in this regression analysis, and the areas where the circles overlap represent the shared variances, or the r^2 's, among the three variables. The shaded area marked 1 (including the area marked 3) represents the variance shared by Grades and Homework, and the shaded area marked 2 (including the area marked 3) represents the variance shared by Parent Education and Grades. Note, however, that these areas of overlap also overlap each other in the doubly shaded area marked 3. This overlap occurs because Homework and Parent Education are themselves correlated. The combined area of overlap between Homework and Grades and between Parent Education and Grades (areas 1 and 2, including 3) represents the variance of Grades *jointly accounted for* by Homework and Parent Education, or the R^2 . As a result of the joint overlap (3), however, the total area of overlap is not equal to the sum of areas 1 and 2; area 3 is counted once, not twice. In other words, R^2 is not equal to $r^2 + r^2$.

Using this logic, it follows that *if* the correlation between the two independent variables is zero then R^2 will equal $r^2 + r^2$. Such a situation is depicted in Figure 3.2, where the area of

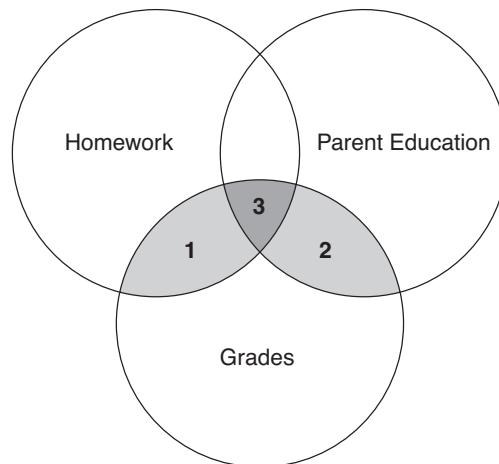


Figure 3.1 Venn diagram illustrating the shared variance (covariance) among three variables. The shaded areas show the variance shared by each independent variable with the dependent variable. Area 3 shows the variance shared by all three variables.

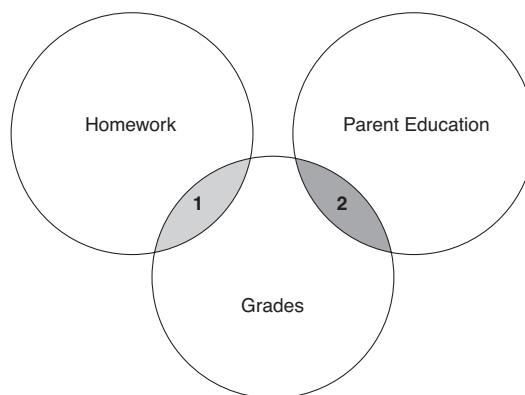


Figure 3.2 Venn diagram illustrating the shared variance among three variables. In this example, there is no correlation (and no shared variance) between the two independent variables.

46 • MULTIPLE REGRESSION

overlap is indeed equal to the sum of areas 1 and 2, because the two independent variables do not themselves overlap. Likewise, turning to the formula for R^2 , you can see what happens when r_{12} is equal to zero. The formula is

$$R_{y_{12}}^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}$$

When zero is substituted for r_{12} ,

$$R_{y_{12}}^2 = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2} \times 0}{1 - 0}$$

the formula reduces to $R_{y_{12}}^2 = r_{y1}^2 + r_{y2}^2$.

Let's double-check. Figure 3.3 shows the results of the regression of Grades on Homework and Parent Education in the (unlikely) event that the correlation between Homework and Parent Education is zero (the data are simulated). Note that the correlations between Parent

Correlations

		GRADES	PARED	HWORK
Pearson Correlation	GRADES	1.000	.294	.327
	PARED	.294	1.000	.000
	HWORK	.327	.000	1.000
Sig. (1-tailed)	GRADES	.	.001	.000
	PARED	.001	.	.500
	HWORK	.000	.500	.
N	GRADES	100	100	100
	PARED	100	100	100
	HWORK	100	100	100

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.440 ^a	.193	.177	6.916656

a. Predictors: (Constant), HWORK, PARED

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	58.008	5.382		
	PARED	1.161	.360	.294	.002
	HWORK	1.213	.338	.327	.001

a. Dependent Variable: GRADES

Figure 3.3 Multiple regression results when there is no correlation between the two independent variables.

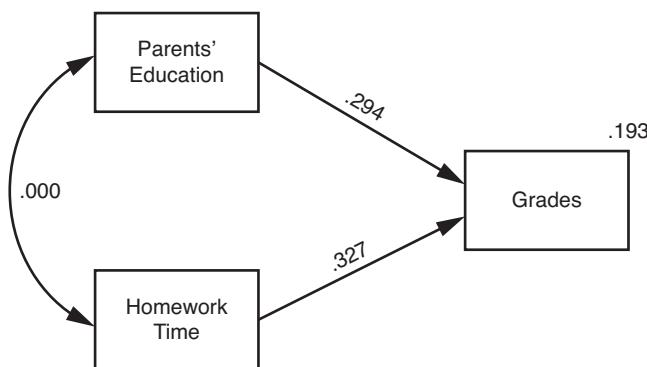


Figure 3.4 Path representation of the effects of Parents' Education and Homework on Grades when there is a correlation of zero between the two predictors. Note that the standardized regression coefficients are the same as their correlations with Grades.

Education and Grades (.294) and between Homework and Grades (.327) are the same as in Chapter 2, but that the correlation between Parent Education and Homework is now zero. And consistent with our reasoning above, R^2 now equals $r^2 + r^2$.

$$\begin{aligned}
 R^2_{\text{Grades-HWork-ParEd}} &= r^2_{\text{Grades-ParEd}} + r^2_{\text{Grades-HWork}} \\
 .193 &= .294^2 + .327^2 \\
 .193 &= .193
 \end{aligned}$$

Also note that when the independent variables are uncorrelated, the β 's are again equal to the correlations (as with simple regression). The reason why is, of course, that the formula for β ,

$$\left(\beta_1 = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \right)$$

reduces to $\beta_1 = r_{y1}$ when $r_{12} = 0$.

Figure 3.4 shows this regression in path format. Notice the correlation of zero between the two independent variables. With this lack of relation between the two variables, the standardized coefficients are the same as the correlations and the $R^2 = r^2 + r^2$.

To reiterate, the R^2 depends not only on the correlations of the independent variable with the dependent variable, but also on the correlations among the independent variables. As a general rule, the R^2 will be less than the sum of the squared correlations of the independent variables with the dependent variable.¹ The only time R^2 will equal $r^2 + r^2$ is when the independent variables are uncorrelated, and this happens rarely in the real world.

PREDICTED SCORES AND RESIDUALS

It is worth spending some time examining more detailed aspects of multiple regression, such as the residuals and the predicted scores. Understanding these aspects of regression will help you more completely understand what is going on in multiple regression and also provide a good foundation for topics that we will cover later.

Among other things, residuals (the error term from the regression equation) are useful for diagnosing problems in regression, such as the existence of outliers, or extreme values. We will address the use of residuals for diagnostic purposes in Chapter 10.

In Chapter 2 we saw how to use the regression equation to predict an individual's score on the outcome. Simply plug a person's values for the two independent variables (i.e., Parent Education and Homework time) into the regression equation and you get the person's predicted grade-point average. So, using the first regression equation from the previous chapter, an 8th-grade student who reports 5 hours of homework per week, and with a parent education level of 16 (four years of college), would have a predicted GPA of 82.103. We also may be interested in the predicted outcomes for *everyone* in our data set. In this case, it is simple to have our statistics program calculate the predicted scores as a part of the multiple regression analysis. In SPSS, for example, simply click on the Save button in multiple regression and highlight Predicted Values; Unstandardized (see Figure 3.5). While we're at it, we'll also ask

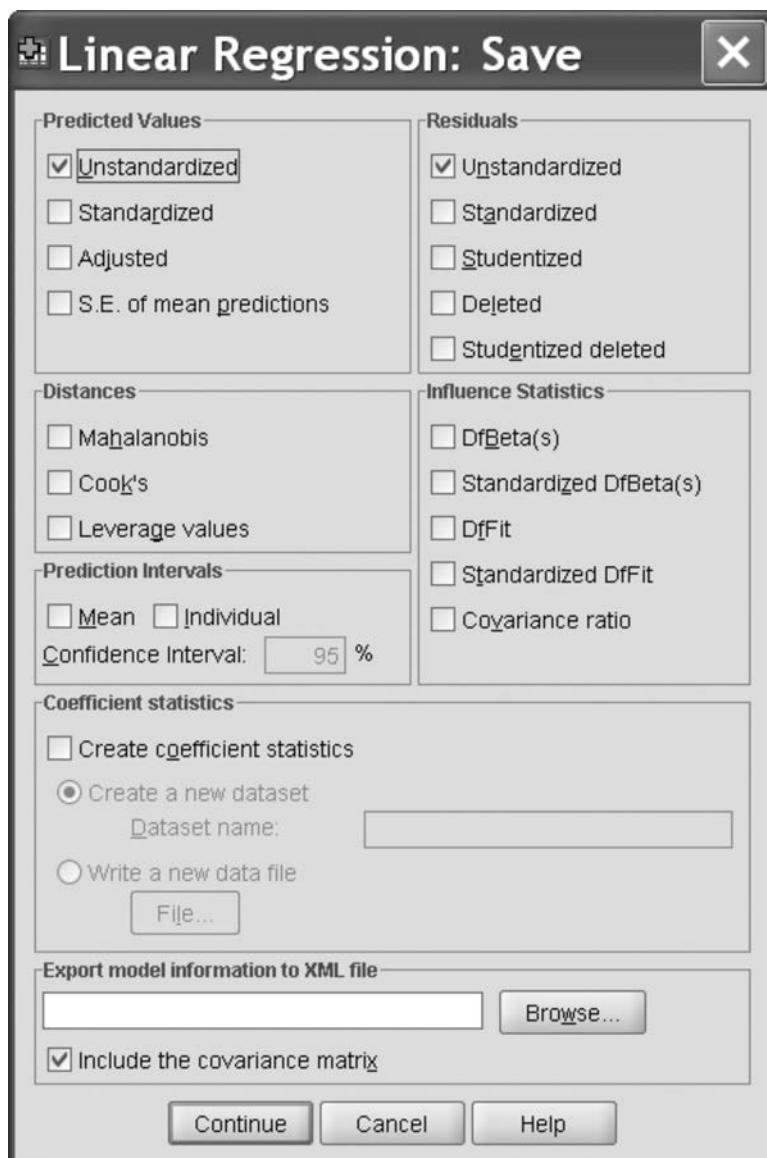


Figure 3.5 Generating predicted values and residuals in SPSS.

for the unstandardized residuals. In SAS, you can get predicted values and residuals using an OUTPUT statement.

I again regressed Grades on Parent Education and Homework using the 8th-grade data from Chapter 2, but this time saved the predicted scores and residuals. Figure 3.6 shows Grades (first column) and the Predicted Grades (PredGrad) for the first 34 cases of our Homework & Grades data from Chapter 2. Note that for some students we predict higher grades based on the regression equation than they actually earned, whereas for other students their actual grades were higher than their predicted grades. Obviously, the prediction is not exact; in other words, there is error in our prediction.

The third column in this figure shows the residuals from this regression (Resid_1). What are the residuals? Conceptually, the residuals are what is left over or unexplained by the regression equation. They are the errors in prediction that we noticed when comparing the actual versus predicted Grades. Remember one form of the regression equation (with two

GRADES	PREDGRAD	RESID_1	ERROR_1
78.00	76.52082	1.47918	1.47918
79.00	81.34282	-2.34282	-2.34282
79.00	75.53297	3.46703	3.46703
89.00	79.48435	9.51565	9.51565
82.00	80.12053	1.87947	1.87947
77.00	78.49651	-1.49651	-1.49651
88.00	79.48435	8.51565	8.51565
70.00	77.50866	-7.50866	-7.50866
86.00	81.22560	4.77440	4.77440
80.00	80.35498	-.35498	-.35498
76.00	78.14484	-2.14484	-2.14484
72.00	79.48435	-7.48435	-7.48435
66.00	76.63804	-10.63804	-10.63804
79.00	79.36713	-.36713	-.36713
76.00	75.88464	.11536	.11536
80.00	86.56656	-6.56656	-6.56656
91.00	84.18914	6.81086	6.81086
85.00	83.08407	1.91593	1.91593
79.00	82.44789	-3.44789	-3.44789
82.00	78.37928	3.62072	3.62072
94.00	81.57727	12.42273	12.42273
91.00	79.60157	11.39843	11.39843
80.00	80.35498	-.35498	-.35498
73.00	82.33067	-9.33067	-9.33067
77.00	78.61373	-1.61373	-1.61373
76.00	82.09622	-6.09622	-6.09622
84.00	76.63804	7.36196	7.36196
81.00	82.09622	-1.09622	-1.09622
97.00	87.03545	9.96455	9.96455
80.00	82.21344	-2.21344	-2.21344
74.00	82.09622	-8.09622	-8.09622
83.00	87.15267	-4.15267	-4.15267
78.00	80.47220	-2.47220	-2.47220
64.00	84.94254	-20.94254	-20.94254

Figure 3.6 Partial listing comparing Grades (Y), Predicted Grades (Y'), and the residuals as output by the computer program, and the error term as computed by subtraction (Y-Y').

independent variables): $Y = a + bX_1 + bX_2 + e$. In this equation, the residuals are equal to e , the error term from the regression.

Remember also the other form of the regression equation, using the *predicted* scores on Y (symbolized as Y'), in this case the predicted grades: $Y' = a + bX_1 + bX_2$. We can subtract this formula from the first formula to figure out how to solve for e , the residuals:

$$\begin{aligned} Y &= a + bX_1 + bX_2 + e \\ -Y' &= a + bX_1 + bX_2 \\ \hline Y - Y' &= e \end{aligned}$$

Thus, in the present example, the residuals are simply the predicted grades subtracted from the actual grades. The final column in Figure 3.6 (Error_1) shows the results of $Y - Y'$, in which I simply subtracted the predicted grades from actual grades. Notice that this error term is identical to the residuals (RESID_1). The residuals are what are left over after the predicted outcome variable is removed from the actual outcome variable; they are the inaccuracies, or errors of prediction. Another way of thinking of the residuals is that they are equivalent to the original dependent variable (Grades) with the effects of the independent variables (Parent Education and Homework) removed. This way of thinking about the residuals will come in handy later.

REGRESSION LINE

As with simple regression, we can also understand the predicted scores and residuals using the regression line. With simple regression, we can find the predicted scores using the regression line: find the value of the independent variable on the X -axis, go straight up to the regression line, and then find the value of the dependent variable (Y -axis) that corresponds to that point on the regression line. The regression line, with simple regression, is simply a line connecting the *predicted* Y 's for each value of X . With multiple regression, however, there are multiple regression lines (one for each independent variable). But wait: if the regression line is equivalent to the predicted scores, then the predicted scores are equivalent to the regression line. In other words, with multiple regression, we can, in essence, get an overall, single regression line by plotting the predicted scores (X -axis) against the actual scores (Y -axis). This has been done in Figure 3.7, which includes both the regression line of the plot of predicted versus actual GPA, and each data point. Why are predicted scores on the X axis? Because another way of thinking about them is that they are a linear combination of the X variables (Parents' Education and Homework).

First note that the r^2 (.152, shown in the lower right of Figure 3.7) from the regression of Grades on Predicted Grades (with Grades predicted by Homework and Parent Education) is identical to the R^2 from the multiple regression of Grades on Homework and Parent Education (.1521), further evidence that this can be thought of as the overall regression line for our multiple regression of Grades on Homework and Parent Education. This finding also points to another way of thinking about R^2 : as the squared correlation between Y and predicted Y (Y').

If the line represents the predicted Grades, then the deviation of each actual Grade (each data point) from the line represents what? The residuals—if you subtract the value (on the Y -axis) of the *regression line* from the value of each *data point* (on the Y -axis), you will find the same values for the residuals as shown in Figure 3.6. Again, this is simply $Y - Y'$. You can see this most easily by focusing on the data point in the lower-right corner of the graph, defined as $X = 84.94$ and $Y = 64$. This is also the final data point in Figure 3.6. Follow the line from this point to the regression line and then over to the Y -axis. The value on the Y -axis is also 84.94. The residual is thus $64 - 84.92 = -20.94$, also the same value shown for the residual in Figure 3.6. (Here's an extra-credit question for you: since in this figure every point on the regression line has the same values for both the X - and the Y -axes, what is the value for b ?

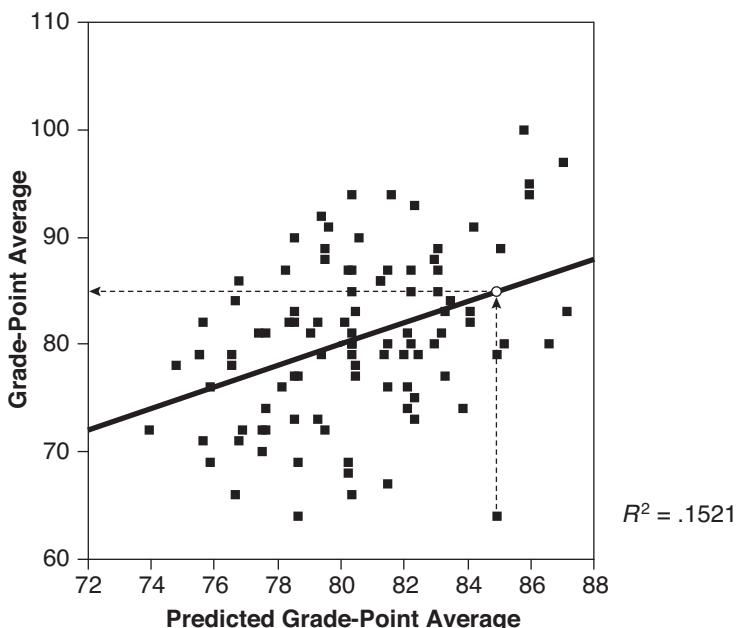


Figure 3.7 Plot, with regression line, of Grades (Y) versus Predicted Grades (Y').

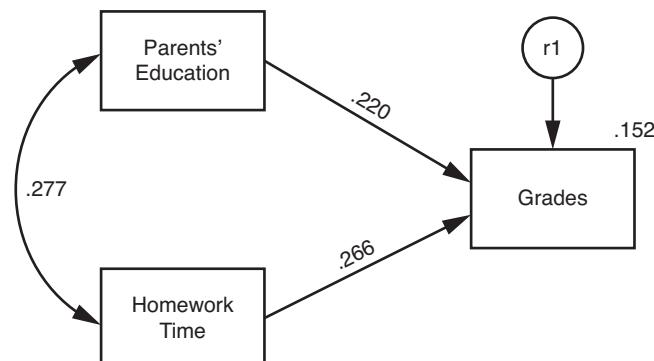


Figure 3.8 Figural (path) display of residuals. The variable $r1$ is in a circle rather than a rectangle to show that it is an unmeasured variable. Such variables are explored in depth in Part 2.

Remember that b is the slope of the regression line.) It is also possible to depict the residuals in a path display of regression results, as is done in Figure 3.8. There, the small circle labeled $r1$ represents the residual. A circle is used rather than a rectangle to indicate that we don't have actual measures of the residuals in our original data set. We could (and just have) generated estimates of the residuals, but these are a product of our regression, not a part of our original data. In the framework of path models (Part 2), this is an “unmeasured” variable, and you can think of it as all other influences on Grades other than the two variables (Parent Education and Homework) shown in the model. This depiction will come in handy when you learn about partial and semipartial correlations (Appendix C).

LEAST SQUARES

Recall that I said in Chapter 2 that the two independent variables were “optimally weighted.” What does this mean? Why not just weight each of the two variables by $\frac{1}{2}$ to predict GPA; in other words, why not just standardize the two independent variables, average them, and use that composite to predict GPA using simple regression? Or why not weight them in some other logical combination? The reason is that the prediction will not be as good, as accurate. The explained variance (R^2) will not be as high, and the unexplained variance ($1 - R^2$) will be higher. Another way of saying this is to state that the regression line shown in Figure 3.7 is the best fitting of all possible (straight) lines that could be drawn through these data points.

So what does *best fitting* mean? Again, it means the line that minimizes the error of prediction, or the unexplained variance. Take a look at the line again. Suppose you were to measure the distance from each data point to the regression line and subtract from it the corresponding point from the regression line. This is what we just did for a single data point (84.94, 64), and we found that these are the same as the residuals. These are the errors in prediction. If you were to sum these values, you would find that they summed to zero; the positive values will be balanced by negative values. To get rid of the negative values, you can square each residual and then sum them. If you do this, you will find that the resulting number is smaller than for *any other possible straight line*. This best fitting line thus minimizes the errors of prediction; it minimizes the squared residuals. You will sometimes hear simple or multiple regression referred to as *least squares regression* or OLS (ordinary least squares) regression. The reason is that the regression weights the independent variables so as to *minimize the squared residuals*, thus *least squares*.

Figure 3.9 displays descriptive statistics for some of the variables we have been discussing: Grades (Y), Predicted Grades (Y'), and the Residuals. Also shown are descriptive statistics for the *squared* Residuals (ResidSq). Note that the means and sums for Grades and Predicted Grades are the same. The Predicted Grades have a narrower range (73.91 to 87.15) than do the actual Grades (64 to 100) and a smaller variance (8.84 compared to 58.11), which should be obvious when looking at the figure that shows the regression line; notice the scale for the X axis versus the Y axis in Figure 3.7. The Y-axis on the figure has a much wider range than does the X-axis. Note that the sum of the residuals is zero in Figure 3.9. Note also the sum of the squared residuals: 4878.17. As mentioned in the previous paragraph, the regression line minimizes this number; any other possible straight line will result in a larger value for the sum of the squared residuals. If you turn back to Chapter 2, you can compare this number to the Sum of Squares for the residual in Figure 2.4; they are the same (4878.17). The residual sums of squares is just that: the sum of the squared residuals.

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Variance
GRADES Grade Point Average	100	64.00	100.00	8047.00	80.4700	58.110
PREDGRAD Unstandardized Predicted Value	100	73.90895	87.15267	8047.000	80.47000	8.836
RESID_1 Unstandardized Residual	100	-20.94254	14.18684	.00000	-8.8E-15	49.274
RESIDSQ Valid N (listwise)	100	.01	438.59	4878.17	48.7817	4431.695

Figure 3.9 Descriptive statistics for Grades, Predicted Grades, the residuals, and the squared residuals.

I argued that the independent variables are weighted so that this sum of squared residuals is minimized and the R^2 is maximized. In our current example, Parent Education was weighted .220 (the standardized regression coefficient), and Homework was weighted by .266, close to a 50/50 ratio. What would happen if we chose a different weighting? Perhaps, for some reason, you believe that Parent Education is not nearly as important as Homework for explaining Grades. Therefore, you decide to weight Parent Education by .25 versus .75 for Homework when predicting Grades. This solution may be satisfying in other ways, but the resulting prediction is not as accurate and is more error laden. Using the least squares solution of multiple regression in Chapter 2, we explained 15.2% of the variance in Grades ($R^2 = .152$). If, however, you regress Grades on a composite that weighted Parent Education by .25 and Homework by .75, our logically determined solution, you will find that this solution explains slightly less variance in Grades: 14% (see Figure 3.10). As noted previously, the error variance (residual sum of squares) was 4878.171 using the least squares solution. In contrast, using this 25/75 solution, the sum of squared residuals is larger: 4949.272 (Figure 3.10). The least squares, multiple regression, solution minimized the residual, or error, sums of squares and maximized the R^2 , or the variance in Grades explained by Parent Education and Homework. As a result (and given the adherence to necessary assumptions), the estimates produced by the least squares solution will be the best possible estimates and the least biased (meaning the most likely to reproduce the population values).

Perhaps it is obvious that the variability of points around the regression line is closely related to the accuracy in prediction. The closer the data points in Figure 3.7 cluster around the regression line, the less error involved in prediction. In addition, the closer the data points are to the regression line, the more likely the regression is to be statistically significant, because this lowered variability will reduce the variation in residuals. The value of F depends, in part, on the variability in the residuals:

$$F = \frac{SS_{\text{regression}} / df_{\text{regression}}}{SS_{\text{residual}} / df_{\text{residual}}}$$

In addition, the variability in the residuals is related to the standard error of the regression coefficient (se_b), which is used to calculate the statistical significance of b ($t = b/se_b$).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.374	.140	.131	7.1065

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	803.638	1	803.638	15.913	.000
	Residual	4949.272	98	50.503		
	Total	5752.910	99			

b. Dependent Variable: GRADES Grade Point Average

Figure 3.10 Regression results with Parent Education weighted at 25% and Homework weighted at 75%. Note that the R^2 decreases, and the sum of squared residuals increases.

REGRESSION EQUATION = CREATING A COMPOSITE?

These last few sections hint at a different way of conceptualizing what happens in multiple regression. We saw under Predicted Scores and Residuals that we could create a single score, the predicted dependent variable (in this case, predicted Grades), that functions the same way in a simple regression analysis as do the multiple independent variables in a multiple regression analysis. We saw, for example, that the R^2 from multiple regression is the same as r^2 between Y and \hat{Y} . We hinted in the section Least Squares that we could also create such a single independent variable by weighting the multiple independent variables. Is it therefore possible to use this weighting method to create a single independent variable that matches the predicted score?

The answer is yes. Instead of weighting the standardized Homework and Parent Education variables by .75 and .25, we could have weighted them by β 's from the multiple regression equation (.266 and .220) to create a composite. Even more directly, we could create a composite using the unstandardized values of Homework and Parent Education, weighting each according to its b from the multiple regression in Chapter 2 (.988 and .871, respectively). Either of these approaches would have created a composite of Homework and Parent Education that predicted Grades just as well as did our predicted Grades variable and just as well as did the original Homework and Parent Education variables.

I am not suggesting that you do this in practice; the multiple regression does it for you. Instead, you should understand that this is one way of thinking about how multiple regression works: MR provides an optimally weighted composite, a synthetic variable, of the independent variables and regresses the dependent variable on this single composite variable. This realization will stand you in good stead as you ponder the similarities between multiple regression and other statistical methods. In fact, this is what virtually all our statistical methods do, from ANOVA to structural equation modeling. “All statistical analyses of scores on measured/observed variables actually focus on correlational analyses of scores on synthetic/latent variables derived by applying weights to the observed variables” (Thompson, 1999, p. 5). Thompson goes on to note—tongue in cheek—what you may have long suspected, that we simply give these weights different names (e.g., factor loadings, regression coefficients) in different analyses so as to confuse graduate students.

ASSUMPTIONS OF REGRESSION AND REGRESSION DIAGNOSTICS

Given the conceptual nature of this book, I have just touched on the issues of residuals and least squares regression. Analysis of residuals is also a useful method for detecting violations of the assumptions underlying multiple regression and outliers and other problems with data. I want to postpone discussion of the assumptions underlying regression until you have a deeper understanding of how to develop, analyze, and interpret regression analyses. These assumptions are presented in Chapter 10 as an important topic and one worthy of additional study. Likewise, we will postpone discussion of the regression diagnostics until that time, along with diagnosis of other potential problems in regression (e.g., multicollinearity).

It is worth noting that residuals have other uses, as well. Suppose, for example, that you were studying student performance on a test across various age levels, but wanted to remove the effects of age from consideration in these analyses. One possible solution would be to regress the test scores on age and use the residuals as age-corrected test scores (e.g., Keith, Kranzler, & Flanagan, 2001). Darlington and Hayes (2017) discussed using residuals for other research purposes. In Appendix C we will use the residuals to understand partial and semipartial correlations.

SUMMARY

This chapter has focused on some of the nitty-gritty of multiple regression analysis, including the nature of R^2 compared to r^2 , the conceptual and statistical meaning of predicted scores and residuals, and the method by which multiple regression produces the “optimal” prediction. We found that R^2 depends not only on the original correlation of each independent variable with the dependent variable but also on the correlations of the independent variables with each other. As a result, R^2 is usually less than the sum of the r^2 ’s and only equals the sum of the r^2 ’s when the independent variables are themselves uncorrelated. Likewise, the β ’s are not equivalent to and are usually smaller than the original r ’s. Only when the correlations among the independent variables are zero do the β ’s equal the r ’s.

Residuals are the errors in prediction of a regression equation and the result of subtracting the predicted scores on the dependent variable (predicted via the regression equation) from the actual values of participants on the dependent variable. Multiple regression works to minimize these errors of prediction so that the residual sums of squares, the sum of the squared residuals, is the smallest possible number. For this reason, you will sometimes see regression referred to as least squares regression. One way of thinking about multiple regression is that it is creating a synthetic variable that is an optimally weighted composite of the individual independent variables. This composite, weighting each independent variable by its regression weight, is then used to predict the outcome variable.

Do not be overly worried if all the concepts presented in this chapter are not crystal clear; opaque will do for now! I do encourage you, however, to return to this chapter periodically as you become more familiar and fluent in multiple regression; I believe this chapter will make more sense each time you read it and will also deepen your understanding of other topics.

EXERCISES

1. Use the Grades, Parent Education, and Homework example from Chapter 2; make sure you can reproduce the residual analyses from this chapter (i.e., those summarized in Figures 3.6, 3.7, and 3.9). Output residuals and predicted scores, and examine their descriptive statistics and correlations with Grades and each other. Make sure you understand why you obtain the relations you find.
2. Create a composite variable weighting Parent Education and Homework by their regression weights as found in Exercise 1. Regress Grades on this composite. Note that you can weight the original variables using the unstandardized regression weights, or you can first standardize Parent Education and Homework (convert them to z-scores) and then weight them by the appropriate β ’s. How do the R^2 and sums of squares compare to the multiple regression results?
3. Now try creating a composite that weights the Parent Education and Homework by some other values (e.g., 25% and 75%). Note that to do this you will need to standardize the variables first. What happens to the R^2 and sum of squared residuals?
4. Reanalyze the regression of Grades on Parent Education and Family Income from Chapter 2 (Exercise 4). Output the unstandardized predicted values and residuals. Compute the correlation between Grades and Predicted Grades. Is the value the same as the R from the multiple regression? Explain why it should be. Create a scatterplot of Predicted Grades with Grades, along with a regression line. Pick a data point in the raw data and note the actual value for Grades, Predicted Grades, and the Residual. Is the residual equal to Grades minus Predicted Grades? Now find the same data point on

the scatterplot and mark the value on the graph for that person's Grades and Predicted Grades. Show graphically the residual.

Note

- 1 If the two independent variables are *negatively* correlated with each other, but correlate positively with the dependent variable, R^2 will actually be larger than $r^2 + r'^2$. The β 's will also be larger than the r 's. This phenomenon may be considered a form of what is called statistical *suppression*. Suppression is discussed in more detail in several sources (e.g., Cohen et al., 2003, chap. 3; MacKinnon, Krull, & Lockwood, 2000; Pedhazur, 1997, chap. 7; Thompson, 2006, chap. 8).

4

Three and More Independent Variables and Related Issues

Three Predictor Variables	57
<i>Regression Results</i>	60
<i>Interpretation</i>	61
Rules of Thumb: Magnitude of Effects	62
Testing the Difference Between Two Regression Coefficients	63
Four Independent Variables	64
<i>Another Control Variable</i>	65
<i>Regression Results</i>	65
<i>Trouble in Paradise</i>	65
Common Causes and Indirect Effects	68
The Importance of R^2 ?	70
Prediction and Explanation	72
Summary	73
Exercises	74
<i>Notes</i>	75

This chapter will present two more examples of multiple regression, one using three independent variables and the other, four. The intent is to increase your comfort with conducting and interpreting multiple regression and to solidify the concepts presented so far. You will see that the addition of explanatory variables makes the regression no more difficult, with the exception that you have more to discuss when explaining the results. We will use these examples to confront several looming issues.

THREE PREDICTOR VARIABLES

Let's take our homework example a little further. Suppose that you become interested in the effects of homework completed in school versus the effects of homework completed out of school. I actually *did* become interested in this topic thanks to our children. When we asked them if they had homework to complete, they began to respond "I did it in school." Our response, in turn, was "that's not homework, that's schoolwork!" Beyond our little parent-child exchanges, I began to wonder if "homework" completed in school had the same effects on learning and achievement as homework completed at home. The results of this research are described in Keith, Hallam, and Fine (2004); the research

used structural equation modeling, rather than multiple regression, but MR could have been used.

At any rate, suppose that you share at least some of my interest in this topic (we will switch to other examples in subsequent chapters). We will use the NELS data to examine the relative effects of homework completed in school versus homework completed out of school on students' grades. Our research question is something along these lines: does homework completed in school have the same effect on high school students' grades as homework completed out of school? To answer the question, we regress Grades on a measure of time spent on Homework In School and a measure of time spent on Homework Out of School. As in the previous example, we control for Parents' level of Education. Grades (FFUGrad) in this example are an average of students' 10th-grade grades in English, Math, Science, and Social Studies. Parent Education (BYParEd) is the education level of the father or mother (whichever is higher) for each student. The Homework variables are students' 10th-grade reports of the amount of time they spend, on average, per week doing homework, across subjects, In School (F1S36A1) and Out of School (F1S36A2). All variables are included in your copy of the NELS data; you should examine the descriptive statistics for all these variables and should also examine the frequencies of each predictor variable. It is good practice to run the multiple regression analysis, too! It might also be worth rereading Appendix A and its discussion of the NELS data set. The figural representation of the regression model is shown in Figure 4.1. The paths represent the regression weights and the curved arrows represent the correlations among the independent variables (in multiple regression the predictor variables are correlated with one another).

Figure 4.2 shows the frequencies of the independent variables. Notice that the scales of these variables are different from those in previous chapters. In the current example, Parent Education ranges from a value of 1, representing "Did not finish High School," up to a value of 6, representing an advanced graduate degree (PhD, MD, etc.). The Homework variables are no longer hours but rather chunks of hours, ranging from 0 (No homework) to 7 (Over 15 hours per week). Figure 4.3 shows the descriptive statistics for the dependent variable, 10th-grade (First Follow-Up) Grade Average. Its scale has also changed, from the common 0 to 100 scale to a 1 to 8 scale, with 1 representing low grades and 8 representing high grades. The NELS developers had justifiable reasons for scaling these variables in this manner, but the variables no longer have the nice logical scales (e.g., years or hours) from the previous examples. This deficiency is more than made up for, in my opinion, by the fact that these are real and nationally representative data, whereas the previous examples had used simulated data. Figure 4.4 shows the correlations among these variables.

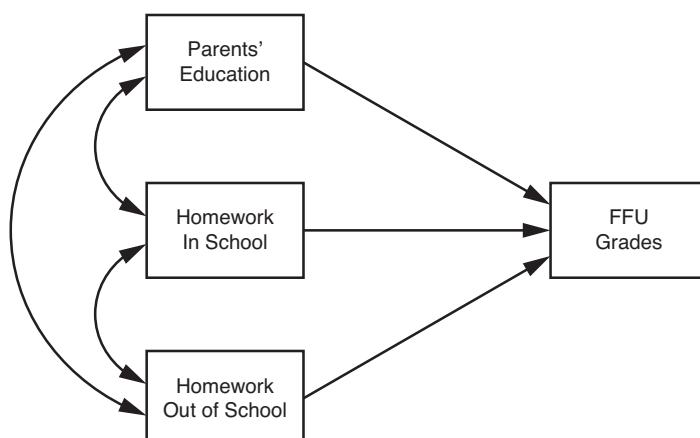


Figure 4.1 Multiple regression example in path format.

BYPARED PARENTS' HIGHEST EDUCATION LEVEL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 did not finish HS	97	9.7	9.7	9.7
	2 HS Grad or GED	181	18.1	18.1	27.8
	3 lt 4 year degree	404	40.4	40.4	68.3
	4 college grad	168	16.8	16.8	85.1
	5 M.A. or equiv.	86	8.6	8.6	93.7
	6 PhD., M.D. or other	63	6.3	6.3	
	Total	999	99.9	100.0	
Missing	8 missing	1	.1		
Total		1000	100.0		

F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 NONE	63	6.3	6.7	6.7
	1 1 HOUR OR LESS	232	23.2	24.6	31.3
	2 2-3 HOURS	264	26.4	28.0	59.3
	3 4-6 HOURS	168	16.8	17.8	77.1
	4 7-9 HOURS	80	8.0	8.5	85.6
	5 10-12 HOURS	66	6.6	7.0	92.6
	6 13-15 HOURS	31	3.1	3.3	95.9
	7 OVER 15 HOURS	39	3.9	4.1	100.0
Total		943	94.3	100.0	
Missing	96 MULTIPLE RESPONSE	7	.7		
	98 MISSING	17	1.7		
	System	33	3.3		
	Total	57	5.7		
Total		1000	100.0		

F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 NONE	76	7.6	8.1	8.1
	1 1 HOUR OR LESS	341	34.1	36.5	44.6
	2 2-3 HOURS	242	24.2	25.9	70.5
	3 4-6 HOURS	158	15.8	16.9	87.4
	4 7-9 HOURS	42	4.2	4.5	91.9
	5 10-12 HOURS	37	3.7	4.0	95.8
	6 13-15 HOURS	14	1.4	1.5	97.3
	7 OVER 15 HOURS	25	2.5	2.7	100.0
Total		935	93.5	100.0	
Missing	96 MULTIPLE RESPONSE	9	.9		
	98 MISSING	23	2.3		
	System	33	3.3		
	Total	65	6.5		
Total		1000	100.0		

Figure 4.2 Frequencies for the independent variables in the three-predictor MR example.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Deviation Std.	Variance
FFUGRAD ffu grades	950	1.00	8.00	5.6661	1.4713	2.165
Valid N (listwise)	950					

Figure 4.3 Descriptive statistics for the dependent variable Grades.**Correlations^a**

		FFUGRAD ffu grades	F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL	F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL	BYPARED PARENTS' HIGHEST EDUCATION LEVEL
FFUGRAD ffu grades	Pearson Correlation Sig. (2-tailed)	1 .004	.096 1	.323 .000	.304 .000
F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL	Pearson Correlation Sig. (2-tailed)	.096 .004	1 .000	.275 .000	.059 .075
F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL	Pearson Correlation Sig. (2-tailed)	.323 .000	.275 .000	1 .000	.271 .000
BYPARED PARENTS' HIGHEST EDUCATION LEVEL	Pearson Correlation Sig. (2-tailed)	.304 .000	.059 .075	.271 .000	1 .000

a. Listwise N=909

Figure 4.4 Correlations among the independent and dependent variables.

Regression Results

Figure 4.5 shows some of the results of the regression analysis. As you can see, the three variables, Parent Education, time spent on Homework In School, and time spent on Homework Out of School explained 15.5% of the variance in students' 10th-grade GPA ($R^2 = .155$), and the overall regression equation was statistically significant ($F[3, 905] = 55.450, p < .001$). The third table in the figure, however, shows that not all the variables are important in this regression. In fact, Parent Education had a substantial and statistically significant effect on Grades ($b = .271, \beta = .234, p < .001$), as did time spent on Homework Out of School ($b = .218, \beta = .256, p < .001$). In contrast, the effect of time spent on Homework In School was tiny and was not statistically significant ($b = .012, \beta = .012, p = .704$). (When you do this regression yourself, you may get a value of 1.16E-02 in the b column. Don't panic; the coefficient is simply displayed as an exponential number; move the decimal point two places to the left, that is, .0116.) Note also the 95% confidence intervals for unstandardized coefficients. The CI for Homework In School encompasses zero; again, we cannot reject the hypothesis that the population value is zero.

The results are fairly similar to those we found using the simulated data in Chapter 2 (which were, of course, designed to mimic reality). Focusing on the β 's, we conclude that each standard deviation increase in time spent on Homework Out of School led to a .256 SD increase in GPA, with Parent Education and In-School Homework controlled. Each additional SD in Parent Education resulted in a .234 SD increase in student GPA (controlling for Homework). As noted previously, the scales of the three independent variables (Homework and Parent Education) are not particularly meaningful. Parent Education ranged from 1 (did not finish high school) to 6 (PhD, MD, or other doctoral degree). The two Homework variables had values that ranged from 0, for "None" as the average amount of time spent on Homework per week, to 7 for "over 15 hours per week." Because the scales for these variables do not follow any naturally interpretable scale, such as years for Education or hours for

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.394 ^a	.155	.152	1.3500

a. Predictors: (Constant), F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL, BYPARED PARENTS' HIGHEST EDUCATION LEVEL, F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL

b. Dependent Variable: FFUGRAD ffu grades

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	303.167	3	101.056	55.450	.000
	Residual	1649.320	905	1.822		
	Total	1952.486	908			

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	4.242	.135	31.337	.000	3.977	4.508
	BYPARED PARENTS' HIGHEST EDUCATION LEVEL	.271	.037	.234	.7375	.000	.199 .343
	F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL	.012	.031	.012	.379	.704	-.048 .072
	F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL	.218	.028	.256	7.780	.000	.163 .273

a. Dependent Variable: FFUGRAD ffu grades

Figure 4.5 Results of a multiple regression with three independent variables.

Homework, the b 's are not readily interpretable. We could say “each unit of Homework Time Out of School resulted in a .218-point increase in GPA,” but this would not tell us much, because we would then need to explain what a “unit” increase in Homework meant. Likewise, it does not help that the scale used for Grades is nontraditional, as well (it ranges from 1, “Mostly below D” to 8 “Mostly As”). When the scales of the variables are in a nonmeaningful metric, it makes more sense to interpret standardized regression coefficients, β 's, than it does to interpret unstandardized regression coefficients, or b 's. You should still report both, however, along with standard errors or confidence intervals of the b 's. Such a practice will allow comparability with other studies using similar scales. Strictly speaking, it is also the b 's that are tested for statistical significance.

Interpretation

Assuming that you trusted these findings as reflecting reality, how might you interpret them to parents, or to high school students? The important finding, the finding of primary interest, is the difference in the effects of In-School versus Out-of-School Homework. Parent Education was used primarily as a control variable that was included to improve the accuracy of our estimates of the effects of the Homework variables on GPA (see the original reasoning for including this variable in Chapter 2); its interpretation is of less interest. With these caveats in mind, I might interpret these finding as follows (a real world interpretation, say to parents):

As you may know, many high school students complete a part or all of their homework while in school, whereas others complete all or most of their homework at home or

out of school. Some do a little of both. I was interested in whether these two types of homework—in school versus out of school—were equally effective in producing learning. To find out, I conducted research to examine the relative influence of time spent on homework in school and out of school on high school students' grades in school. I also took the education of the parents into account. The results suggest that these two types of homework indeed have different effects. Homework completed In School had virtually no effect on students' Grades. In contrast, Homework completed Out of School, presumably at home, had a fairly strong effect on Grades; students who completed more Homework Out of School achieved higher Grades, even after the Education level of their parents was taken into account. I encourage you to encourage your high schoolers to complete their homework at home rather than in school. If they do, that homework is likely to show an important payoff in their grades: the more homework they complete, the higher their grades are likely to be.

Again, this explanation would be worthwhile if you believed that these results explained the relations among these variables; as you will see, the findings from our next regression will create doubts about this. This explanation also side steps an important question: *why* does homework completed outside of school have an effect on GPA while homework completed in school does not? I can think of at least two possibilities. First, it may be that the process of doing homework out of school requires a greater degree of initiative and independence and that initiative and independence, in turn, improve grades. Second, it may be that when students complete homework in school that homework is essentially displacing instructional time, thus resulting in no net gain in time spent on learning. These possibilities are explored further in the research article (Keith et al., 2004), and you may have other ideas why this difference exists. These possible reasons for the difference are all testable in future research!

RULES OF THUMB: MAGNITUDE OF EFFECTS

Another issue that I should address is the criteria by which I argued that some effects were "tiny," whereas others were "substantial." One historical criticism of research in psychology is that many researchers have focused on and reported only statistical significance, ignoring the magnitude of effects (Cohen, 1994; Thompson, 1999). There is a consensus in psychology that researchers should report and interpret effect sizes in addition to statistical significance, and many journals now require such reporting (American Psychological Association, 2010). One advantage of multiple regression is that its statistics focus naturally on the magnitude of effects. R^2 and regression coefficients (b) can certainly be tested for their statistical significance. But R , R^2 and the regression coefficients (especially β) are scales that range from low to high, from zero to 1.0 for R and R^2 (and β usually, although not always, ranges between ± 1), and thus it is natural to focus on the magnitude of these effects. So, what constitutes a large versus a small effect? Although there are general rules of thumb for a variety of statistics (e.g., Cohen, 1988), it is also the case that Cohen and others have urged that each area of inquiry should develop its own criteria for judging the magnitude of effects.

Much of my research focuses on the influences on school learning, influences like homework, parent involvement, academic coursework, and so forth. Based on my research and reading in this area, I use the following rules of thumb for judging the magnitude of effects on learning outcomes (e.g., achievement, Grades). I consider β 's below .05 as too small to be considered meaningful influences on school learning, even when they are statistically significant. β 's above .05 are considered small but meaningful; those above .10 are considered moderate, and those above .25 are considered large. Using these criteria, the β associated with time spent on Homework Out of School is large, whereas the β associated

with time spent on Homework In School would be considered tiny, even if it were statistically significant. Keep in mind, however, that these rules of thumb apply to research on learning and achievement, and that I have little idea how well they generalize to other areas. You will need to use your and others' expertise in your own area of research to develop similar guidelines.

Regression findings can also be converted easily to other measures of effect size, notably Cohen's f^2 via the formula $f^2 = \frac{R^2}{1-R^2}$. A common rule of thumb for f^2 is that .02 represents a small effect, .15 a medium effect, and .35 a large effect (Cohen et al., 2003, p. 95). These authors also recommend that researchers develop rules of thumb for their own substantive areas of research, however. Using this conversion the overall regression would result in a Cohen's f^2 of .183, a medium effect. We will discuss f^2 in more detail in Chapter 5, where we will also see how to examine the effect of each predictor.

TESTING THE DIFFERENCE BETWEEN TWO REGRESSION COEFFICIENTS

In Chapter 2 we compared the magnitude of regression coefficients from two different equations. We also compared, qualitatively, the magnitude of two standardized coefficients from a single regression equation (Parent Education versus Homework), but postponed conducting a statistical comparison. We will do so now. Interestingly, this is a topic that I get asked about often in class, but few regression books address.

Note the regression results shown in Figure 4.5. Studying the β 's, you might note that Out-of-School Homework had a slightly larger standardized effect on Grades than did Parent Education. You might also reasonably wonder whether the effect for Homework was *statistically significantly* larger than the effect for Parent Education. Here is a simple way to make such comparisons (derived from postings on sci-tech.archive.net and allexperts.com).

First, standardize the two variables you want to compare, in this case BYParEd and F1S36A2. An easy way to do so in SPSS is to ask for descriptive statistics and click the box to save standardized versions of these variables (see Figure 4.6). Next, create two new composite variables, one that is the sum of the two new standardized variables, and one that is the difference between the two:

```
sum_pe_hw_z=zbypared+zf1s36a2 and  
dif_pe_hw_z=zbypared-zf1s36a2
```

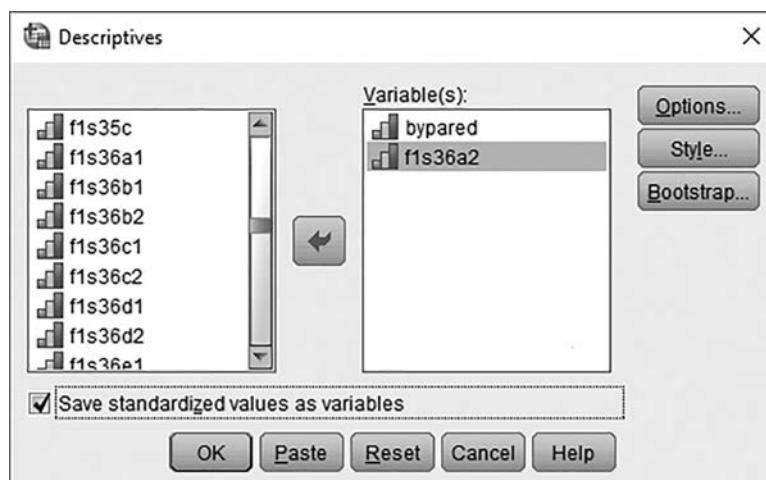


Figure 4.6 Saving standardized values (z-scores) of Parent Education and Out-of-School Homework (in SPSS).

64 • MULTIPLE REGRESSION

(here I have used the convention, from SPSS, that standardized versions of variables have the same names preceded by a z).

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.394 ^a	.155	.152	1.34998	

a. Predictors: (Constant), dif_pe_hw_zParent Education–Out HW Difference, sum_pe_hw_zParent Education–Out HW Sum, f1s36a1 TIME SPENT ON HOMEWORK IN SCHOOL

ANOVA					
Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 303.167	3	101.056	55.450	.000
	Residual 1649.320	905	1.822		
	Total 1952.486	908			

Model	Coefficients ^a						
	Unstandardized Coefficients			Standardized Coefficients	t	Sig.	95% Confidence Interval for B
	B	Std. Error	Beta				Lower Bound Upper Bound
1	(Constant) 5.644	.077			72.913	.000	5.492 5.796
	f1s36a1 TIME SPENT ON HOMEWORK IN SCHOOL .012	.031	.012	.379	.704		-.048 .072
	sum_pe_hw_zParent Education–Out HW Sum .360	.029	.391	12.506	.000		.303 .416
	dif_pe_hw_zParent Education–Out HW Difference -.015	.038	-.012	-.392	.695		-.089 .059

a. Dependent Variable: ffugrad ffu grades

Figure 4.7 Testing the difference between two regression coefficients from the same regression equation.

Second, re-conduct the regression, but substitute the two new composite variables instead of the variables from which they were created. That is, in the current example, regress 10th-grade GPA on In-School Homework, the Parent Education\Out-of-School Homework summed variable, and the Parent Education\Out-of-School Homework difference variable. The statistical significance of the coefficient associated with the difference between these two standardized variables (*diff_pe_hw_z*) is a test of the statistical significance of the difference between their standardized effects.

The relevant output from this regression is shown in Figure 4.7. Note that the model summary and the ANOVA table, showing the R^2 and its statistical significance, are identical to those for the earlier regression (Figure 4.5). In the table of coefficients, however, the coefficient associated with the difference between the two variables (*diff_pe_hw_z*) is not statistically significant. This means that the difference between these (standardized) variables is not statistically significant. Thus, while it appears that the effect of Out-of-School Homework is slightly larger than the effect of Parent Education (Figure 4.4), that difference is not statistically significant.

FOUR INDEPENDENT VARIABLES

Our next example will also continue our exploration of the effects of time spent on homework on high school students' achievement. The purpose of this example is twofold. First, it will extend our analyses to an example with four independent variables; you should feel quite comfortable with this extension by now, because the analysis and interpretation are very similar to those completed previously. Second, however, this extension of our example will

erect a few speed bumps in the merry analysis and interpretation road I have been leading you down. You should be troubled by the differences in these results and those presented previously. And although we will eventually resolve these problems, this example should begin to illustrate the importance of theory and thought in multiple regression analysis. On to the example.

Another Control Variable

For our previous examples, we have added a variable representing Parent Education to our regressions to “control,” to some extent, students’ family backgrounds. Our reasoning went something like this: parents who value education for themselves likely value education for their children, as well. Such parents are likely to emphasize learning, schooling, and studying more than are parents who place a lower value on education (Walberg, 1981, referred to such an orientation as the “curriculum of the home”). As a result, children in such homes are likely to spend more time studying; they are also likely to earn higher grades. I noted previously that we needed to include Parent Education in the regression because, if our speculation about the effects of Parent Education is correct, it is a potential *common cause* of both Homework and Grades.

What about other potential common causes? It seems likely that students’ academic aptitude, or ability, or previous achievement might also function in this manner. In other words, doesn’t it seem likely that more able students should not only earn higher grades but might also be inclined to spend more time studying? If so, shouldn’t some measure of students’ prior achievement be included in the regression as well?

Our next multiple regression example is designed with this speculation in mind. In it, I have regressed students’ 10th-grade GPA (FFUGrade) on both In-School (F1S36A1) and Out-of-School Homework (F1S36A2), as in the previous example. Also included is a measure of Parents’ highest level of Education (BYParEd), again as in the previous example. This new regression, however, also includes a measure of students’ Previous Achievement (BYTests), an average of students’ scores on a series of academic achievement tests in Reading, Mathematics, Science, and Social Studies administered in the 8th grade. This new regression, then, tests the effects on Grades of In-School Homework versus Out-of-School Homework, while controlling for Parents’ highest levels of Education and students’ Previous Achievement.

Regression Results

Descriptive statistics and the results of the multiple regression are shown in Figures 4.8 and 4.9. As shown in Figure 4.8, the linear combination of the variables representing parents’ education, previous achievement, time spent on in-school homework, and the time spent on out-of-school homework accounted for 28.2% of the variance in 10th-grade GPA ($R^2 = .282$), which appears to be quite an improvement over the 15.5% of the variance explained by the previous multiple regression (in subsequent chapters we will learn how to test this change in R^2 for statistical significance). The overall regression, as in the previous example, is statistically significant ($F[4, 874] = 85.935, p < .001$).¹ Everything seems to be in order.

Trouble in Paradise

When we focus on Figure 4.9, however, we are led to different conclusions than in our previous analysis. Parents’ Education level and time spent on Homework Out of School still had

Descriptive Statistics

	Mean	Std. Deviation	N
FFUGRAD ffu grades	5.7033	1.4641	879
BYPARED PARENTS' HIGHEST EDUCATION LEVEL	3.22	1.27	879
F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL	2.09	1.53	879
F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL	2.55	1.72	879
BYTESTS Eighth grade achievement tests (mean)	51.9449	8.6598	879

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.531	.282	.279	1.2432

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	531.234	4	132.808	85.935	.000 ^a
	Residual	1350.728	874	1.545		
	Total	1881.962	878			

a. Predictors: (Constant), BYTESTS 8th-grade achievement tests (mean), F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL, F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL, BYPARED PARENTS' HIGHEST EDUCATION LEVEL

b. Dependent Variable: FFUGRAD ffu grades

Figure 4.8 Multiple regression with four independent variables: descriptive statistics and model summary.

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.
	B	Std. Error			
1	(Constant)	1.497	.257	5.819	.000
	BYPARED PARENTS' HIGHEST EDUCATION LEVEL	9.12E-02	.037	.079	2.443
	F1S36A1 TIME SPENT ON HOMEWORK IN SCHOOL	-1.2E-02	.029	-.013	-.423
	F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL	.158	.027	.186	5.876
	BYTESTS 8th-grade achievement tests (mean)	6.80E-02	.006	.402	12.283

a. Dependent Variable: FFUGRAD ffu grades

Figure 4.9 Multiple regression with four independent variables: regression coefficients.

Table 4.1 Comparison of Regression Coefficients for the Three-Variable versus Four-Variable Multiple Regression

Variable	<i>Three independent variables</i>		<i>Four independent variables</i>	
	<i>b</i> (SE_b)	β	<i>b</i> (SE_b)	β
Parent Education	.271 (.037)	.234	.091 (.037)	.079
Previous Achievement	—	—	.068 (.006)	.402
In-School Homework	.012 (.031)	.012	-.012 (.029)	-.013
Out-of-School Homework	.218 (.028)	.256	.158 (.027)	.186

statistically significant effects on Grades, but the magnitude of these effects are very different. In the previous example, the unstandardized and standardized regression coefficients associated with Parent Education were .271 and .234, and now they are .091 and .079, respectively. Indeed, all the coefficients changed, as shown in Table 4.1, with the addition of the new independent variable.

What this means is that the conclusions we draw from these regressions will also be very different. Focusing on the variables of primary interest (the two Homework variables), we conclude from the three-independent-variable multiple regression that Homework time Out of School had a large effect on Grades, using my rules of thumb, but in the four-variable regression, we conclude that Homework had only a moderate effect on Grades. The effect of In-School Homework was small and not statistically significant in both analyses, although the sign switched from positive to negative. What is going on? Were all our conclusions from the earlier analysis erroneous? (You may be tempted to conclude that we should instead focus only on statistical significance, since the same variables remain statistically significant versus not statistically significant in the two regressions. This conclusion is incorrect, however, because it is not always the case that the same variables will remain statistically significant with the addition of new independent variables.)

You should be troubled by this development. It suggests that our conclusions about the effects of one variable on another change depending on which other variables are included in our analyses. Focusing on the three-variable regression, you would conclude that each additional unit (whatever it is) of time spent on Homework out of school results in a .218-point increase in GPA. If you believe the four-variable regression, however, you might argue that each additional unit of time spent on Homework out of school results in a .158-point increase in GPA. Which conclusion is correct?

This example illustrates a danger of multiple regression as illustrated so far: *The regression coefficients will often (although not always) change depending on the variables included in our regression equation.* This development certainly does not argue for the scientific respectability of our findings, however, nor does it bode well for the scientific respectability of multiple regression! If our conclusions change depending on the variables we include in our analyses, then knowledge and conclusions depend on our skill and honesty in selecting variables for analysis. Research findings should be more constant and less ephemeral if they are to form the basis for understanding, knowledge, and theory. Furthermore, this change in findings and conclusions means that, to some extent, we can find what we want by choosing the variables to include in our regression. Want to find that Parent Education has a moderate

to strong effect on GPA? Don't include previous achievement in your analysis. Want to conclude, instead, that Parent Education only has a small effect? Then do include a measure of previous achievement in your regression!

It may be small comfort if I tell you that this danger is not entirely a result of multiple regression, *per se*. Instead, it is a danger in most nonexperimental research, whatever statistical technique is used for analysis. This conundrum is one reason that many researchers argue against making causal conclusions from nonexperimental research: the results change depending on the variables analyzed. Of course, an admonition against nonexperimental research means that much scientific inquiry is simply not possible, because many worthy scientific questions—and especially questions in the behavioral sciences—are simply not testable through other means. This danger is also one reason that many researchers focus on *prediction* rather than explanation. We may be on slightly more stable ground if we make statements like “when GPA was regressed on Parent Education and Homework In and Out of School, Homework Out of School and Parent Education were statistically significant *predictors* of GPA, whereas Homework In School was not.” Such a predictive conclusion avoids the implied causal connection in my statements (e.g., that time spent on Homework Out of School has a strong effect (as in cause and effect) on Grade-point average). But a focus on prediction rather than explanation is also scientifically less valuable; it does not allow us to use our findings for the development of theory or to change the status quo. If all we can conclude is that homework predicts achievement, then we cannot legitimately encourage children, parents, or teachers to use homework as a method of improving learning. *Intervention thinking requires causal thinking!* (cf. Tufte, 2001). Causal thinking, in turn, requires careful thought and knowledge of previous research and theory.

COMMON CAUSES AND INDIRECT EFFECTS

Fortunately, there is a resolution to this dilemma. Ironically, the solution requires additional, more formal, causal thinking, rather than less, and will be dealt with in depth in the beginning of Part 2. In the meantime, I will present a brief preview of what is to come, enough, I hope, to quell your fears to some extent.

Figure 4.10 shows a model of the thinking underlying our four-independent-variable multiple regression, with the arrows or paths in the model representing the presumed influence of one variable on another. We have been using such models from the beginning of our regression journey. A more detailed model is shown in Figure 4.11. The explicitness of this model may seem surprising, but it should not; most of the paths simply present in figural form my earlier explanations concerning the reasons for the variables included in the multiple regression. The inclusion of the four independent variables in the multiple regression implies that we believe these variables may affect Grades and that we want to estimate the magnitude of these effects; it makes sense, then, to include paths representing these possible effects in the model. This portion of the model is implied, whether we realize it or not, every time we conduct a multiple regression analysis. Recall that we included the variables Parent Education and Previous Achievement in our regression because we thought that these variables might affect both Grades and Homework, and thus the paths from these variables to the two Homework variables make sense as well. This reasoning accounts for all the paths except two: the path from Parent Education to Previous Achievement and the arrow from In-School Homework to Out-of-School Homework. Yet it makes sense that if Parent Education affects Grades it should affect achievement as well. My reasoning for drawing the path from In School to Out-of-School Homework is that students who work on homework in school will take home that portion of their homework that is not completed in school. Although not discussed here, most of these decisions are also supported by relevant theory.

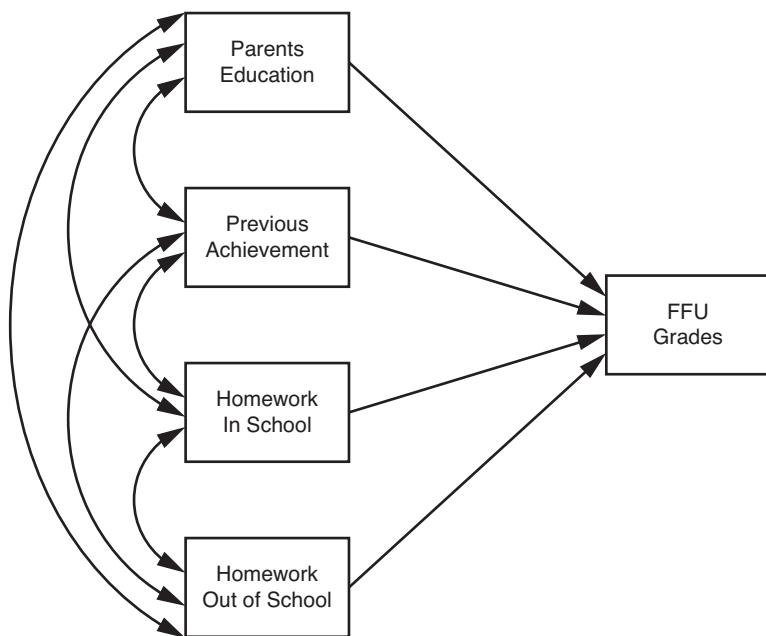


Figure 4.10 MR with four independent variables, in figural (path) format.

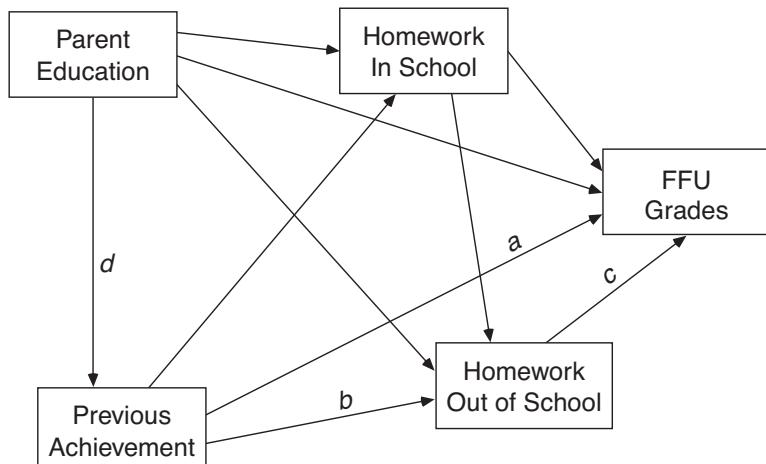


Figure 4.11 A more complete version of the four independent variable path model. This model makes explicit the presumed ordering of the independent variables.

The paths labeled *a* and *b* make explicit what we mean by a *common cause*: Our model assumes that Previous Achievement affects Grades directly (path *a*) and that it also affects Homework completed out of school (path *b*). If so, and if both of these paths are statistically significant and meaningful, then Previous Achievement *must* be included in the model to get an accurate estimate of the effects of Out-of-School Homework on Grades. *To interpret*

regression coefficients as effects, all common causes of the presumed cause and the presumed effect must be included in the model. If they are not, then the regression coefficients will be inaccurate estimates of the effects; in many cases, they will be overestimates of these effects. This, then, was the reason for the drop in the apparent effect of Out-of-School Homework on Grades from the three-independent-variable regression to the four-independent-variable regression: Previous Achievement, a common cause of both Homework and Grades, was erroneously excluded from the three-independent-variable model. With this common cause included in the model, we get smaller and more accurate estimates of effects.

If you include in the regression a variable that logically is prior to Homework Out of School and Grades, but is *not* a common cause of Homework and Grades, the regression weight for Out-of-School Homework will *not* change. That is, a variable that affects only Out-of-School Homework but not Grades will not change the Homework regression coefficient. Likewise, a variable that affects Grades but not Out-of-School Homework will not change the Homework regression coefficient.²

There is a different reason for the drop in the apparent effects of Parent Education on Grades (from $b = .271$ to $b = .091$) in moving from the first to the second multiple regression. A portion of the effect we initially attributed to the effect of Parent Education on Grades now appears as an *indirect* effect. In our current model, Parent Education affects Previous Achievement (path d), and Previous Achievement affects Grades. Thus Parent Education affects Grades *indirectly* through Previous Achievement. Another way of saying this is that Previous Achievement partially *mediates* the effect of Parent Education on Grades. When we discuss structural equation modeling and path analysis, we will learn how to calculate these indirect effects, and you will find that for both models the *total* effects of Parent Education on Grades are the same. For now, just remember that it is not necessary to include mediating effects for multiple regression to provide accurate estimates of effects, but that the regression coefficients from simultaneous regression (the type of multiple regression we are currently doing) only focus on *direct*, not mediating, effects. In Chapter 5 we will see that another type of multiple regression, sequential regression, can be used to focus on total effects. We will discuss mediation more completely in Chapter 9, and more so in Part 2 of the text.

To reiterate, to interpret regression coefficients as the effects of one variable on another, the common causes of the presumed cause and presumed effect must be included in the regression. If your regression *does* include these common causes, you can indeed make such interpretations (well, there are a few other assumptions that we will deal with later in Chapter 10). You can thus rest somewhat easier, because this requirement, although difficult, is not impossible to satisfy. In contrast, it is not necessary to include intervening variables in your regression, although you should keep in mind that if you do include mediating variables between your presumed cause and presumed effect, your regression results are estimating only a portion of the total effect of one variable on another. We will deal with these topics more extensively in the next chapter and at the beginning of Part 2 of this book.

THE IMPORTANCE OF R^2 ?

As we switched from a regression with three independent variables to one with four, we also noted that the R^2 increased; we explained more variance in students' GPAs. As you will discover in your reading, some researchers make much of the size of R^2 and try to explain as much variance as possible in any given regression. I do not, and with a little reflection on the previous few sections, you may understand why. It is relatively easy to increase R^2 ; just add more variables to the regression that predict the outcome. In the current example, we could add previous GPA to the regression, which would likely lead to another healthy increase in the variance explained, and we might also add a measure of motivation, assuming that it,

too, should increase the variance explained in students' Grades. But recall that our primary interest in these regressions was to understand the effects of In-School and Out-of-School Homework on Grades. Indeed, it makes sense to add variables to the regression *if* these variables are common causes of Homework and Grades. In contrast, given our purpose, it makes little sense to add variables to the regression if they are not common causes of Homework and Grades. Thus, although we can inflate the R^2 by adding to the regression variables that affected students' Grades (but not Homework), such additions serve little purpose other than inflating the R^2 ; they will not help us to better understand the effects of Homework on Grades. We can also increase R^2 by adding intervening variables between Homework and Grades, but unless our interest is in understanding the indirect effects of Homework on Grades, this addition will also make little sense.

It is tempting to think that the more variance you have explained the closer you have come to understanding and explaining some phenomenon. But this is true only if you have included the *proper* variables in your regression. It is also tempting to think that if you find a high R^2 you must have included the proper variables in your regression. This is also not necessarily the case. Suppose you regress students' High School GPAs on their college GPAs. You will likely get a fairly high R^2 , but college grades do not influence High School grades (you have confused cause and effect). You have not explained the phenomenon. Or perhaps you decide to regress reading proficiency of elementary students on their shoe size. Again you would likely get a high R^2 , but you have not explained reading skill, nor have you chosen the right variables for your regression. The high R^2 is the result of a spurious association (there is a common cause of reading proficiency and shoe size: growth or age). The high R^2 did not assure that you chose the correct variables for inclusion in the regression.

Should you then ignore R^2 ? No, of course not, and we have not done so here. My point is simply this: yes, other things being equal, the higher the R^2 the better, but a high R^2 is generally not the most important criterion if we are conducting regression for the purposes of explanation. What I suggest is that you make sure that R^2 is reasonable, which depends on the constructs you are studying and requires some knowledge of research in this area. For the dependent variable of Grades, for example, I generally expect to explain 25% or so of the variance in Grades, and our four-independent-variable regression is consistent with this expectation. If we are focusing on achievement test scores, I expect a higher R^2 , whereas if our dependent variable is self-concept, I expect to explain less variance. Some phenomena are easier to explain than others; more reliable dependent variables should also lead to more explained variance. Likewise, prior to interpretation you should make sure the other regression results are reasonable. Suppose in the Grades–Homework regression used in this chapter you found a negative regression coefficient associated with Previous Achievement. That result is so implausible that I likely would not interpret the findings no matter how high the R^2 and certainly not without additional investigation.

Are you surprised that I think it's reasonable to explain only 25% of the variance in Grades? This means that 75% of the variance in Grades is unexplained! I have several responses. First, yes, it's difficult to explain human behavior; we are unpredictable creatures. To put it differently, you'd probably be insulted if I declared that I could predict your behavior with a high degree of accuracy. If I rattle off the influences of your grades and tell you that I can predict your future grades very accurately from these variables, you might even feel angry or defeated. When you think of it this way, you might thus be relieved that we are explaining only 25% of the variance in Grades. Or, as Kenny put it, "human freedom may then rest in the error term" (i.e., the unexplained variance) (1979, p. 9).

I should also note for the benefit of those more familiar with ANOVA that we will generally consider explaining 25% of the variance in a dependent variable as a *large* effect size in an ANOVA. "A good rule of thumb is that one is fooling oneself if more than 50% of the

variance is predicted" (Kenny, 1979, p. 9). It happens, but surprisingly infrequently. When it does happen, it's often the case that we have analyzed longitudinal data with the same variable measured at two points in time (as both the dependent and an independent variable). Finally, I should note that others place a greater emphasis on R^2 than I do. Ask the instructor teaching your class: What's his or her position on the importance of R^2 ?

PREDICTION AND EXPLANATION

Let's spend a little more time on a topic that we have broached a few times so far: the distinction between prediction and explanation. The underlying purpose of our research , and whether that purpose is related to prediction or explanation, has important implications for how we choose the variables for regression, conduct the analysis, and interpret the results. As you will see in Chapter 5, some methods of multiple regression are better suited for one purpose than another. I am assuming that most readers will be interested in using multiple regression for explanatory purposes, and most of my examples have been set up accordingly. Many researchers blur these two purposes, however, and I may have done the same in previous chapters. It is time, however, to make the distinction sharper.

In most of the examples so far, we have been interested in the *effects*, or *influences*, of one or more variables on an outcome. Such an interest denotes an explanatory purpose; we want to *explain*, partially, how an effect comes about, and we use our independent, or explanatory, variables to accomplish this purpose. The explanatory intent of these examples is further revealed in our interpretations; we talk of the effects of homework, for example, on grades. Even more revealing, we discuss the probable results if students were to increase the time they spent on homework. Such an interpretation reveals a clear inference of cause and effect, and such an inference is the essence of explanation.

It is also possible to use multiple regression for the purpose of prediction. You may be an admissions officer of a college interested in predicting, in advance, which applicants to your college are most likely to perform well in school so that you can accept these students and reject those who are likely to perform poorly. In such an example, you have no real interest in explanation and no interest in making cause and effect interpretations. Your only interest is in making as accurate a prediction as is possible from the various predictor variables available to you. If prediction is your goal, you *will* want to maximize the R^2 (in contrast to our earlier discussion).

As discussed previously in this chapter, if your interest is in explanation, you need to choose the variables for the regression equation very carefully. In addition to the dependent and independent variables of primary interest, your regression should include any likely common causes of these variables. At the same time, you should refrain from including any irrelevant variables (unless you wish to demonstrate that they are not common causes), because they are likely to dilute your power and muddy your findings. Because of the care needed in choosing the variables to include in an explanatory regression analysis, the researcher needs a firm grounding in relevant theory and previous research. Theory and previous research go far in telling you which variables should be included in such regressions.

If your interest is in prediction, however, you have much less need to fret over your selection of variables. Certainly, a knowledge of theory and previous research can help you maximize successful prediction, but it is not critical. In fact, if your purpose is simple prediction, then you could even use an "effect" to predict a "cause."

An example will help illustrate this point. Intelligence (or aptitude or previous achievement) commonly appears in theories of school learning as an important influence on students' learning. Thus, explanatory regressions with achievement tests, grades, or some other measure of learning as an outcome often include a measure of one of these constructs

(intelligence, aptitude, etc.). It would make little sense to include a measure of grades as an independent variable in a regression analysis attempting to explain intelligence, because the analysis would reverse the “cause” and the “effect.” If our interest were only in *predicting* intelligence, however, including grades among the predictors would be perfectly acceptable. If grades made prediction more accurate, why not use them?

As noted previously, many researchers confuse these purposes, and thus don’t think through the variables carefully when conducting an explanatory regression. Others may end up using an approach more suited to prediction, when their real interest is in explanation. Even worse, it is not unusual for a researcher to set up and conduct a prediction-oriented regression, but then interpret the results in an explanatory fashion. For example, I have seen researchers speak of prediction throughout a research article, carefully eschewing any sort of causal language. But then, in the discussion, the researchers argue that programs or interventions are needed to change the level of a variable from their study to effect change in the outcome. But such an argument requires, and is predicated on causal, explanatory thinking. This bait and switch, while presumably unintentional, is poor practice and may lead to wildly erroneous conclusions (think about an erroneous, explanatory interpretation of our previous prediction example). Don’t fall prey to this bait and switch in your own research or in reading the research of others. Be clear as to whether your purpose is explanatory or predictive, choose your method accordingly, and interpret your findings properly.

SUMMARY

This chapter extended the example used in previous chapters to illustrate multiple regression with three and four independent variables. First, we regressed Grades on Parent Education, time spent on In-School Homework, and time spent on Out-of-School Homework. The results suggested that Out-of-School Homework had a strong effect on Grades, whereas In-School Homework had no such effect. In the second example, we added another variable to the regression, students’ Previous Achievement. In this regression, Homework Out of School had only a moderate effect on Grades. As we have seen, the analysis and interpretation of these examples were very similar to those in earlier chapters. We can easily add additional independent variables, with straightforward analysis and interpretation.

We made a disturbing discovery, however: the regression coefficients changed in magnitude as we added new variables to the multiple regression equation. I argued that there may be two reasons for such changes: first, if a common cause of a presumed cause and a presumed effect is included in a regression, the regression coefficients will change from those found when such a variable is excluded from the regression. Second, if an intervening variable is included in a regression between the presumed cause and the presumed effect, the regression coefficients will change in magnitude, because the regression coefficients focus only on direct effects. The first reason for the change in the regression coefficients constitutes a serious error in the analysis, but the second does not.

I discouraged a fixation on R^2 , as well as a temptation to maximize R^2 . We should include the relevant variables and not load up our regressions with irrelevant variables. You might, in fact, reasonably be suspicious when you obtain R^2 ’s above .50.

These problems and concerns only apply to regression for which you are interested in explanation, that is, when you are interested in the magnitude of the effect of one variable on another. They are less applicable when your chief interest is in the simple prediction of one variable from a group of others. I have argued, however, that such simple prediction is scientifically less appealing, because it does not allow you to think in terms of theory, interventions, policy, or changes to the status quo. One thing you must *not* do is to pretend you are interested in simple prediction but then switch to an explanatory conclusion (e.g., if you

spent more time on homework out of school, your grades would likely improve). Unfortunately, such bait and switch tactics are depressingly common in the research literature. I have also argued that it is necessary to think causally to understand what is happening in explanatory analyses; path diagrams are a useful heuristic aid to such thinking. We will use such diagrams throughout this text to explicate important concepts. We will cover these topics in more detail in Part 2; for now, we will continue to focus on the proper analysis and interpretation of multiple regression. You have no doubt noticed that there are several topics, including prediction and explanation, understanding which variables should be included in a regression, and the proper interpretation of regression coefficients, that we will revisit on a regular basis. I believe these issues are important to introduce early to get you thinking about them. We will revisit them as your knowledge increases and eventually resolve them.

EXERCISES

1. If you have not done so already, conduct the two multiple regression analyses presented in this chapter. Compare your results to mine. Analyze the descriptive statistics for the variables used (e.g., means, standard deviations, variances, minimum and maximum for all variables, frequency distributions of Parent Education, Previous Achievement, Homework In School, and Homework Out of School) to make sure you understand the metric of the variables. Provide a formal, English, and real-world (e.g., to parents) interpretation of the findings.
2. Does the size of an adolescent's family influence his or her self-esteem? Does TV viewing affect self-esteem? Using the NELS data, regress 10th-graders' Self-Concept scores (F1Cncpt1) on Parent Education (BYParEd), Achievement (BYTests), Family Size (BYFamSiz), and TV Time (create a composite by calculation of the mean of BYS42A and BYS42B). Check the variables to make sure that you understand their metric (F1Cncpt1 has positive and negative values because it is a mean of z scores; positive scores represent more positive self-concept). Clean up the data, as needed, and run the multiple regression. Interpret your findings. Do any of your findings surprise you? Are you willing to interpret the regression coefficients as effects? Why or why not?
3. Does age affect eating disorders in women? Tiggemann and Lynch (2001) studied the effect of women's body image on eating disorders across the life-span. The file labeled "Tiggeman & Lynch simulated.sav" includes a simulated version of some of the variables from this research (the data are also contained in an Excel file and a text file with the same name, but the extension ".xls" or ".dat"). The variables in the file are Age (21 to 78), the extent to which the women habitually monitored their bodies and how it looked (Monitor), the extent to which the women felt shame when their bodies did not look the way they expected (Shame), the extent to which women felt anxiety about their bodies (Anxiety), and the extent to which the women endorsed eating disorder symptoms (Eat_Dis). Is the correlation between age and eating disorders statistically significant? When you regress eating disorders on Age and these other variables (Monitor, Shame, and Anxiety), does age have an effect on eating disorders? Which of these variables are most important for explaining eating disorders? Interpret your findings.
4. Conduct a multiple regression on four or five variables of your choice. Look through the NELS data and find a variable you are interested in explaining. Pick several independent variables you think may help in explaining this dependent variable. Examine descriptive statistics for these variables and frequencies for variables with a limited number of response options to make sure you understand their scales. Clean up the data as needed; that is, make sure the variables are coded in the proper order and that

missing values are dealt with properly (e.g., “Don’t know” responses are coded as missing, rather than as a value that will be analyzed). Conduct the multiple regression and interpret the results. Are there any threats to your analysis and interpretation (e.g., neglected likely common causes)?

5. Do all those advertisements you see for drugs affect your perceptions of risk for disease? Park and Grow (2008) questioned whether exposure to direct-to-consumer advertising for antidepressants affected people’s perceptions of the prevalence of depression and their own risk of depression. The file “depression advertising.sav” includes a simulated version of some of the variables used in Park & Grow’s research. The file includes data from 221 (simulated) undergraduate students enrolled in introductory advertising classes. The variables in the file are Age (in years), Experience (whether they knew people with and treated for depression, the sum of 3 questions each coded 1 [no] or 2 [yes]), Familiarity with advertisements for depression drugs (a sum of familiarity on a 1 to 7 scale for six antidepressant ads, with high scores representing greater familiarity), perceived Prevalence of depression (students’ perceptions of U.S. prevalence of depression, in percentages, so that a high score represents a higher estimate of the prevalence of depression), and Risk, students’ perceptions of their own risk of suffering from depression in their lives (percentage). Examine descriptive statistics for these variables to make sure they are reasonable (given the brief description of each). The key variables are perceived Risk and Familiarity with advertising. What is the correlation between these two variables; are they related? Regress Risk on Familiarity, while controlling for Age, Experience, and Prevalence. Does familiarity with antidepressant advertising affect perceived Risk of depression? Which variable is most important for explaining perceived Risk? Provide a formal, English, and real world interpretation of your findings.
6. This exercise is designed to explore further the nature of common causes, and what happens when non-common causes are included in a multiple regression. We will begin our analysis of these data here, and will return to them in Chapter 9 and in Part 2 when we have the tools to explore them more completely. There are two data files for this exercise, both including variables labeled X1 X2 X3 and Y1. In both files, the three X variables are intercorrelated, but variable X2 is not a common cause of variables Y1 and X3. For the data in the first file (common cause 2.sav), variable X2 has no effect on Y1. In the second file (common cause 3.sav), variable X2 has no effect on variable X3. For “common cause 2.sav” regress variable Y1 on variables X1 X2 and X3. Next, regress variable Y1 on just X1 and X3. Compare the regression coefficients for variable X3 in the first versus the second analysis. Did it change substantially? Now do the same analysis for “common cause 3.sav.” Again, does the coefficient for variable X3 change from the first to the second regression? Discuss the meaning of these findings in class.

Notes

- 1 You may—and should—wonder about the substantial change in degrees of freedom. In the previous example, we had 3 and 905 degrees of freedom, with 3 degrees of freedom for the regression of Grades on three independent variables ($df = k = 3$) and 905 df for the residual ($df = N - k - 1 = 909 - 3 - 1 = 905$). Now we have 4 and 874 df . The 4 makes sense ($df = k = 4$), but the 874 does not. The reason is due to our treatment of missing data. All large-scale surveys have missing data, as does NELS. In these regressions I have used *listwise deletion* of missing data, meaning that any person who had any missing data on any one of the five variables in the analysis was not included in the analysis. Apparently, some students who had complete data when four variables were used (Grades, Parent Education, Homework In School, and Homework Out of School) had missing data for the new variable Previous Achievement. When this new variable was added, our sample size (using listwise deletion) decreased from 909 to 879, so our new residual degrees of freedom

equals $879 - 4 - 1 = 874$. Listwise deletion is the default treatment of missing data in SPSS and most other programs, but there are other options as well. Better options for dealing with missing data are discussed in Part 2 of this book.

- 2 To demonstrate why this is so, I need to skip ahead to some concepts presented in Part 2, Chapter 12. Our formula for calculating β from correlations [e.g., $\beta_1 = (r_{y1} - r_{y2}r_{12}) / (1 - r_{12}^2)$] will not work because we are talking about *effects*, not correlations, equal to zero. Thus, you can either take these statements on faith for now or continue with this note. Focus on the model shown in Figure 4.12, a much simplified version of the model from Figure 4.11. The variables of primary interest are Homework Out of School (Homework) and Grades. The variable labeled X is a potential common cause of Homework and Grades. The paths are equivalent to β 's. The path c is equal to the regression weight for Homework when Grades is regressed on X and Homework, and the path b is equal to the regression weight for X for this same regression. We will see in Chapter 12 that $r_{\text{Grades} \cdot \text{Homework}} = c + ab$. But if X has no effect on Grades (the first way by which X would not be a common cause), then $b = 0$, and, as a result, $r_{\text{Grades} \cdot \text{Homework}} = c$. In other words, in this case the β is the same as the r . This means that when X affects Homework but not Grades the β for Homework from the regression of Grades on Homework and X will be the same as the correlation between Grades and Homework. Recall from Chapter 1 that β is equal to r when there is only a single independent variable. Thus my comment “a variable that affected only Out-of-School Homework but not Grades would not change the Homework regression coefficient.” My second statement was “a variable that affected Grades but not Out-of-School Homework would not change the Homework regression coefficient.” In this case the path a would be equal to zero. Using the same formula as above $r_{\text{Grades} \cdot \text{Homework}} = c + ab$, if a were equal to zero, then $r_{\text{Grades} \cdot \text{Homework}} = c$ again. If X has no effect on Homework, then with the regression of Grades on Homework and X , the regression coefficient for Homework will be the same as if X were not included in the regression. If a variable is not a common cause of Homework and Grades, then its inclusion in the regression will not change the regression coefficients. For more information see the discussion of common cause in Chapter 9.

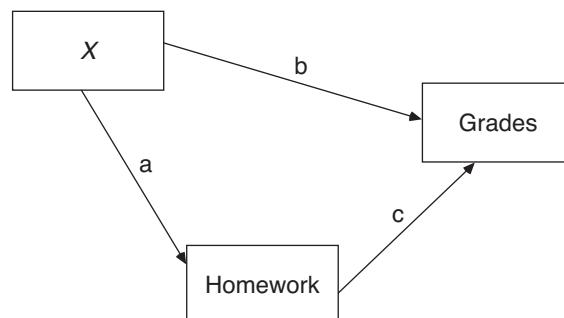


Figure 4.12 Simplified version of Figure 4.7, used to demonstrate what happens to regression coefficient equations in the absence of common causes.

5

Three Types of Multiple Regression

Simultaneous Multiple Regression	79
<i>The Analysis</i>	79
<i>Purpose</i>	80
<i>What to Interpret</i>	80
<i>Strengths and Weaknesses</i>	81
Sequential Multiple Regression	81
<i>The Analysis</i>	81
<i>Comparison to Simultaneous Regression</i>	83
<i>Problems With R^2 as a Measure of Effect</i>	87
<i>Cohen's f^2 as a Measure of Effect</i>	89
<i>Other Uses of Sequential Regression</i>	89
<i>Interpretation</i>	93
<i>Summary: Sequential Regression</i>	94
Stepwise Multiple Regression	95
<i>The Analysis</i>	96
<i>Danger: Stepwise Regression Is Inappropriate for Explanation</i>	97
<i>A Predictive Approach</i>	98
<i>Cross-Validation</i>	99
<i>Adjusted R^2</i>	100
<i>Additional Dangers</i>	100
<i>Alternatives to Stepwise Regression</i>	101
<i>Summary: Stepwise Regression</i>	101
The Purpose of the Research	102
<i>Explanation</i>	104
<i>Prediction</i>	104
Combining Methods	105
Summary	105
Exercises	106
<i>Notes</i>	107

The type of multiple regression that we have been using so far in this book is one of three major types or flavors of multiple regression, commonly called simultaneous or forced entry regression. In this chapter, we will compare simultaneous regression to two other types of multiple regression, sequential (hierarchical) regression and stepwise regression. As you will see, the different types of multiple regression serve different purposes and have different interpretations and different strengths and weaknesses.

We will analyze one problem several different ways to illustrate the differences in the three regression approaches. Suppose you are interested in the effect of self-perceptions on some aspect of academic performance. Specifically, you are interested in achievement in Social Studies. We will use the NELS data and the 10th-grade achievement standardized scores on the History, Civics, and Geography (or Social Studies) test (F1TxHStd). For measures of self-perceptions, the examples will use a short measure of 10th-grade self-esteem (F1Cncpt2), made up of seven items such as "I feel I am a person of worth, the equal of other people" and "On the whole, I am satisfied with myself." The items were reversed, if necessary, so that high scores represented higher self-esteem. The items were converted to z scores and then averaged to create the composite. (NELS also includes another self-esteem variable, labeled F1Cncpt1, which uses fewer items than does the F1Cncpt2 composite that we are using.) Also included in the regressions is a short measure of locus of control (F1Locus2), a measure of the degree to which people believe they control their own destiny (an internal locus of control) versus the extent to which they believe external forces control them (an external locus). Sample items include "In my life, good luck is more important than hard work for success" and "Every time I try to get ahead, something or somebody stops me." F1Locus2 included six items, with higher scores representing a more internal locus of control.

In the regressions, we will also include two control variables in the spirit of our previous discussion of the importance of including common causes in our analyses. Instead of parent education level, we turn to a broader socioeconomic status (SES) variable (BySES). This SES variable includes a measure of the parents' level of education, but also includes measures of the parents' occupational status and family income. BySES is a mean of z -scores of these items. Such SES variables are common in regression analyses of educational outcomes, although they may go by the name of Family Background, rather than SES. We will call this variable SES for now; remember, however, that it is much more than a measure of income. Students' Grade-Point Average from grades 6 to 8 (ByGrads), on a standard 4.0 scale, was included in the regressions as a measure of students' previous academic performance. Descriptive statistics for the five variables are shown in Figure 5.1, and the correlation matrix of the variables is shown in Table 5.1. The path model version of the regression set-up is shown in Figure 5.2.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE	923	28.94	69.16	50.9181	9.9415	98.834
BYSES SOCIO-ECONOMIC STATUS COMPOSITE	1000	-2.414	1.874	-3.1E-02	.77880	.607
BYGRADS GRADES COMPOSITE	983	.5	4.0	2.970	.752	.566
F1CNCPT2 SELF-CONCEPT 2	941	-2.30	1.35	3.97E-02	.6729	.453
F1LOCUS2 LOCUS OF CONTROL 2	940	-2.16	1.43	4.70E-02	.6236	.389
Valid N (listwise)	887					

Figure 5.1 Descriptive statistics for the NELS variables used in the chapter.

Table 5.1 Intercorrelations among 10th-Grade Social Studies Test Score, Parent SES, Previous GPA, Self-Esteem, and Locus of Control

Variables		F1TxHStd	BySES	ByGrads	F1Cncpt2	F1Locus2
F1TxHStd	10th-Grade Standardized Test	1.000				
BySES	Socioeconomic Status Composite	.430	1.000			
ByGrads	Grades Composite	.498	.325	1.000		
F1Cncpt2	Self-Concept 2 Composite	.173	.132	.167	1.000	
F1Locus2	Locus of Control 2 Composite	.248	.194	.228	.585	1.000

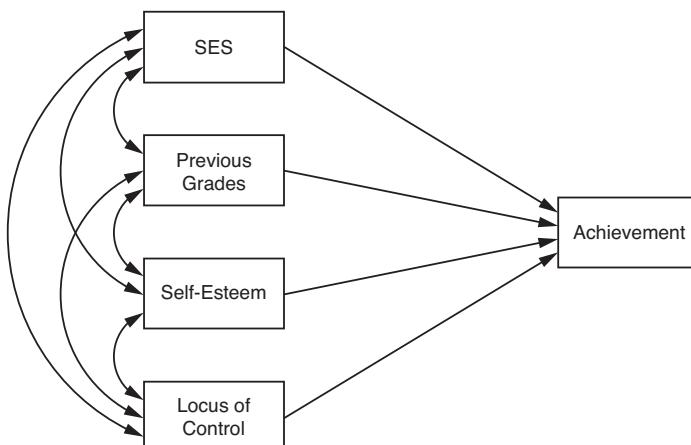


Figure 5.2 Path representation of the simultaneous regression of Social Studies Achievement on SES, Previous Grades, Self-Esteem, and Locus of Control.

SIMULTANEOUS MULTIPLE REGRESSION

The Analysis

In the type of multiple regression we have been using so far, all the independent variables were entered into the regression equation at the same time, thus the label *simultaneous regression*. This type of regression is also referred to as *forced entry regression*, because all variables are forced to enter the equation at the same time, or *standard* multiple regression. The simultaneous multiple regression results are shown in Figure 5.3. These are the type of results you are used to looking at, so we will not spend much time with them. First, we focus on the *R* and the *R*² and their statistical significance; the four explanatory variables in combination account for 34% of the variance in 10th-grade Social Studies test scores. The overall regression is statistically significant ($F = 112.846 [4, 882], p < .001$). The next step is to focus on the unstandardized regression coefficients, their statistical significance and confidence intervals, and the standardized regression coefficients. From this portion of Figure 5.3, you can see that all of the variables except Self-Esteem have a statistically significant effect on the Social Studies test score. SES and previous Grades have a strong effect, whereas Locus of Control has a small to moderate effect.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.582 ^a	.339	.336	8.0412

a. Predictors: (Constant), F1LOCUS2 LOCUS OF CONTROL 2, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1CNCPT2 SELF-CONCEPT 2

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29186.88	4	7296.721	112.846	.000 ^a
	Residual	57031.03	882	64.661		
	Total	86217.92	886			

a. Predictors: (Constant), F1LOCUS2 LOCUS OF CONTROL 2, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1CNCPT2 SELF-CONCEPT 2

b. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Coefficients^a

Model	Unstandardized Coefficients			t	Sig.	95% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	35.517	1.226	28.981	.000	33.112	37.923	
	SES	3.690	.378	.285	9.772	.000	2.949	4.431
	Previous Grades	5.150	.399	.380	12.910	.000	4.367	5.933
	Self-Esteem	.218	.501	.015	.436	.663	-.764	1.201
	Locus of Control	1.554	.552	.097	2.814	.005	.470	2.638

a. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Figure 5.3 Simultaneous regression of Social Studies test scores on SES, Previous Grades, Self-Esteem, and Locus of Control.

Purpose

Simultaneous regression is primarily useful for explanatory research to determine the extent of the influence of one or more variables on some outcome. In the present example, we could use simultaneous regression to determine the extent of the influence of Self-Esteem and Locus of Control on social studies achievement, while controlling for SES and previous academic performance. Simultaneous regression is also useful for determining the *relative* influence of each of the variables studied; indeed, it may be the best method for making this determination. As noted in the previous chapter, simultaneous regression estimates the direct effects of each independent variable on the dependent variable.

Because explanation subsumes prediction, however, simultaneous regression can also be used to determine the extent to which a set of variables *predicts* an outcome and the relative importance of the various predictors. For the current example, we can examine the β 's to conclude that previous Grades is the best predictor among this set of variables, followed by SES and Locus of Control. Simultaneous regression can also be used to develop a prediction equation; for the current example, the b 's could be used in an equation with a new sample of students to predict 10th-grade social studies achievement.

What to Interpret

In simultaneous multiple regression, the R^2 and associated statistics are used to determine the statistical significance and importance of the overall regression. The regression coefficients

are used to determine the magnitude of effect of each variable (controlling for the other variables) and, as we have seen, can be used to make policy or intervention recommendations. Such recommendations are particularly useful when the variables used have a meaningful metric (unlike the current example), using the unstandardized regression coefficients. The standardized coefficients are useful for determining the *relative* importance of each explanatory variable.

Strengths and Weaknesses

As we will see, simultaneous MR is very useful when the goal of research is explanation, because of the ability to focus on both the overall effect of all variables and the effect of each variable with the others controlled. The regression coefficients are useful for making predictions concerning what would happen if interventions or policy changes were made (e.g., how much would achievement increase if one were able to effect a change in locus of control from external to internal), and the standardized coefficients can provide information concerning the relative importance of various influences. If one has used theory and previous research to choose the variables to include in the regression, simultaneous regression can indeed provide estimates of the effects of the independent on the dependent variables. We have already broached the primary weakness of simultaneous MR: the regression coefficients can change, perhaps drastically, depending on the variables included in the regression equation.

SEQUENTIAL MULTIPLE REGRESSION

Sequential (also called hierarchical) regression is another common method of multiple regression and, like simultaneous regression, is often used in an explanatory manner. We will spend considerable time discussing the method, its interpretation, strengths, and weaknesses. This discussion will also point toward similarities and differences with simultaneous regression. We will end with a summary of this presentation.

The Analysis

With sequential multiple regression, the variables are entered into the regression equation one at a time, in some order determined in advance by the researcher. For our current example, I entered SES into the equation in the first block, then previous GPA in the second block, then Self-Esteem, and finally Locus of Control.

The primary results of interest are shown in Figure 5.4. The first table in the figure shows that the variables were entered in four blocks (rather than one) and the order of entry of the variables. The second portion of the figure provides the statistics related to each block of the sequential regression. With sequential regression, instead of focusing on the regression coefficients, we often focus on the change in R^2 (ΔR^2) to determine whether a variable is important and to test the statistical significance of each variable in the equation.

SES was the first variable entered, and the ΔR^2 associated with SES was .185 (.185 minus 0, because no variance was explained prior to the entry of SES into the equation). With the entry of previous Grades in the equation, the R^2 increased to .328, so ΔR^2 for previous Grades is .143 (.328 – .185), and the addition of Self-Esteem increased the variance explained by .5% ($\Delta R^2 = .005$), and so on.

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	BYSES SOCIO-ECONOMIC STATUS COMPOSITE ^a	.	Enter
2	BYGRADS GRADES COMPOSITE ^a	.	Enter
3	F1CNCPT2 SELF-CONCEPT 2 ^a	.	Enter
4	F1LOCUS2 LOCUS OF CONTROL 2 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Model Summary

Added to the Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
SES	.430 ^a	.185	.185	200.709	1	885	.000
Previous Grades	.573 ^b	.328	.143	188.361	1	884	.000
Self-Esteem	.577 ^c	.333	.005	6.009	1	883	.014
Locus of Control	.582 ^d	.339	.006	7.918	1	882	.005

a. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE

b. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE

c. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1CNCPT2 SELF-CONCEPT 2

d. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1CNCPT2 SELF-CONCEPT 2, F1LOCUS2 LOCUS OF CONTROL 2

Figure 5.4 Sequential regression of Social Studies test scores on SES, Previous Grades, Self-Esteem, and Locus of Control.

Are these increases in explained variance statistically significant? The formula we use is a simple extension of one of our earlier formulas:

$$F = \frac{R_{12}^2 - R_1^2 / k_{12} - k_1}{1 - R_{12}^2 / (N - k_{12} - 1)}$$

In other words, we subtract the R^2 from the equation with fewer variables from the R^2 with more variables (ΔR^2). This is divided by the unexplained variance (from the equation with more variables). The numerator uses the change in degrees of freedom (which is often 1), and the denominator uses the degrees of freedom for the equation with more variables. As in the earlier simultaneous regression, $N = 887$.

We'll use the formula to calculate the F associated with the final step of the sequential multiple regression:

$$\begin{aligned} F &= \frac{R_{1234}^2 - R_{123}^2 / k_{1234} - k_{123}}{1 - R_{1234}^2 / (N - k_{1234} - 1)} \\ &= \frac{.339 - .333 / 1}{(1 - .339) / (887 - 4 - 1)} \\ &= \frac{.006}{.661 / 882} = 8.006 \end{aligned}$$

which matches the value shown in the figure (7.918), within errors of rounding. In other words, the addition of Locus of Control to the equation leads to an increase in R^2 of .006, or a 6/10 of 1% increase, in explained variance. This seemingly tiny increase in variance explained is statistically significant, however ($F = 7.918 [1, 882], p = .005$).

Of course, we can also test the overall regression equation, with all variables entered, for statistical significance. The overall $R^2 = .339, F = 112.846 [4, 882], p < .001$, which is the same result we got with the simultaneous regression.

Comparison to Simultaneous Regression

It will be instructive to compare the results of our sequential regression to those of the simultaneous regression using the same variables (Figures 5.3 versus 5.4). One of the most striking differences is that for the simultaneous regression Self-Esteem was not statistically significant, whereas for the sequential regression it was statistically significant ($\Delta R^2 = .005, F = 6.009 [1, 883], p = .014$). Why do we get different answers with the different methods? The second difference is that the magnitude of effects, as suggested by the ΔR^2 's in the sequential regression, seems so different and so much smaller than the effects suggested by the β 's in the simultaneous regression. In the simultaneous regression, for example, we found a small to moderate effect for Locus of Control on the Achievement tests ($\beta = .097$), but in the sequential regression, Locus of Control accounted for only a .6% increase in the variance explained in Social Studies achievement, a seemingly minuscule amount. We will deal with the first problem (statistical significance) first and with the second issue (magnitude of effects) following.

The Importance of Order of Entry

As you will soon discover, the statistical significance (and the apparent magnitude of effect) of the variables in a sequential regression depends on their order of entry into the equation. Look at Figure 5.5. In this sequential regression, the first two variables were entered in the same order, but Locus of Control was entered at step 3 and Self-Esteem at step 4. With this order of entry, the Self-Esteem variable was again not statistically significant ($p = .663$). The primary reason that Self-Esteem was statistically significant in one sequential regression and not the other is, of course, the difference in variance accounted for by Self-Esteem in one regression versus the other ($\Delta R^2 = .005$ in Figure 5.4 versus .001 in Figure 5.5).

Next, focus on Figure 5.6. For this regression, the order of entry into the sequential regression was Locus of Control, Self-Esteem, previous Grades, and SES. Notice the drastic change in the variance accounted for by the different variables. When entered first in the regression equation, SES accounted for 18.5% of the variance in Achievement (Figures 5.3 and 5.4), but when entered last, SES only accounted for 7.2% of the variance (Figure 5.6). The bottom line is this: with sequential multiple regression, the variance accounted for by each independent variable (i.e., ΔR^2) changes depending on the order of entry of the variables in the regression equation. Because the ΔR^2 changes depending on order of entry, variables will sometimes switch from being statistically significant to being not significant, or vice versa. As our results have shown, when variables are entered earlier in the regression will generally show larger explained variance than if they are entered later in the regression. Again, we have encountered a disconcerting discrepancy.

If the order of entry makes such a big difference in sequential regression results, what then is the *correct* order of entry? A cynical, unethical answer might be to enter the variables you want to show as important first in the regression equation, but this is an indefensible and inferior solution. What are the options? What was my thinking for various orders of entry in Figures 5.4 through 5.6? One common and defensible solution is to input the variables in order of presumed or actual *time precedence*. This was my thinking for the first example. SES, a parent variable largely in place when many children are born, should logically precede the

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	BYSES SOCIO-ECONOMIC STATUS COMPOSITE ^a	.	Enter
2	BYGRADS GRADES COMPOSITE ^a	.	Enter
3	F1LOCUS2 LOCUS OF CONTROL 2 ^a	.	Enter
4	F1CNCPT2 SELF-CONCEPT 2 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Model Summary

Added to the Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
SES	.430 ^a	.185	.185	200.709	1	885	.000
Previous Grades	.573 ^b	.328	.143	188.361	1	884	.000
Locus of Control	.582 ^c	.338	.010	13.797	1	883	.000
Self-Esteem	.582 ^d	.339	.001	.190	1	882	.663

a. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE

b. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE

c. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1LOCUS2 LOCUS OF CONTROL 2

d. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1LOCUS2 LOCUS OF CONTROL 2, F1CNCPT2 SELF-CONCEPT 2

Figure 5.5 Sequential regression of Social Studies test scores on SES, Previous Grades, Locus of Control, and Self-Esteem. With sequential regression, the order of entry of the variables affects their apparent importance.

Model Summary

Added to the Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
Locus of Control	.248 ^a	.061	.061	57.867	1	885	.000
Self-Esteem	.250 ^b	.063	.001	1.099	1	884	.295
Previous Grades	.517 ^c	.267	.204	246.158	1	883	.000
SES	.582 ^d	.339	.072	95.495	1	882	.000

a. Predictors: (Constant), F1LOCUS2 LOCUS OF CONTROL 2

b. Predictors: (Constant), F1LOCUS2 LOCUS OF CONTROL 2, F1CNCPT2 SELF-CONCEPT 2

c. Predictors: (Constant), F1LOCUS2 LOCUS OF CONTROL 2, F1CNCPT2 SELF-CONCEPT 2, BYGRADS GRADES COMPOSITE

d. Predictors: (Constant), F1LOCUS2 LOCUS OF CONTROL 2, F1CNCPT2 SELF-CONCEPT 2, BYGRADS GRADES COMPOSITE, BYSES SOCIO-ECONOMIC STATUS COMPOSITE

Figure 5.6 Sequential regression of Social Studies test scores on Locus of Control, Self-Esteem, Previous Grades, and SES. Again, the order of entry makes a big difference in the apparent effects of the variables with sequential regression.

student variables, measured in 8th and 10th grades. Previous Grades, from grades 6 through 8, is prior to both Self-Esteem and Locus of Control measured in 10th grade. Self-Esteem and Locus of Control are a little more difficult, but it seems to me that conceptions of one's worth should come about and thus be causally prior to conceptions of internal versus external control. It would also be possible to argue that one's conception of internal versus external control may be prior to feelings of self-worth, reasoning that was operationalized in the second sequential regression example of Figure 5.5. Beyond actual time precedence and logic, previous research can also help make such decisions. I know of one study that tested these competing hypotheses, and it supported either a reciprocal relation or self-esteem as prior to locus of control (Eberhart & Keith, 1989).

For Figure 5.6, variables were entered in possible *reverse* time precedence, with strikingly different findings. And there are undoubtedly other methods for deciding the order of entry: perceived importance, background variables versus variables of interest, static versus manipulable variables, and so on. But again, which method is correct? And why is order so important?

Why Is Order of Entry So Important?

One way of understanding why different orders of entry make such a difference in findings is through a return to Venn diagrams. Figure 5.7 shows the relations among three hypothetical variables: a dependent variable Y , and two independent variables X_1 and X_2 . The areas of overlap represent the shared variance among the three variables. The shaded area of overlap marked 1 (including area 3) represents the variance shared by X_1 and Y , and the shaded area marked 2 (including area 3) represents the variance shared by X_2 and Y . But these variances overlap, and area 3 represents the variance shared by all three variables. It is the area of double shading (3) that is treated differently depending on the order of entry of the variables in sequential regression. If X_1 is entered first into the equation to predict Y , this variance (area 1 and area 3) is attributed to variable X_1 . When variable X_2 is added, the ΔR^2 is equal to area 2 (excluding area 3). In contrast, if Y is first regressed on variable X_2 , then both areas 2 and 3 will be attributable to variable X_2 , and when variable X_1 is subsequently added, only the

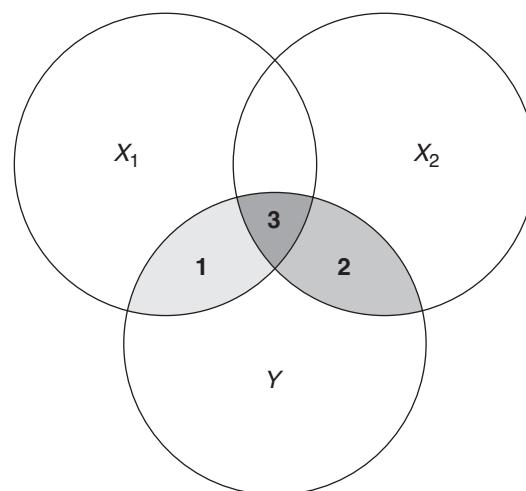


Figure 5.7 Venn diagram illustrating why the order of entry is so important in sequential regression. The variance shared by all three variables (area 3) is attributed to whichever variable is first entered in the MR.

variance of area 1 (excluding area 3) will be attributed to it. This heuristic aid helps explain *why* order of entry makes a difference, but the question of which order is correct is still not answered.

Total Effects

It is again useful to turn to path diagrams to further understand why we get a difference depending on order of entry and to help us to understand the proper order of entry. Figure 5.8 shows a model that represents the ordering used for the initial sequential analysis (Figure 5.4). As discussed in previous chapters, the regression coefficients from the simultaneous multiple regression in fact are estimates of the direct paths, or direct effects, to the final outcome, Social Studies Achievement. These paths are marked *a*, *b*, *c*, and *d*, and we could simply insert the standardized or unstandardized coefficients from the simultaneous regression (Figure 5.3) in place of these letters. Simultaneous regression estimates the direct effects in such models.

But we chose a particular order of entry in the sequential regression. SES was entered first, then Previous Grades, then Self-Esteem, and then Locus of Control. This ordering is represented in two ways in Figure 5.8: first, in the left-to-right sequencing of the variables, and second, by the dashed arrows from one variable to the next suggesting that SES comes first, followed by Previous Grades, and so on.

Figure 5.9 shows an even more complete representation of our sequential regression. The model still includes the direct effects (paths *a* through *d*), and, as we will see, these are, in fact, estimated in the last block of the sequential regression. Note that this model, unlike those in previous chapters, also includes indirect pathways from earlier to later variables. That is, there are also *indirect effects* in this model. Thus, in addition to the possible direct effect of Self-Esteem on Achievement (path *a*), Self-Esteem also has an indirect effect on Achievement, through Locus of Control, symbolized by the heavier arrows in the figure (paths *e* and *d*). If we were to estimate path *e*, we could actually multiply path *e* times path *d* to produce an estimate of this indirect effect. And we could also sum the direct and indirect effects of Self-Esteem on Achievement to estimate the *total effect* of Self-Esteem on Achievement.

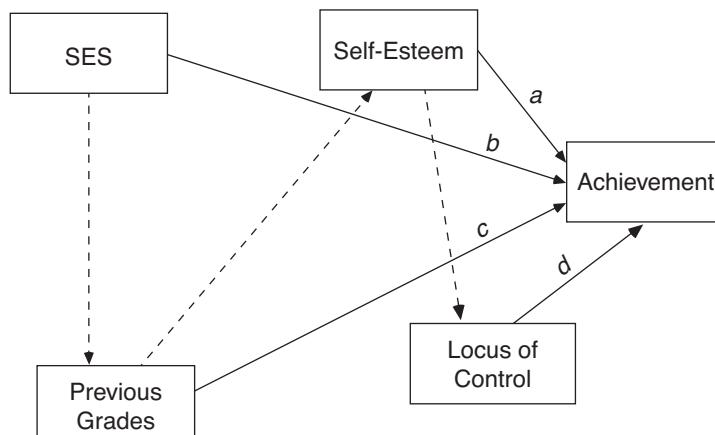


Figure 5.8 Path representation of a sequential regression. The model shows the sequencing of the first sequential regression (the regression in Figure 5.4).

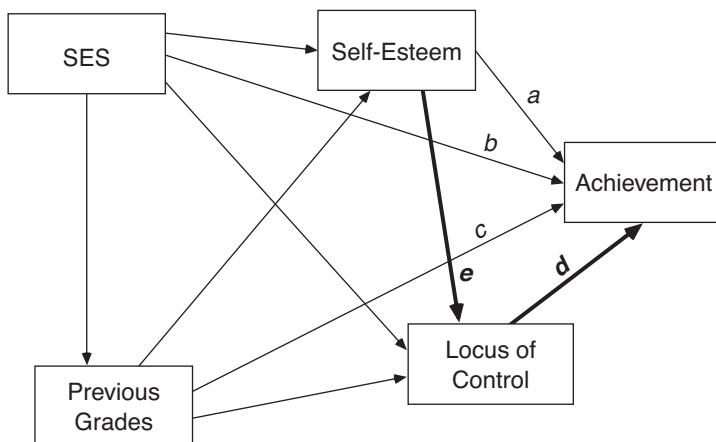


Figure 5.9 A more complete version of the initial sequential regression model.

As it turns out, sequential multiple regression estimates the *variance* accounted for by these *total effects*. (Note that the ΔR^2 's do *not* estimate the total paths directly but rather the variance attributable to these total effects.) Thus the reason that sequential and simultaneous multiple regression may give us different answers concerning the importance of variables is that they focus on two different aspects of the multiple regression. Simultaneous regression focuses on estimates of *direct effects*, whereas sequential regression focuses on the variance accounted for by *total effects*.

Variables entered first in a sequential regression have larger effects, other things being equal, than do variables entered later in the equation, because there are many more (indirect) ways variables entered early can affect the final outcome. Thus SES, for example, seems to have a relatively large effect when entered first in a sequential regression, because it can thus affect Social Studies Achievement through Grades, Self-Esteem, and Locus of Control.

Armed with this understanding, the question of the proper order of entry becomes clearer. Whether we realize it or not, *any time* we use sequential regression in an explanatory manner, we have implied a model such as that shown in Figure 5.9! The *proper* order of entry is the order implied by our models, assuming those models are set up correctly. Thus, if you use sequential regression, you had better first spend some time thinking through the model that underlies your analysis. If you do not think through your model correctly, your analysis will produce inaccurate estimates of the effects of one variable on another. When variables are entered prematurely, sequential regression will overestimate their effects. The effects of variables that are entered later than they should be in the analysis will be underestimated. If you use sequential regression, you should be prepared to defend the model underlying this regression. As you read the results of others' sequential regressions, you should sketch out the models underlying these regressions to make sure they are reasonable.

Problems With R^2 as a Measure of Effect

We have discussed the problem of variables changing in apparent importance depending on their order of entry in a sequential regression. Let's now return to our second concern: why do all the effects in sequential regression appear so much smaller than in simultaneous regression (e.g., Figure 5.3 versus 5.4)? The reason, of course, is that in simultaneous

regression we focus on the regression coefficients as the measure of the importance of effects, whereas with sequential regression, we focus on the ΔR^2 's, the increments to the explained variance, as indicators of the importance of effects. And although variances have a number of desirable properties—variances are easy to work with in formulas, and they are a familiar metric—explained variance is a stingy and misleading measure of the effect of one variable on another (Rosenthal & Rubin, 1979; Schmidt & Hunter, 2014, chap. 5). There are many possible examples of this truism. For example, everyone knows the importance of smoking on lung cancer; smoking is one of the primary causes of lung cancer. How much variance, then, do you think smoking explains in lung cancer: 30%? 50%? No matter what your answer, you will be surprised: smoking explains 1% to 2% of the variance in lung cancer (Gage, 1978)! The point is not that smoking is unimportant; the point is that this seemingly small amount of variance is important.

What statistic comes closer to representing “importance” than R^2 ? Darlington and Hayes (2017) suggested using the unsquared metric, rather than the squared metric. It does make a difference. A multiple correlation of .40 is twice as large as a multiple correlation of .20, but if we squared these correlations, the first accounts for four times the variance as the second (.16 versus .04). In sequential multiple regression, the unsquared counterpart to ΔR^2 is $\sqrt{\Delta R^2}$ (which is not the same as ΔR). $\sqrt{\Delta R^2}$, as it turns out, is equal to what is known as the *semipartial correlation* of Y with X , controlling for the other variables in the equation. Conceptually, a semipartial correlation is the correlation of Y with X_1 , with the effects of X_2 , X_3 , and so on, removed from X_1 . It may be symbolized as $sr_{y(1-23)}$, with the parentheses showing that the effects of X_2 and X_3 are removed from X_1 but not from Y . (Partial and semipartial correlations are presented in more depth in Appendix C.) Turning to the sequential regression from Figure 5.4, the $\sqrt{\Delta R^2}$'s will equal .430, .378, .071, and .077 for SES, Grades, Self-Esteem, and Locus of Control. These values are at least a little more consistent with the β 's from Figure 5.3 than are the ΔR^2 's.

There is another reason for preferring $\sqrt{\Delta R^2}$ to ΔR^2 (and R to R^2 , and r to r^2 , and so on): these unsquared coefficients generally come closer to representing most definitions of “importance” than do the squared coefficients (Darlington & Hayes, 2017). Darlington and Hayes provided several useful examples. To use one example (p. 216), suppose I ask you to flip two coins, a dime and a nickel. If either coin comes up heads, I will pay you the amount of that coin (10 cents or 5 cents). Over a long series of coin flips, you will earn 5 cents 25% of the time (i.e., nickel = heads, dime = tails), 10 cents 25% of the time, 15 cents 25% of the time, and nothing 25% of the time (dime and nickel both = tails). Obviously, the dimes are twice as important in determining your earnings as the nickels, since dimes are worth twice as much as nickels. If you conduct a multiple regression of this problem, regressing your earnings on the results of each coin (heads = 1 and tails = 0; these are examples of *dummy* variables to be discussed in later chapters), the ΔR^2 associated with dimes will not be twice as large as that associated with nickels, but *four times as large* (ΔR^2 for nickels = .20, ΔR^2 for dimes = .80). The $\sqrt{\Delta R^2}$'s put importance back in the proper metric, however. The $\sqrt{\Delta R^2}$ associated with dimes (.894) is, in fact, twice as large as that associated with nickels (.447). These data are summarized in Table 5.2. Dimes are twice as important as nickels in determining the amount of money received; $\sqrt{\Delta R^2}$ demonstrates this importance, but R^2 does not. See Darlington and Hayes for other examples, as well. In sequential multiple regression (and in other types of regression), the unsquared coefficient generally provides a better indicator of importance than does the squared coefficient. We will still test the statistical significance of ΔR^2 , but if we are interested in comparing the magnitude of effects, we will use $\sqrt{\Delta R^2}$.¹

Table 5.2 Comparison of ΔR^2 versus $\sqrt{\Delta R^2}$ as Measures of the Importance of Effects

<i>Measure of Importance</i>	<i>Importance of Nickels</i>	<i>Importance of Dimes</i>
ΔR^2	.200	.800
$\sqrt{\Delta R^2}$.447	.894

Cohen's f^2 as a Measure of Effect

As noted in Chapter 4, f^2 is another common measure of effect size. Just as R^2 may be converted to f^2 , so may ΔR^2 :

$$f^2 = \frac{R_{y,12}^2 - R_{y,1}^2}{1 - R_{y,12}^2}$$

Note, of course, that f^2 , because it is a squared metric, suffers from the same problems as a measure of "importance" as does ΔR^2 .

Other Uses of Sequential Regression

There are other ways of conducting sequential multiple regression; the method I have outlined here is, in my experience, far and away the most common method. You may also see this method referred to as *variance partitioning* (Pedhazur, 1997) or *sequential variance decomposition* (Darlington, 1990).

Interpretation of Regression Coefficients

It is also possible to use regression coefficients from each step of the sequential regression as estimates of the *total* effects of each variable on the outcome. In this case, we would use the b or β associated with the variable entered at that step as the estimate of the total effect, ignoring the coefficients for variable entered at earlier steps in the equation. For example, Figure 5.10 shows a table of such coefficients as generated by SPSS. The relevant coefficients are in italic boldface; these are the coefficients that I would report in a write-up of the research, perhaps accompanied by a table such as Table 5.3. I have rarely seen this approach used outside of path analysis, however, so we will discuss it in more detail when we get to that topic (see Chapter 12).

Table 5.3 Total Effects of SES, Previous Grades, Self-Esteem, and Locus of Control on 10th-Grade Social Studies Achievement, Estimated through Sequential Regression

<i>Variable</i>	<i>b (SE_b)</i>	<i>β</i>
SES	5.558 (.392)**	.430
Previous Grades	5.420 (.395)**	.400
Self-Esteem	1.016 (.414)*	.069
Locus of Control	1.554 (.552)**	.097

* $p < .01$.

** $p < .05$.

Model		Coefficients ^a				
		B	Std. Error	Beta	t	Sig.
1	(Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE	51.090 5.558	.299 .392		170.745 .430	.000 .000
2	(Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE BYGRADS GRADES COMPOSITE	34.793 3.875 5.420	1.218 .377 .395		28.561 10.280 13.724	.000 .000 .000
3	(Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE BYGRADS GRADES COMPOSITE F1CNCPT2 SELF-CONCEPT 2	35.138 3.798 5.291 1.016	1.223 .377 .397 .414		28.734 10.068 13.318 .069	.000 .000 .000 .014
4	(Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE BYGRADS GRADES COMPOSITE F1CNCPT2 SELF-CONCEPT 2 F1LOCUS2 LOCUS OF CONTROL 2	35.517 3.690 5.150 .218 1.554	1.226 .378 .399 .501 .552		28.981 9.772 12.910 .436 .097	.000 .000 .000 .663 .005

a. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Figure 5.10 Sequential regression used to estimate the total effects of each variable on the outcome. The italicized boldface coefficients are estimates of total unstandardized and standardized effects.

Block Entry

It is possible to enter groups of variables in blocks, or groups of variables, as well as one at a time. A primary reason for entering variables in blocks might be to estimate the effect of a type or category of variables on the outcome. Using our current example, we might be interested in the effect of the psychological variables together, and above and beyond the effect of the background variables, on Achievement. If this were our interest, we could enter the two background variables sequentially, followed by the two psychological variables in a block. (It would also be possible to enter the two background variables in one block and the two psychological variables in a second block.) The statistical significance of the resulting ΔR^2 associated with the second block could be examined to determine whether these psychological variables explained statistically significantly more variance, and the resulting $\sqrt{\Delta R^2}$ could be examined to determine the relative importance of the effect of these psychological variables. Some of the output from such an analysis is shown in Figure 5.11. These results suggest that the two psychological variables, in combination, are important for Social Studies

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	BYSES SOCIO-ECONOMIC STATUS COMPOSITE ^a	.	Enter
2	BYGRADS GRADES COMPOSITE ^a	.	Enter
3	F1CNCPT2 SELF-CONCEPT 2, F1LOCUS2 LOCUS OF CONTROL 2 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Model Summary

Added to the Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
SES	.430 ^a	.185	.185	200.709	1	885	.000
Previous Grades	.573 ^b	.328	.143	188.361	1	884	.000
Self-Esteem & Locus of Control	.582 ^c	.339	.010	6.987	2	882	.001

a. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE

b. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE

c. Predictors: (Constant), BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYGRADS GRADES COMPOSITE, F1CNCPT2 SELF-CONCEPT 2, F1LOCUS2 LOCUS OF CONTROL 2

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE	51.090	.299		170.745	.000
	5.558	.392	.430	14.167	.000
2 (Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE BYGRADS GRADES COMPOSITE	34.793	1.218		28.561	.000
	3.875	.377	.300	10.280	.000
	5.420	.395	.400	13.724	.000
3 (Constant) BYSES SOCIO-ECONOMIC STATUS COMPOSITE BYGRADS GRADES COMPOSITE F1CNCPT2 SELF-CONCEPT 2 F1LOCUS2 LOCUS OF CONTROL 2	35.517	1.226		28.981	.000
	3.690	.378	.285	9.772	.000
	5.150	.399	.380	12.910	.000
	.218	.501	.015	.436	.663
	1.554	.552	.097	2.814	.005

a. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Figure 5.11 Self-Concept and Locus of Control entered as a block in a sequential regression.

Achievement (the example will be interpreted in more detail later in this chapter). In essence, this example illustrates one possible combination of sequential and simultaneous regression. Another possible reason for entering variables in blocks is if you were unsure of the proper order of some of your variables. Using our current example, if we could not decide whether

Self-Esteem should follow or precede Locus of Control, we might enter the two variables in the same block.

Unique Variance

Another use of sequential regression is for researchers who wish to isolate the unique variance in a dependent variable accounted for by each variable in a regression, after taking all other variables into account. Return to the Venn diagram in Figure 5.5. If you are interested in the unique variance attributable to a variable, you will be interested in variance associated with area 1 (excluding area 3) as the unique variance attributable to variable X_1 and the variance associated with area 2 (excluding area 3) as the unique variance of variable X_2 . Conceptually, this approach is like conducting a series of sequential regressions, entering each variable last in one of these equations. In practice, isolating the unique variance for a variable can be accomplished this way, but there are simpler methods. If your primary interest is the statistical significance of each unique variance, it is simpler to conduct a simultaneous regression. The statistical significance of the regression coefficients is equal to the statistical significance of the ΔR^2 's with each variable entered last in the regression equation. You can demonstrate this to yourself by comparing the statistical significance for the ΔR^2 from the last variable entered in Figures 5.4 through 5.6 with that of the regression coefficients from Figure 5.3. So, for example, with the simultaneous regression, Self-Esteem had a probability of .663. When entered last in a sequential regression (Figure 5.5), Self-Esteem had a probability of .663. Likewise, in Figure 5.3, the t associated with Self-Esteem was .436; in Figure 5.5, the F was .190 (recall that $t^2 = F$). Compare the t and p associated with SES in Figure 5.3 with the F and p associated with SES when it was entered last in a sequential regression (Figure 5.6). The equivalence of the statistical significance of variables in simultaneous regression with variables entered last in sequential regression will prove useful in later chapters.

If you are interested in the *values* of $\sqrt{\Delta R^2}$ for each variable when entered last in the regression equation, recall that these are equal to the semipartial correlations. Thus you can conduct a simultaneous regression requesting the semipartial correlations (also called *part* correlations). The last column of the table in Figure 5.12, for example, shows the semipartial correlations of each variable with the Social Studies test, controlling for all other variables (SPSS labels these as part correlations). So, for example, the value of the "part" correlation shown in Figure 5.12 for SES is .268. The value shown for the ΔR^2 for SES when it was entered last in the equation was .072 (Figure 5.6). And $\sqrt{\Delta R^2} = \sqrt{.072} = .268$.

Model	Coefficients ^a							
	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	Correlations		
	B	Std. Error				Zero-order	Partial	Part
1 (Constant)	35.517	1.226		28.981	.000			
BYSES	3.690	.378	.285	9.772	.000	.430	.313	.268
SOCIO-ECONOMIC STATUS COMPOSITE	5.150	.399	.380	12.910	.000	.498	.399	.354
BYGRADS GRADES COMPOSITE	.218	.501	.015	.436	.663	.173	.015	.012
F1CNCPT2	1.554	.552	.097	2.814	.005	.248	.094	.077
SELF-CONCEPT 2								
F1LOCUS2 LOCUS OF CONTROL 2								

a. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Figure 5.12 Semipartial (part) correlations of each variable with the Social Studies Achievement outcome. Semipartial correlations are equal to $\sqrt{\Delta R^2}$ with each variable entered last in a sequential regression.

If the program you use does not easily produce semipartial correlations but you want information about unique variance, you can get this information by conducting a series of sequential regressions, entering, in turn, each variable last in the equation. I will refer to this approach as *sequential unique regression*.

Interactions and Curves

Finally, we can use sequential regression to test for interactions and curves in the regression line by adding these terms last in a sequential regression. This is a common use and one we will discuss in later chapters.

Interpretation

Throughout this book, I've interpreted the results of a number of simultaneous regressions. Here is a brief interpretation of a sequential regression, one that also illustrates a plausible use of the methodology. For this analysis, we'll use the analysis and output from Figure 5.11. Here's a possible interpretation:

The purpose of this research was to determine whether students' psychological characteristics have an effect on high school students' social studies achievement, even after controlling for the effects of relevant background variables. To accomplish this purpose, students' scores on a 10th-grade standardized social studies (history, citizenship, geography) were regressed on SES, previous (8th grade) Grades, and two psychological variables, Locus of Control and Self-Esteem, using a sequential multiple regression analysis.

The results of the analysis are shown in Table 5.4. The first background variable entered in the regression, SES, resulted in a statistically significant increase in explained variance ($\Delta R^2 = .185$, $F[1, 885] = 200.709$, $p < .001$), as did the second background variable entered into the regression equation, Previous Grades ($\Delta R^2 = .143$, $F[1, 884] = 188.361$, $p < .001$). Of greater interest are the results of the third step of the sequential regression. In this step, the psychological variables of Locus of Control and Self-Esteem were entered as a block. As shown in the table, these psychological variables explained a statistically significant increase in the variance of Achievement ($\Delta R^2 = .010$, $F[2, 882] = 6.987$, $p = .001$). These findings suggest that personal, psychological variables may indeed be important for students' high school achievement. If so, focusing on high school students' psychological well-being may be important for their achievement as well as their wellbeing.

Table 5.4 Effects of SES, Previous Grades, Self-Esteem, and Psychological Characteristics on 10th-Grade Social Studies Achievement

<i>Block</i>	ΔR^2	<i>Probability</i>
1 SES	.185	<.001
2 Previous Grades	.143	<.001
3 Locus of Control and Self-Esteem	.010	.001

In this interpretation, I could have included in the table the semipartial correlations (or $\sqrt{\Delta R^2}$) or the β 's from each block, but without a discussion and interpretation of total effects, I think these statistics would be more misleading than illuminating. It is not uncommon for researchers to report also the β 's from the final step of the regression.

Summary: Sequential Regression

Analysis

With sequential regression, variables are added one at a time or in blocks. The order of entry of the variables should be consistent with an underlying causal model, or the results will not provide accurate estimates of the effects of the variables on the outcome.

Purpose

The primary purpose of sequential regression is explanation. A researcher is interested in determining which variables are important influences on some outcome. Given the adequacy of the underlying causal model, one can also use sequential regression to determine the extent of the total influence of each variable on the outcome. This use requires the use of some of the b 's or the β 's from each block in the analysis, however, rather than the more common use of ΔR^2 ; it will be presented in more detail in Part 2 when we discuss path analysis (Chapter 12). Sequential unique regression can also be used to determine the unique contribution of each variable to some outcome, after controlling for the other variables in a model.

Sequential regression can also be used in the service of prediction, for example, to determine which variables are statistically significant predictors of some outcome. You may also be interested in rank ordering the importance of predictors. In this case, order of entry makes a difference, so the best approach is to focus on semipartial correlations or to add each variable last in the equation to determine its unique contribution to the prediction (sequential unique regression). For both of these purposes, however, simultaneous regression may accomplish the same goals more simply. The danger of using sequential regression for prediction is that you or the readers of your research may be sorely tempted to interpret the results in an explanatory fashion. Remember that any time you start thinking along the lines of "this means that if we were to increase X , then Y would increase" you have crossed the line from prediction to explanation.

What to Interpret

In sequential regression, we generally focus on the statistical significance of the change in explained variance (ΔR^2) as the measure of the statistical significance of each variable. It is common to see ΔR^2 also used as an indicator of the importance of each variable, but, as we have seen, $\sqrt{\Delta R^2}$ is a better measure of importance. In addition, any reference to the relative importance of variables in sequential regression is implicitly or explicitly based on a causal model. It is also possible to interpret the regression coefficients associated with the variables entered at each block of a sequential regression.

The exception to the rule that sequential regression requires an implicit causal model is when each variable is added last to the equation to determine each variable's unique contribution to the outcome variable (sequential unique regression). This approach is analogous to simultaneous regression.

Strengths

If based on a defensible model, sequential regression can provide good estimates of the *total* effects of a series of variables on some outcome (although examining b or β rather than ΔR^2). Sequential regression, with its focus on change in explained variance, may be more comfortable than simultaneous regression for those more familiar with ANOVA methods. Sequential regression is useful for determining whether some new variable improves the prediction of some outcome over and above an existing set of variables; we will use sequential regression in this fashion to test the statistical significance of interaction terms and curve components.

Weaknesses

Sequential regression will give different estimates of the importance of variables in the regression depending on the order of entry of these variables. Other things being equal, variables entered earlier in a sequential regression will appear more important than those entered later. This is because sequential regression estimates total effects, including indirect effects through variables entered later in an analysis. If not based on an implicit, reasonable model sequential regression can give misleading estimates of effects. In my experience, the use of such models underlying sequential regression is rare. Sequential regression will underestimate effects entered too late and overestimate the effects of variables entered too early.

Conclusion

As you can see, I have suggested fairly constrained uses for sequential multiple regression: testing the statistical significance of curves and interactions (discussed in more detail in Chapters 7 and 8), testing whether single variables or blocks of variables are important additions to a regression equation, and for calculating total effects within a causal model (discussed in more detail in Part 2). Simultaneous regression is generally my default regression approach for most problems. Why, then, have I spent so much time on the topic? The primary reason is that, depending on your area of research interest, you may encounter sequential regression commonly in your research reading. Different research traditions have different norms as to what methods are most common. Unfortunately, many such presentations will use sequential regression poorly and in ways I have argued against in this chapter. I have here tried to present the most common uses of sequential regression and explain why some are appropriate and others are not. Given the overlap between simultaneous and sequential regression, it is also possible to use sequential regression as your default method, and this may well be the norm in your area of research. The important point is to understand how the two methods relate to one another and the degree to which they focus on different aspects of the regression approach.

STEPWISE MULTIPLE REGRESSION

In your reading, you may encounter a multiple regression variation called stepwise regression (or one of its variations, e.g., forward selection or backward elimination). Unlike simultaneous or sequential regression, stepwise multiple regression should be used only for prediction. Unfortunately, because of its apparent ease, stepwise regression is often used in attempts at explanation. I will admonish you over and over not to make this mistake and will generally discourage the use of stepwise methods. The presentation of stepwise regression will follow the format used for sequential regression, with an extended discussion followed by a summary.

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	BYGRADS GRADES COMPOSITE	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	BYSES SOCIO-ECONOMIC STATUS COMPOSITE	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	F1LOCUS2 LOCUS OF CONTROL 2	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Model Summary

Added to the Model	R	R Square	Adjusted R Square	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
Previous Grades	.498 ^a	.248	.247	.248	291.410	1	885	.000
SES	.573 ^b	.328	.327	.080	105.682	1	884	.000
Locus of Control	.582 ^c	.338	.336	.010	13.797	1	883	.000

a. Predictors: (Constant), BYGRADS GRADES COMPOSITE

b. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYSES SOCIO-ECONOMIC STATUS COMPOSITE

c. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, F1LOCUS2 LOCUS OF CONTROL 2

ANOVA^d

Model	Sum of Squares	df	Mean Square	F	Sig.	
1	Regression	21357.16	1	21357.164	291.410	.000 ^a
	Residual	64860.76	885	73.289		
	Total	86217.92	886			
2	Regression	28283.26	2	14141.630	215.781	.000 ^b
	Residual	57934.66	884	65.537		
	Total	86217.92	886			
3	Regression	29174.59	3	9724.864	150.536	.000 ^c
	Residual	57043.33	883	64.602		
	Total	86217.92	886			

a. Predictors: (Constant), BYGRADS GRADES COMPOSITE

b. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYSES SOCIO-ECONOMIC STATUS COMPOSITE

c. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, F1LOCUS2 LOCUS OF CONTROL 2

d. Dependent Variable: F1TXHSTD HIST/CIT/GEOG STANDARDIZED SCORE

Figure 5.13 Stepwise regression of Social Studies Achievement on SES, Previous Grades, Self-Esteem, and Locus of Control.

The Analysis

Stepwise multiple regression is similar to sequential regression in that predictor variables are entered one at a time in a sequential order. The difference is that with stepwise multiple regression the computer chooses the order of entry, rather than the researcher.

Figure 5.13 shows the primary output from a stepwise regression using the variables and data used throughout this chapter. Previous Grades were entered at step 1, SES at step 2, and Locus of Control at step 3. Note from the model summary table that each of these additions to the equation resulted in a statistically significant increase in ΔR^2 . Self-Esteem, in contrast, was not added to the equation because its addition would not have led to a statistically significant increase in R^2 .

How Are Variables Added to the Equation?

As shown in the last column of the table labeled Variables Entered/Removed, variables are entered into the equation if the probability associated with the ΔR^2 is less than .05. If this were the only way variables could be included in the equation, we would call this *forward entry* (stepwise) regression. It is also possible, however, that the variance of a variable entered at one step of the equation is reproduced by that of several variables entered in later steps of the equation. If this occurs (it does not in the current example) and the p associated with the earlier-entered variable increased to .10 or greater, this variable would be dropped from the regression equation. This is, of course, more likely in problems with many possible predictor variables. If this were the only approach to be used (i.e., all variables entered and the statistically-not-significant ones dropped), we would be conducting a *backward elimination* (stepwise) regression. The term *stepwise regression* usually refers to a combination of these two methods, but is also used to refer to the forward entry method alone. The probability values for entry and removal can be changed. It is also possible to limit the number of steps; we could have set a maximum of two steps, for example, thus allowing only Grades and SES to enter the equation.

How Does the Program Decide What Variable to Add at Each Step?

The first variable to enter is the variable with the largest correlation with the outcome variable. In the current example, Grades had the largest correlation with Social Studies Achievement and was the first variable to enter the equation. The program then calculates the semipartial correlation of each remaining variable with the outcome, controlling for the variable(s) already in the equation, and the variable with the largest semipartial correlation with the outcome is entered next. In the current example, SES had the largest semipartial correlation with Social Studies, after controlling for Grades. Said differently, the variable that will lead to the largest increase in ΔR^2 will be added next to the regression equation. The program then continues to cycle through these steps—add a variable, compute semipartial correlations of the excluded variables controlling for the entered variables—until no more excluded variables fulfill the requirement for entry, or the maximum number of steps is reached. Since we know that the squared semipartial correlations are equal to the ΔR^2 , this process is the same as calculating the possible ΔR^2 for each variable at each step.

Danger: Stepwise Regression Is Inappropriate for Explanation

This sounds great, doesn't it? No more need to do the hard work of thinking through models, no more embarrassment if these models are proved wrong! All you have to do is decide which variables to include in the analysis, not which are important. Just let the computer decide! There are no substitutes for the hard work, however. Stepwise regression may indeed help you determine a useful subset of variables for predicting some outcome (and we will even question this statement later), but that is all. Stepwise regression cannot tell you which variables influence some outcome; to decide this, you must start with a defensible, theory and research-derived notion of the plausible influences on some outcome, what we have been calling a model. Even with such a model, what a proper, explanatory, regression analysis reveals is the extent of the influence of one variable on another, *given the adequacy of your model*. In other words, an implicit or explicit model is required for explanatory interpretation of multiple regression, and such models do not come from statistics programs, but from knowledge of theory and research on a topic, combined with careful thought.

Perhaps the answer is to start with an informal explanatory model, one that includes the important, relevant variables, and then conduct a stepwise regression. This technique does

not help either, since stepwise regression does nothing to tell you the proper order of the variables in the model. Yes, stepwise regression orders the variables but only in reference to the degree to which the variables sequentially explain variance. This ordering may be entirely different from the *causal* ordering of variables. Note that the stepwise results in Figure 5.13 don't tell us the proper time precedence of the variables (compare the results in Figure 5.13 with the model in Figure 5.9). And we can get even more ridiculous. We could just as easily regress Previous Grades on SES, self-perceptions, and 10th-grade Social Studies Achievement, and the stepwise regression would dutifully tell us that Social Studies achievement was the best single predictor of Grades, followed by SES, and so on, even though our predictor (Social Studies Achievement) happened after the criterion (Grades)! The bottom line is this: stepwise regression results do not help us understand how variables affect an outcome. For these reasons, methodologists routinely condemn stepwise regression as an explanatory method: "variables mindlessly enter into the analysis in the absence of theory and the results, therefore, are theoretical garbage" (Wolfle, 1980, p. 206). To make this point in class, I tell my students, tongue in cheek, that stepwise regression is a tool of the devil. Do *not* use stepwise regression if you wish to understand the influence of a group of variables on an outcome; do *not* use stepwise regression if you wish to make policy or intervention recommendations based on your results.

A Predictive Approach

What can stepwise regression tell you, then? Stepwise regression can tell you which subset of a group of predictors may be used to predict some criterion. It may be used to develop an equation to predict some criterion, using a given group of predictors. Stepwise regression may be used for prediction. Several examples will help illustrate these points. One of the most common uses of regression for prediction is in selection. Suppose, for example, you are a college admissions officer and want to improve your accuracy in admitting students who will do well at your college. Suppose further that you have a number of predictors available: high school grades, rank in class, SAT or ACT scores, participation in academic clubs and athletics, even personality measures. You could develop a prediction equation using your current student body by regressing students' current GPAs on this information, and you could then use this prediction equation as an aid in selecting new students. This equation will look just like the equations we developed earlier in this book:

$$\text{Grades (predicted)} = a + b_1 \text{HSGrades} + b_2 \text{HSRank} + b_3 \text{SAT} + \dots$$

Note that this example illustrates the use of simultaneous regression in the service of prediction. But suppose further that it is difficult to collect all this information, and it would be more cost effective if you could predict almost as well using fewer predictors. In this case, stepwise regression might be a method of reducing the number of variables in the equation while still improving prediction accuracy over the status quo.

Psychologists often use individually administered tests to select participants for a treatment (e.g., special education services or participation in intervention programs). These tests are expensive and time consuming. If shorter versions could be developed with little loss of reliability or validity, we might consider this a worthwhile trade-off. If so, you could use stepwise regression to find out which 4 of the 10 subtests, for example, best predicted the overall score on the test. Future selection could then use the generated equation to predict the overall score from these 4 tests.

Note that for these examples of prediction, theories and models are unimportant. The admissions officer does not care which of these variables *affect* college success; she only cares that the prediction improves the admissions process. The psychologist who is searching for

a valid, but shorter version of a test of intelligence does not care why subtests help in the prediction of the total score. Indeed, he would probably be willing to use an *achievement* test to aid in prediction, even though relevant theory would argue that intelligence affects achievement, rather than the reverse. Likewise, if you could develop a reliable method for predicting the future price of a stock or which horse would win a race, you probably would not care why your equation worked (at least until it stopped working). If our goal is simply prediction, the theoretical relation of the predictors to the criterion does not matter. What is important, however, is that we are not subsequently tempted to interpret our predictive results in an explanatory fashion. The college admissions officer is therefore not justified in telling a potential applicant that if he raises his High School GPA this, in turn, will likely improve his subsequent college GPA.

Cross-Validation

Just as the researchers in these examples don't care why the variables enter the equation in stepwise regression, neither does the program "care" why variables enter the equation. The variance that a predictor accounts for in a criterion may be reliable, valid variation, or it may be due to error, or chance variation. In other words, stepwise regression capitalizes on chance. As a result, the accuracy of prediction, as measured by the variance explained in the criterion by the predictors, or R^2 , is likely to be inflated. Likewise, the regression coefficients used in subsequent prediction may be less accurate than is acceptable.

One way of exploring and improving such prediction is through a method called *cross-validation*. In this method, one sample is used to develop the regression equation, which is then cross-validated on a second sample. The two samples can be separate samples from the same population, or one larger sample split at random. The regression equation from the first sample is used to create a composite, a weighted predicted criterion score, for the second sample (e.g., via a "compute" statement in SPSS). This is similar to the composite variable creation we did in Chapter 3. This predicted criterion is then correlated with the actual criterion in the second sample. If this correlation is considerably smaller than the R from the initial equation, it means that the equation does not generalize and is therefore suspect. Double cross-validation is also possible, in which each sample is used to generate an equation that is then tested in the other sample. If the cross-validation is successful (the r for the second regression is close to the R from the first regression), it is common to combine the two samples to generate even more stable regression weights.

We could split our NELS data set into two samples of 500. For the first exploratory sample, we could use stepwise regression to predict Social Studies Achievement from SES, Grades, Self-Esteem, and Locus of Control. The generated regression equation could then be used to create a composite predicted Social Studies Achievement score in the second, or cross-validation, sample. We could then compare the correlation of this composite with actual Social Studies Achievement in the cross-validation sample with the value of the R for the exploratory sample. If the two were close, we could have confidence that the b 's in Figure 5.13 can be used in another sample to predict the Social Studies Achievement score. It is also common to split the samples in a ratio of two-thirds (exploratory) and one-third (cross-validation).²

Obviously, cross-validation requires a larger sample (or second samples). Ironically, the way to ensure that regression weights are stable, that equations generalize, is through large samples and fewer predictors. This, then, should be another major lesson of stepwise regression: use large samples and relatively few predictors. Unfortunately, this advice often runs counter to the use of stepwise regression in practice. The irony is that researchers often use stepwise regression when they have a small sample and want to reduce the number of predictors in the regression equation.

Adjusted R^2

Just as R^2 is likely to be smaller in a new sample, it is also likely to be smaller in the population than in the sample. There are a number of methods for estimating the population R^2 from the sample R^2 . A common formula is:

$$R_{adjusted}^2 = R^2 - \frac{k(1-R^2)}{N-k-1}$$

If you do the calculations, you will see this is the “adjusted R^2 ” reported in the table in Figure 5.13 (and it appears to be the one used by most computer programs). For much more detail and alternatives, see Darlington and Hayes (2017, chap. 7; also see Raju, Bilgic, Edwards, and Fleer, 1999, for a comparison of methods). The point I want to make is this: the R^2 we would likely get in the population and the one that we would likely get upon cross-validation depend on sample size and the number of predictors used. Other things being equal, your results will be more stable with larger samples and fewer predictors.

I should note that neither of these issues, cross-validation or adjusted R^2 , apply only to stepwise regression or even to regression in the service of prediction. Although less common, we could just as easily and fruitfully cross-validate explanatory regression results. Indeed, such cross-validation can be considered a form of replication; both cross-validation and replication should be conducted more commonly than they are.

Additional Dangers

I hope I have succeeded in convincing you that you should not use stepwise regression in explanatory research. Unfortunately, there are also dangers when using stepwise regression in the service of prediction. I will outline a few of them briefly here; for a more complete treatment, see Thompson (1998).

Degrees of Freedom

At each step in stepwise regression, the program examines *all* variables in the set of predictors, not just the variable added in that step. The degrees of freedom for the regression and residual *should* recognize this use of the data, but computer programs generally print degrees of freedom as if only one variable had been considered, ignoring all the variables that were considered but not entered into the equation. In other words, the degrees of freedom for every step of the regression shown in Figure 5.13 should be 4 and 882, because four variables were either entered or evaluated at every step (and these same *dfs* apply to the final equation, even though only three variables were used in the final equation). The result of such adjustments is that the actual F values are smaller than that listed on most printouts. Likewise, the *adjusted R²* should take into account the total number of predictors used. If sample size is small and the number of predictors large, the actual adjusted R^2 may be much smaller than that shown on the printout.

Not Necessarily the Best Predictors

Stepwise regression is commonly used when the researcher wishes to find the best subset of predictors to accomplish the prediction (indeed, this was the reasoning behind my prediction examples above). Yet, because of the way stepwise regression works, entering one variable at a time, the final set of predictors may not even be the “best” subset. That is, it may not be the subset with the highest R^2 . Thompson illustrates this point nicely (1998).

Lack of Generalizability

More than other regression methods, stepwise regression is especially likely to produce coefficients and equations that generalize poorly to other samples or situations. Cross-validation is especially important for stepwise regressions.

Alternatives to Stepwise Regression

In my experience, most researchers who use stepwise regression are interested in finding which variables are “most important,” in some vague sense, for the outcome. If one digs deeper, it usually turns out that the intended purpose is explanatory in nature. As we have already discussed, simultaneous regression or sequential regression are more appropriate for explanation.

In cases where prediction is the goal, stepwise regression may be acceptable. As noted by Cohen and colleagues (2003, pp. 161–162), the problems of stepwise regression become less severe when the researcher is interested only in prediction, the sample size is large and the number of predictors relatively small, and the results are cross-validated. Even in these cases, there may be better alternatives.

I have already argued that both simultaneous and sequential regression can be used for prediction. If a researcher is interested in developing an equation from a set of predictors, this can be obtained via simultaneous regression. If used for purposes of prediction, one could enter variables in a sequential regression based on the ease of obtaining them, using the coefficients from the final equation to develop the prediction equation.

Even when one simply wants to get the “best” subset of predictors from among a larger group, there are alternatives. *All subsets* regression, for example, will test all possible subsets of a set of predictors to determine which subset offers the best prediction. Say you want the best 10 out of 25 predictors for some outcome; all subsets regression will likely give you this information more accurately than will stepwise regression. This method is one of the options for variable selection in SAS (MaxR). It can be conducted manually in other statistical programs by using a series of regressions and comparing the variance explained by all possible subsets of variables.

One final caveat concerning my disdain for stepwise regression: occasionally you may encounter research that used sequential regression, but referred to it as stepwise regression. Presumably this confusion occurs because the variables were added in steps (here I have tried to use the verbiage “blocks” rather than “steps” when referring to sequential regression to avoid this confusion). Referring to sequential regression as stepwise regression is not common usage, but I see it on occasion. Thus you will need to read the research details to determine for sure which method was used. But don’t assume it was really sequential regression; many researchers use stepwise regression when simultaneous or sequential regression would be a better approach.

Summary: Stepwise Regression

Analysis

In stepwise regression, variables are added one at a time. The order of entry of the variables is controlled by the statistics program; the variable that will lead to the largest increase in ΔR^2 is entered at each step. If an earlier variable becomes statistically not significant with the addition of later variables, it can be dropped from the equation.

Purpose

The primary purpose of stepwise regression is prediction. It is often used to select a subset of available variables that provides efficient prediction of some criterion. Stepwise regression should not be used when you want to understand the effect of a group of variables on some outcome (explanation).

What to Interpret

The statistical significance associated with the change in variance explained (ΔR^2) is the primary focus with stepwise regression. You may also use the produced regression coefficients (b 's) in subsequent prediction equations.

Strengths

Stepwise regression may be useful when you have a large number of possible predictors and don't know which you should keep and which you should discard (of course you should also have a large N). Stepwise regression can help you reduce the number of predictors and still predict the outcome efficiently. It is tempting to think that stepwise regression's ability to choose predictors, thus allowing you to avoid a lot of difficult thinking, is a strength of this method. I believe it is, instead, a weakness.

Weaknesses

It should be obvious that I am no fan of stepwise regression. It should not be used for explanatory research and, if it is used in this manner, the results will likely be useless. Stepwise regression can be used for predictive research, but even then other approaches may be more productive. I believe there are few uses for this method.

Why spend time discussing this method when I and many others discourage its use? In my experience, the use of stepwise regression, although diminishing, is still all too common. And this assessment is not confined to the areas of research with which I am most familiar. As part of the preparation for this chapter, I conducted a series of literature searches for the word *stepwise*, and it was amazing how often the term showed up in connection with stepwise regression. It appears stepwise regression is common in all areas of psychology that use regression, in education, other social sciences, and even medicine. I present, but condemn, stepwise regression because you are likely to encounter it in your reading. I want to discourage you from using the method, however, so I do not present an interpretation here.

THE PURPOSE OF THE RESEARCH

This chapter introduced two new flavors of multiple regression, sequential and stepwise regression, and compared them to simultaneous regression and to each other. We focused on the method of analysis, interpretation, purpose, strengths, and weaknesses of each method. The three general regression approaches, their purposes, strengths, and weaknesses, are summarized in Table 5.5.

Now, the important question: How should you decide which approach to use? The first step is careful thinking about your purpose in conducting the research. What do you want to be able to say about your findings; how do you plan to use them? An examination of your intended purpose will first help you understand whether you are interested in explanation or in prediction. Following this decision, you can focus on more specific questions to help you make an informed choice as to the most appropriate method.

Table 5.5 Three Types of Multiple Regression: Summary Table

Method	Simultaneous	Sequential	Stepwise
Procedure	All variables forced to enter the regression equation at the same time	<i>Researcher</i> enters one variable at a time based on previous knowledge or theory	<i>Computer</i> enters one variable at a time based on increases in variance explained
Purpose	<i>Explanation:</i> relative importance, effects of each variable <i>Prediction:</i> generating prediction equations	<i>Explanation:</i> Is variable important for the outcome? <i>Explanation:</i> test for statistical significance of interaction, curve components <i>Prediction:</i> Does variable aid in prediction?	<i>Prediction:</i> Which variables help predict criterion?
What to Interpret	Overall R^2 , statistical significance of bs , magnitude of bs and β s	Statistical significance of ΔR^2 , magnitude of $\sqrt{\Delta R^2}$	Statistical significance of ΔR^2
Strengths	1. Very useful for explanation, especially when combined with theory 2. Allows conclusions about relative effects of variables 3. Allows conclusions about policy, intervention implications 4. Estimates direct effects in implied model 5. Order of variables in implied model unimportant	1. Useful for explanation if combined with theory 2. Useful for testing for curves and interactions 3. Estimates total effects in implied model (see Part 2 of this book for more information)	1. May tell you which variables can be used for efficient prediction 2. Doesn't require thought or theory
Weaknesses	1. Regression weights can change depending on which variables are entered 2. Implies a theoretical model 3. Estimates only direct effects	1. ΔR^2 changes depending on the order of entry of variables 2. Can over- or under-estimate the importance of variable depending on order of entry 3. Order of entry implies an ordered, theoretical model 4. Estimates only total effects	1. Doesn't require thought or theory 2. Give up control to computer 3. Cannot use for explanation 4. "Theoretical garbage"

I am grateful to Bettina Franzese, who developed the original version of this table.

Explanation

Are you interested in understanding some phenomenon? Do you want to be able to explain how something comes about? Do you wish to make policy recommendations based on your research? Do you want to be able to provide information on what variables should be changed to maximize some worthwhile outcome? Do you want to describe the likely effects of increasing (or decreasing) a variable?

If any of these questions describe your research focus, then you are primarily interested in the goal of explanation, and either simultaneous regression or sequential regression may be an appropriate method. As a general rule, I find simultaneous regression more often useful than sequential regression, but this is, in part, personal preference. There is also considerable overlap between the two methods and the information they provide (more so than most researchers realize). They do have distinct advantages and disadvantages for different problems, however.

I have argued that explanatory research implies a causal model and that you will be on much firmer ground if you think through this causal model prior to conducting research. One way that simultaneous and sequential regression differ is that they focus on different portions of this implied model. Simultaneous regression estimates the *direct* effects from this model, whereas sequential regression focuses on the *total* effects. As a result, the order of the variables in the model and in the regression is very important for sequential regression, but unimportant for simultaneous regression. The practical upshot of this difference is that if you are confident about which variables should appear in your model, but less sure about their ordering, simultaneous regression will be more appropriate. If you are confident in the ordering, either approach can be used, depending on whether your interest is in direct or total effects. In Part 2 of this book we will focus on estimating both direct and total effects in a single model. This topic is intimately related to the issue of mediation, discussed in Chapter 9 (and further in Part 2).

Are you interested in the effects of one variable on another, for example, so that you can make statements about what happens if we change a key variable (as in our earlier homework-achievement examples)? If so, the unstandardized regression coefficients from simultaneous regression are probably your primary interest. The β 's from simultaneous regression can be used to determine the relative importance of the variables in the model.

Are you interested in the unique variance accounted for by a variable? Said differently, perhaps you wonder if a variable is important, after controlling for some already existing variables? Sequential regression is the common method for answering these types of questions, although, as we have seen, simultaneous regression can provide the same information.

Prediction

If your primary interest is prediction, you have more options, including all three methods of multiple regression. I encourage you to spend some time thinking through this basic question, however, because it is often the case that researchers assume they are interested in prediction when, in fact, their real interest is in explanation. Don't be guilty of the bait and switch (i.e., suggesting that you are only interested in prediction but then switching to an explanatory interpretation in your discussion of findings)!

Are you simply interested in generating a prediction equation for a set of variables? In this case, the regression coefficients from simultaneous regression should work well. Are you interested in whether a new variable improves prediction over and above that offered by a given set of predictors? Either sequential or simultaneous regression will work.

Or are you interested in finding a smaller subset of predictors that works well? If so, is it possible to rank order them on some relevant criteria (e.g., ease or cost of obtaining measures of these predictors)? If you can accomplish such rank ordering, sequential regression may be your best bet, with the ranking providing you information on the order of entry. If not, if

you simply have a group of variables from which you want a smaller subset for prediction, stepwise regression may fit the bill (but all subsets regression would probably work better).

COMBINING METHODS

This chapter has necessarily focused on three methods as distinct categories of multiple regression. It is also quite possible to combine the approaches, however. We already broached this topic in the discussion of sequential regression, when we added two variables (simultaneously) in one step of a sequential analysis. Other combinations are possible, as well. We could force one group of variables into a regression equation and then use stepwise regression to choose one additional variable from several possibilities. *Blocks* of variables can be added at every step of a sequential regression.

The important lesson, whatever your approach, is to make sure you understand thoroughly your intent in conducting the research. Once you have this understanding, make sure that your regression method allows you to accomplish this intention.

SUMMARY

In this chapter we expanded our repertoire of MR methods. The method that we have been using for MR so far—simultaneous or forced entry regression—is, in fact, one of several types of MR. Other methods include sequential (or hierarchical) MR and stepwise MR. With simultaneous MR, all variables are entered into the regression equation at the same time. The overall R^2 and the regression coefficients are generally used for interpretation. The b 's and β 's in simultaneous regression represent the direct effects of the variables on the outcome, with the other variables in the equation taken into account. Simultaneous regression is very useful for explanatory research and can provide estimates of the relative effects of the variables on the outcome. Simultaneous regression can also be used for prediction, in which case the standardized regression coefficients estimate the relative importance of the predictors. The primary weakness of simultaneous regression is that the regression coefficients may change depending on the variables entered in the regression.

In sequential or hierarchical multiple regression, variables are entered in steps or blocks predetermined by the researcher; time precedence is a common basis for such order of entry. The change in R^2 from one step to the next is generally used to test the statistical significance of each variable, and $\sqrt{\Delta R^2}$ may be interpreted as the measure of the relative importance of each variable's total effect (given the correct order of entry of the variables). The regression coefficients from each step of the regression may be interpreted as the total effects of each variable on the outcome, if the variables have been entered in accordance with a theoretical model. A variation, sequential unique MR, is used to determine whether one or several variables are important (explain additional variance) after taking an original set of variables into account. This form of sequential regression is commonly used to test the statistical significance of interactions and curves in the regression line. Sequential regression may be useful for explanation, if the variables are entered in accordance with theory; it can also be used to determine if a variable is useful in prediction. The primary weakness of sequential regression is that the ΔR^2 changes depending on the order of entry of the variables, and thus it can over- or underestimate the importance of variables, depending on the order of entry of variables in the regression.

In stepwise multiple regression and its variations, variables are also entered one at a time, but the computer program chooses the order of entry based on the degree to which each variable increases ΔR^2 . Although this solution may seem to avoid problems in simultaneous or sequential regression in determining the "importance" of variables, it does not. The reason

that stepwise regression does not help in determining the importance of variables is because *using ΔR^2 as a measure of the importance of variables is predicated on the assumption that the variables have been entered in the correct order.* It would be circular reasoning (and a statistical version of the logical fallacy of begging the question) to also use ΔR^2 to determine the order of entry. For this reason, stepwise regression should not be used for explanation. Stepwise regression is only appropriate when the purpose is prediction and, even then, simultaneous and sequential regression may be more appropriate. ΔR^2 and its statistical significance are the primary focus of interpretation in stepwise regression. If used to develop a prediction equation, the b 's from the final regression equation will also be used.

It is also possible, and indeed common, to combine these methods. In the next few chapters, for example, we will combine simultaneous and sequential regression to test for interactions and curves. The chapter ended with a plea that you thoroughly understand your purpose for using multiple regression. This purpose, in turn, will help you decide which method or methods of MR you should use.

EXERCISES

1. Choose an outcome variable from NELS and four or five variables you think may help explain this outcome. Conduct a simultaneous regression, a sequential regression, and a stepwise regression using your variables. Provide an appropriate interpretation of each regression. Make sure you understand and can explain any differences in the three solutions.
2. Pair with a classmate; analyze his or her variables from Exercise 1 in your own sequential regression. Did you both choose the same ordering? Make sure you can explain to your partner the reasons for choosing your ordering. Draw a “model” that explains the ordering you chose for your problem.
3. In an interesting study of a controversial topic, Sethi and Seligman studied the effect of religious fundamentalism on optimism (1993). Think about this problem: are religious fundamentalists likely to be less optimistic or more so than those with a more “liberal” religious orientation? Perhaps fundamentalists have a strict and stern religious orientation that will lead to greater pessimism (and thus less optimism). Or perhaps those with more fundamentalist views decide to let God worry about the problems of the world, thus leading to a more optimistic view. What do you think?

The files titled “Sethi & Seligman simulated” (there are SPSS, Excel, and text [.dat] files) are designed to simulate the Sethi & Seligman data from a MR perspective.³ The primary variables of interest are Fundamentalism (coded so that a high score represents high religious fundamentalism, a low score religious liberalism) and Optimism (high score = optimistic, low score = pessimistic). Also included are several measures of religiosity: the extent of influence of religion in one’s daily life (Influence), religious involvement and attendance (Involve), and religious hope (Hope). It may be important to control for these variables in examining the effect of Fundamentalism on Optimism.

Rgress Optimism on these variables, using both simultaneous and sequential regression. For the sequential regression, design your regression to determine whether Fundamentalism affects Optimism above and beyond the effects of Involvement, Hope, and Influence. Could you get the same information from the simultaneous regression? Interpret your results.

4. Use a library research database (e.g., PsycINFO, Sociological Abstracts, ERIC, Google Scholar) to find an article in your area of interest that used stepwise regression in the analysis. Read the article: are the authors more interested in prediction or explanation?

Pay special attention to the Discussion: do the authors make inferences from their regression that if a predictor variable were increased or reduced then people would change on the outcome? Was stepwise regression appropriate? Would some other method have been more appropriate?

5. This exercise is designed to show differences in findings depending on whether you use simultaneous or sequential regression, and depending on the order of entry in sequential regression. Angela Duckworth and Martin Seligman (2005) were interested in the effects of self-discipline on students' academic performance. They measured the self-discipline (degree of self-regulation and lack of impulsiveness across multiple domains) of 154 eighth-graders in the Fall, and used that to predict (or was it explain?) final GPA in the Spring. Also controlled were students' IQs and their previous GPA, also measured in the Fall. The data are in the file "Duckworth Seligman sim data.sav"; the data are simulated but are designed to produce results consistent with the findings of the original study.
 - a. Conduct a simultaneous regression of GPA on IQ, previous GPA (Pre_GPA), and self-discipline (Self). Which variables are important in explaining GPA? In particular, how important are IQ and self-discipline?
 - b. Conduct a sequential regression of GPA on these same variables. For this regression, enter IQ, Self, and Pre_GPA, in that order. Draw the causal model implied by this regression. Using $\sqrt{\Delta R^2}$ and the β for each variable (as it is entered), note the relative importance of the variables. Again, focus in particular on IQ and self-discipline.
 - c. Conduct another sequential regression on these same variables. This time the order of entry should be IQ, Pre_GPA, and Self. Draw the causal model implied by this regression. Compare these results to the results from step b, with particular attention to the self-discipline variable.
 - d. Explain the reason for the differences in the "importance" of the variables from the different regressions.

Notes

- 1 Darlington and Hayes (2017) recommended that semipartial correlations (also known as *part* correlations) be used to compare the effects of different variables in simultaneous regression, as well, and instead of β 's. You can request semipartial correlations as part of the output for some computer programs, but others do not routinely provide them. Semipartial correlations can be calculated

from the values of t given for each regression coefficient, however: $sr_{y(1-234)} = t \sqrt{\frac{1-R^2}{N-k-1}}$ (Dar-

lington & Hayes, 2017, p. 226). Thompson (2006) argued for the interpretation of structure coefficients in addition to the β 's. Structure coefficients are equal to the correlation of each predictor variable with the outcome, divided by the multiple correlation coefficient: $r_{structure} = r_{yx}/R$. A conversion of the original bivariate correlation, the structure coefficient tells us something about whether each predictor is related to the outcome when other variables are not controlled. In contrast, the β 's tell us something about whether each predictor is related to the outcome when the other variables are controlled.

- 2 I have greatly simplified the issue of cross-validation and recommend additional reading if you use the methodology. There are actually a number of different formulas for estimating the true R^2 and the likely cross-validation R^2 ; Raju and colleagues (1999) compared these empirically. Even more interesting, this and other research suggests that equal weighting of predictors often produces better cross-validations than do those based on MR estimates!
- 3 In the original study, members of nine religious groups were categorized into a Fundamentalist, Moderate, or Liberal categorical variable and the results analyzed via ANOVA. For this MR simulation, I instead simulated a continuous Fundamentalism variable. The results, however, are consistent with those in the original research. I used simulation provided by David Howell as the starting point for creating my own simulation data (www.uvm.edu/~dhowell/StatPages/Special%20Topics%20Folder/Fundamentalism/Fundamentalism.html). Howell's Web pages have numerous excellent examples.

6

Analysis of Categorical Variables

Dummy Variables	109
<i>Simple Categorical Variables</i>	109
<i>More Complex Categorical Variables</i>	110
<i>False Memory and Sexual Abuse</i>	110
Other Methods of Coding Categorical Variables	116
<i>Effect Coding</i>	117
<i>Criterion Scaling</i>	118
Unequal Group Sizes	120
<i>Family Structure and Substance Use</i>	120
Additional Methods and Issues	125
Summary	126
Exercises	127
Notes	128

Our analyses to this point have focused on explaining one continuous dependent variable by regressing it on one or more continuous independent variables. In Chapter 1, however, I argued that a major advantage of multiple regression is that it can be used to analyze both continuous and categorical independent variables. We begin our analysis of categorical independent variables in this chapter.

Categorical variables are common in research. Sex, ethnic origin, religious affiliation, region of the country, and many other variables are often of interest to researchers as potential influences or control variables for a multitude of possible outcomes. We may be interested in the effects of sex or ethnic origin on children's self-esteem, or the effects of religious affiliation or place of residence on adults' voting behavior.¹ Yet these variables are substantively different from the variables we have considered in our MR analyses to this point. Those variables—Homework, SES, Locus of Control, and so on—are continuous variables, ranging from low (e.g., no homework) to high (e.g., 15 hours of homework). Variables such as sex or ethnic origin have no high or low values, however. Certainly we can assign “boys” a value of 0 and “girls” a value of 1, but this assignment makes no more sense than assigning boys a value of 1 and girls a value of 0. Likewise, we can assign values of 1, 2, 3, 4, 5 to Northeast, Southeast, Midwest, Southwest, and West, respectively, but any other ordering will make just as much sense. These variables each use a nominal, or naming, scale; names make more sense for the values of the scales than do numbers. How, then, can we analyze such variables in multiple regression analysis?

DUMMY VARIABLES

Simple Categorical Variables

As it turns out, we can analyze such categorical variables by creating a series of scales in which we assign values of 1 for membership in a category and values of 0 for nonmembership. Thus, our initial coding of the sex variable (boys = 0, girls = 1) can be thought of as a “girl” variable, with membership coded as 1 and nonmembership (i.e., boys) coded as 0. Such coding is called *dummy coding*, creating a dummy variable.²

Figure 6.1 shows the results of a *t* test comparing the 8th-grade reading achievement of girls and boys using the NELS data. For this analysis, I recoded the existing sex variable (Sex, boys = 1, girls = 2) into NewSex, with boys = 0 and girls = 1. The average score for boys on the Reading test was 49.58 versus 52.62 for girls. This difference of 3 points is relatively small; measures of effect size, for example, are $d = .304$, and $\eta^2 = .023$. Nevertheless, the difference is statistically significant ($t = 4.78$, $df = 965$, $p < .001$); girls score statistically significantly higher on the 8th-grade Reading tests than do boys.

Now turn to Figure 6.2. For this analysis, I regressed the 8th-grade Reading test score on the NewSex dummy variable. As you can see by comparing Figure 6.2 with Figure 6.1, the results of the regression are identical to those of the *t* test. The *t* associated with the NewSex regression coefficient was 4.78, which, with 965 degrees of freedom, is statistically significant ($p < .001$).

These figures include additional information, as well. Note that the R^2 is equal to the η^2 I reported above (.023); in fact, η^2 is a measure of the variance accounted for in a dependent variable by one or more independent variables (i.e., R^2). In other words, the η^2 commonly reported as a measure of effect size in experimental research is equivalent to the R^2 from MR. Turn next to the table of coefficients in Figure 6.2. Recall from Chapter 1 that the intercept (constant) is equal to the predicted score on the dependent variable for those participants with a score of zero on the independent variable(s). When dummy coding is used, the intercept is the mean on the dependent variable for the group coded 0 on the dummy variable. When dummy variables are used to analyze the results of experimental research, the group coded 0 is often the control group. In the present example, boys were coded 0; thus the mean Reading score for boys is 49.58. With dummy coding, the b , in turn, represents the deviation from the intercept for the other group; in the present example, then,

Group Statistics

	NEWSEX Sex	N	Mean	Std. Deviation	Std. Error Mean
BYTXRSTD READING STANDARDIZED SCORE	1.00 Female .00 Male	464 503	52.61781 49.58206	9.83286 9.90667	.45648 .44172

Independent Samples Test

	t-test for Equality of Means						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
BYTXRSTD READING STANDARDIZED SCORE	4.778	965	.000	3.03576	.63540	1.78884	4.28268

Figure 6.1 *t* test analyzing Reading test score differences for boys and girls.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.152 ^a	.023	.022	9.87132

a. Predictors: (Constant), NEWSEX Sex

ANOVA^b

Model	Sum of Squares		df	Mean Square	F	Sig.
1	Regression	2224.300	1	2224.300	22.827	.000 ^a
	Residual	94032.55	965	97.443		
	Total	96256.85	966			

a. Predictors: (Constant), NEWSEX Sex

b. Dependent Variable: BYTXRSTD READING STANDARDIZED SCORE

Coefficients^a

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	49.582	.440		112.650	.000	48.718	50.446
NEWSEX Sex	3.036	.635	.152	4.778	.000	1.789	4.283

a. Dependent Variable: BYTXRSTD READING STANDARDIZED SCORE

Figure 6.2 Regression of Reading test scores on Sex. The results are the same as for the *t* test.

girls (the group coded 1) scored 3.04 points higher on the Reading test than did boys (the group coded 0). Again, the results match those of the *t* test.

More Complex Categorical Variables

The same technique works with more complex categorical variables, as well. Consider a question about religious affiliation. We could ask “What is your religion?” and then list the possibilities, as shown in the top of Table 6.1. Alternatively we could ask a series of four questions, with yes or no answers, to get the same information, as shown in the bottom of Table 6.1. The two methods are equivalent. If you considered yourself to have some other religious affiliation than those listed, you would choose the final option for the first method or simply answer no to each question for the second method. Essentially, we do something similar to analyze categorical variables in multiple regression by changing them into a series of yes or no, or dummy, variables. An example research study will illustrate the coding and analysis of dummy variables; we will then use the same study to illustrate other possible coding methods.

False Memory and Sexual Abuse

Considerable controversy surrounds adult self-reports of previous childhood sexual abuse: do such reports always represent valid, but repressed, memories, or are they false memories (cf. Alexander et al., 2005; Belli, 2016)? Bremner, Shobe, and Kihlstrom (2000) investigated memory skills in women with self-reported sexual abuse and posttraumatic stress disorder (PTSD). Briefly, women who had and had not been sexually abused as children were read lists of words and were later given a list of words, including words they had heard along with words *implied* by, but not included on, the original lists (“critical lures,” or

Table 6.1 Two Different Methods of Asking (and Coding) Religious Affiliation

What is your religion?

1. Protestant
2. Catholic
3. Jewish
4. Islam
5. Other (or none)

<i>Are you:</i>	<i>Yes</i>	<i>No</i>
Protestant?	1	0
Catholic?	1	0
Jewish?	1	0
Muslim?	1	0

Report

FALSEPOS percent of false positives

GROUP group membership	Mean	N	Std. Deviation
1.00 Abused, PTSD women	94.6000	20	10.3791
2.00 Abused, Non-PTSD women	68.0500	20	39.3800
3.00 Non-abused, non-PTSD women	63.5500	20	27.9143
Total	75.4000	60	31.2395

Figure 6.3 Descriptive statistics for the false memory data.

false positives). Figure 6.3 shows the (simulated) percentage of these false positives remembered by Abused women with PTSD, Abused women without PTSD, and Nonabused, non-PTSD women.³ The data are also on the Web site (“false memory data, 3 groups.sav” or “false.txt”). As the figure shows, Abused, PTSD women falsely recognized more words not on the list than did non-PTSD and Nonabused women; in fact, they “recalled” almost 95% of the false critical lures as being on the lists. The differences are striking, but are they statistically significant?

ANOVA and Follow-Up

The most common way of analyzing such data is through analysis of variance. Such an analysis is shown in Figure 6.4. As the figure shows, there were indeed statistically significant differences in the percentages of false recalls across the three groups ($F = 6.930 [2, 57]$, $p = .002$). Although not shown in the figure, the difference across groups represented a medium to large effect size ($\eta^2 = .196$), one that would presumably be apparent to a careful observer (cf. Cohen, 1988).

Also shown in Figure 6.4 are the results of Dunnett’s test, which is a post hoc test used to compare several groups to one group, usually several experimental groups to a single control group. Here our interest was to compare Abused women (with and without PTSD) to women who had not been abused. As shown in the figure, Abused, PTSD women had statistically significantly more false positives than did women who were not abused, but the difference between Abused (Non-PTSD) and Nonabused women was not statistically significant.

Regression Analysis with Dummy Variables

Our real interest, of course, is not in the ANVOA tables but in how to conduct such an analysis via multiple regression; the ANOVA is included for comparison purposes. The three groups may be considered a single, categorical variable with three categories: 20 participants form the Abused, PTSD group, coded 1; 20 participants form the Abused, Non-PTSD group, coded 2; and so on. We need to convert the categorical Group variable into dummy variables.

To include all the information contained in a single categorical variable, we need to create as many dummy variables as there are categories, **minus 1**. The example includes three categories, or groups, so we need to create two ($g - 1$) dummy variables. Each dummy variable should represent membership in one of the groups. Table 6.2 shows how I translated the original single categorical variable into two dummy variables. The first dummy variable, AbusePTS (meaning Abused, PTSD) has values of 1 for members of the Abused, PTSD group, and thus contrasts members of this group with all others. The second dummy variable (Abuse_NoPTS) is coded so that members of the Abused, Non-PTSD group were coded 1, while all other participants were coded 0. The actual computer manipulations to create these dummy variables can be accomplished in SPSS via RECODE or IF commands, and similar commands in other programs.

ANOVA

FALSEPOS percent of false positives

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	11261.70	2	5630.850	6.930	.002
Within Groups	46316.70	57	812.574		
Total	57578.40	59			

Multiple Comparisons

Dependent Variable: FALSEPOS percent of false positives

Dunnett t (2-sided)^a

(I) GROUP group membership	(J) GROUP group membership	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00 Abused, PTSD women	3.00 Non-abused, non-PTSD women	31.0500*	9.0143	.002	10.6037	51.4963
2.00 Abused, Non-PTSD women	3.00 Non-abused, non-PTSD women	4.5000	9.0143	.836	-15.9463	24.9463

*. The mean difference is significant at the .05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Figure 6.4 Analysis of variance of the false memory data, with Dunnett's test as a follow-up.

Table 6.2 Converting a Group Variable with Three Categories into Two Dummy Variables

Group	AbusePTS	No_PTS
1 Abused, PTSD	1	0
2 Abused, Non-PTSD	0	1
3 Nonabused, Non-PTSD	0	0

You may wonder why there is no third dummy variable that compares the Nonabused, Non-PTSD group with the other two groups. But such a third dummy variable is not needed; it would be redundant. Consider that in multiple regression we examine the effect of each variable, with the other variables in the equation *held constant*. If we regress the proportion of false positives on only the first dummy variable, our results will highlight the comparison of Abused, PTSD participants against the other two groups. We will use *multiple* regression, however, and control for the second dummy variable at the same time, which means the first dummy variable will show the effect of Abuse and PTSD, while controlling for Abuse, Non-PTSD. The result is that in the multiple regression the first dummy variable will contrast the Abused, PTSD group with the Nonabused (and Non-PTSD) group, whereas the second dummy variable will compare the Abused, Non-PTSD with the Nonabused group. We will return to this question (the number of dummy variables) later in the chapter.

Figure 6.5 shows a portion of the data. It is always a good idea to check the raw data after recoding or creating new variables to make sure the results are as intended. Figure 6.5 shows that the two dummy variables were created correctly.

Group	AbusePTS	Abuse_NoPTS
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
2	0	1
2	0	1
2	0	1
2	0	1
2	0	1
3	0	0
3	0	0
3	0	0
3	0	0
3	0	0

Figure 6.5 Portions of the false memory data showing the Group variable converted into two dummy variables, AbusePTS and No_PTSD.

For the multiple regression, I regressed the percentage of false positives on these two dummy variables; the results are shown in Figure 6.6, where these two variables account for 19.6% of the variance in the number of false positives. This R^2 (.196) matches the η^2 from the analysis of variance. Likewise, the F associated with the regression (6.930 [2, 57], $p = .002$) matches that from the ANOVA.

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.442 ^a	.196	.167	28.50568	

a. Predictors: (Constant), Abused, non-PTSD vs other, Abused, PTSD vs other

ANOVA ^a					
Model		Sum of Squares	df	Mean Square	F
1	Regression	11261.700	2	5630.850	6.930
	Residual	46316.700	57	812.574	
	Total	57578.400	59		

a. Dependent Variable: percent of false positives

b. Predictors: (Constant), Abused, non-PTSD vs other, Abused, PTSD vs other

Coefficients ^a						
Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B
	B	Std. Error	Beta			
1	(Constant)	63.550	6.374		.000	50.786 76.314
	Abused, PTSD vs other	31.050	9.014	.472	.001	12.999 49.101
	Abused, non-PTSD vs other	4.500	9.014	.068	.499	-13.551 22.551

a. Dependent Variable: percent of false positives

Figure 6.6 Multiple regression analysis of the false memory data using the two dummy variables.

Post Hoc Probing

The regression coefficients, also shown in Figure 6.6, may be used to perform post hoc comparisons. As in the simpler example, the intercept (constant) provides the mean score on the dependent variable (percentage of false positives) for the group that was assigned zeros for *both* dummy variables. Again, this is often the “control” group and, in this case, is the mean score for those participants who were neither abused nor suffered from PTSD ($M = 63.55$). The regression coefficients, in turn, represent the *deviations* from this mean for *each* of the other two groups. Women who were abused and suffer from PTSD had an average of 94.60 false positives ($63.55 + 31.05$), and abused, non-PTSD women had an average of 68.05 false positives ($63.55 + 4.50$). Compare these calculations of the mean scores for each group with those shown in Figure 6.3.

Dunnett’s Test

The *t*’s associated with the dummy variables can be used in several ways. First, we can use them for Dunnett’s test, as we did with the ANOVA. To do so, you need to ignore the probabilities associated with the *t*’s on the printout; instead look up those values of *t* in Dunnett’s table in a statistics book that contains a variety of such tables (e.g., Howell, 2013; Kirk, 2013), or search for it online, e.g., “table Dunnett’s test.” The critical values for three treatment groups and 60 degrees of freedom (the closest value in the table to the actual value of 57 *df*) are 2.27 ($\alpha = .05$) and 2.90 ($\alpha = .01$) (two-tailed test, Appendix E, Kirk, 2013). Thus, the regression results are again identical to those from the ANOVA: the data suggest that Abused, PTSD participants have statistically significantly more false memories of words than do nonabused women, whereas the difference between Abused and Nonabused women without PTSD is not statistically significant.

Why not simply use the probabilities associated with the *t*’s in Figure 6.6? And why do these probabilities differ from those shown in Figure 6.4? Simply put, Dunnett’s test takes

into account the number of comparisons made in an effort to control the total family-wise error rate. Recall that if you conduct, for example, 20 t tests, each with an error rate of .05, you would likely find one of these comparisons to be statistically significant by chance alone. Many post hoc tests control for this increase in family-wise error rate resulting from multiple comparisons, and Dunnett's test is one such post hoc comparison. The probabilities associated with the t 's in the regression do not take the family-wise error rate into account, but when we look up the t 's in Dunnett's table, we do take these error rates into account.⁴

Other Post Hoc Tests

We could also simply focus on the t 's and associated probabilities in the regression output, uncorrected for the family-wise error. Given the statistical significance of the overall regression, this procedure is equivalent to the Fisher least significant difference (LSD) post hoc procedure. Alternatively, we can use the Dunn–Bonferroni procedure to control the family-wise error rate. That is, we set the overall alpha to .05 and decide to make two comparisons. We would then look at the probabilities associated with each dummy variable and count any with $p < .025$ (.05/2) as statistically significant. With the current example, all three approaches (Dunnett, multiple t tests, and Dunn–Bonferroni) give the same answer, although with different levels of probability. This will not always be the case. The Dunn–Bonferroni procedure is more conservative (meaning that it is least likely to be statistically significant) than is the use of multiple t tests, among the most liberal procedures. Dunnett's test is more conservative than the LSD procedure but is fundamentally different in that it only makes a subset of all possible comparisons.

What if we were interested in the third possible comparison, whether the difference between Abused, PTSD and Abused, Non-PTSD participants was statistically significant? Using the regression results, you can calculate the mean difference between the two results ($94.60 - 68.05 = 26.55$). As long as the n 's in each group are the same, the standard error of this difference is the same for all three possible comparisons; as shown in Figure 6.6, the standard error is 9.014.⁵ Thus the t associated with a comparison of the AbusePTS and the No_PTSD groups is $26.55/9.014 = 2.95$ ($p = .005$). We then either use this value in an LSD-type post hoc comparison or compare it to $\alpha = .0167$ (.05/3) in a Dunn–Bonferroni comparison. In either case, we conclude that abused women with PTSD also have statistically significantly more false memories of words than do abused women without PTSD. Try conducting an ANOVA on these data, followed by both the LSD and Dunn–Bonferroni post hoc analyses to check the accuracy of these statements. Of course, to make this final comparison in MR, you could also simply redo the dummy coding making, for example, group 1 the comparison group to find this final comparison.

Demonstration of the Need for Only $g - 1$ Dummy Variables

I argued earlier that we only need $g - 1$ dummy variables because this number of dummy variables captures all the information contained in the original categorical variable. In the present example, we only need two dummy variables to capture all the information from the three categories used in the research. You may be skeptical that $g - 1$ dummy variables indeed include all the information of the original categorical variable, but we can demonstrate that equivalence easily. To do so, I regressed the original Group variable used in the ANOVA analysis against the two dummy variables that I claim capture all the information contained in that Group variable. If the two dummy variables do indeed include all the information from the original categorical variable, then the dummy variables should account for 100% of the variance in the Group variable; R^2 will equal 1.0. If, however, a third dummy variable is needed to contrast the three groups, then the R^2 should equal something less than 1.0.

Figure 6.7 shows the results of such a multiple regression: the two created dummy variables do indeed explain all the variation in the original categorical variable. Thus, we only

need $g - 1$ dummy variables to correspond to any categorical variable (indeed, if we use g dummy variables, our MR would encounter problems).

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Abuse_NoPTS Abused, non-PTSD vs other, AbusePTS Abused, PTSD vs other ^b	.	Enter

a. Dependent Variable: Group group membership

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000 ^a	1.000	1.000	.00000

a. Predictors: (Constant), Abuse_NoPTS Abused, non-PTSD vs other, AbusePTS Abused, PTSD vs other

Figure 6.7 Multiple regression demonstrating the need for only $g - 1$ dummy variables. These dummy variables explain 100% of the variance of the original Group variable.

Was Multiple Regression Necessary?

Now, was there any reason to use multiple regression to analyze the results of this research? No; for this problem, it would be easier to analyze the data using ANOVA. The simple example is included for several reasons, however. First, it is important to understand the continuity between multiple regression and ANOVA. Second, you may well encounter or develop more complex experimental designs in which it makes more sense to conduct the analysis via MR than ANOVA. Third, you need to understand how to analyze categorical variables in MR as a foundation for conducting MR analyses that include *both* categorical and continuous variables.

This final reason is, I think, paramount. Most of us will likely rarely use MR to analyze either simple or complex experimental data. We will, however, use MR to analyze a mix of continuous and categorical variables. A thorough understanding of the analysis of categorical variables provides a foundation for this type of analysis.

OTHER METHODS OF CODING CATEGORICAL VARIABLES

There are other methods of coding categorical variables besides dummy coding. We will review a few of these briefly. What is important to keep in mind is that these different methods all lead to the same overall outcome (i.e., the same R^2 and level of statistical significance). In other words, the model summary and ANOVA table will be the same across the methods. Different methods, however, can produce differences in the coefficients, in part because the comparisons being made are different. I will use our current example to illustrate several other methods of coding categorical variables.

Effect Coding

Effect coding is another method of coding categorical variables so that they can be analyzed in multiple regression. In dummy coding one group ends up being assigned zeros on all dummy variables (the control or contrast group). Effect coding is similar in that there is this same contrast group, but with effect coding this group is assigned -1 on both effect variables, rather than 0. The contrast group is usually the last group, or may be the group for which you are least interested in making comparisons (Cohen et al., 2003).

Table 6.3 shows effect coding for the three groups for the Abuse/PTSD example. For the first effect variable, the Abused/PTSD group is coded 1 and all other groups are scored 0, except for the final group, which is scored -1. (In this example, these “all other groups” only includes 1 group, Abused/Non-PTSD, but if we had, say, 6 groups and 5 effect coded variables, 4 groups would be coded 0 on this first effect variable.) For the second effect variable, the Abused, Non-PTSD group was coded 1 and all other groups were coded 0, except for the final group, which was coded -1.

Table 6.3 Converting a Group Variable with Three Categories into Two Effect Coded Variables

Group	<i>AbusePTS_E1</i>	<i>Abuse_NoPTS_E2</i>
1 Abused, PTSD	1	0
2 Abused, Non-PTSD	0	1
3 Nonabused, Non-PTSD	-1	-1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.442 ^a	.196	.167	28.50568

a. Predictors: (Constant), Abuse_NoPTS_E2 Abused, Non-PTSD effect variable 2, AbusePTS_E1 Abused, with PTSD effect variable 1

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11261.700	2	5630.850	6.930	.002 ^b
	Residual	46316.700	57	812.574		
	Total	57578.400	59			

a. Dependent Variable: Falsepos percent of false positives

b. Predictors: (Constant), Abuse_NoPTS_E2 Abused, Non-PTSD effect variable 2, AbusePTS_E1 Abused, with PTSD effect variable 1

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	75.400	3.680		20.489	.000	68.031	82.769
	AbusePTS_E1 Abused, with PTSD effect variable 1	19.200	5.204	.506	3.689	.001	8.778	29.622
	Abuse_NoPTS_E2 Abused, Non-PTSD effect variable 2	-7.350	5.204	-.194	-1.412	.163	-17.772	3.072

a. Dependent Variable: Falsepos percent of false positives

Figure 6.8 Multiple regression analysis of the false memory data using two effect coded variables.

Figure 6.8 shows the results of the regression of the percentage of false positives on the two effect variables. The same percentage of variance was accounted for as in the previous

regression (19.6%), with the same resulting F and probability. Differences do show up, however, in the table of coefficients.

Why the differences? The intercept and b 's highlight different comparisons for effect than for dummy coding. Recall that the intercept is the predicted score on the dependent variable for those with a score of zero on all independent variables. But with effect coding no group is coded zero on all the effect variables. With effect coding, the intercept (constant) shown in Figure 6.8 is the *grand* mean of all three groups on the dependent variable (percentage of false positives). The intercept, representing the grand, or overall, mean is 75.40; note that this value is the same as the overall mean listed in Figure 6.3. The b 's, in turn, are the deviation for each group from the grand mean. The b associated with the first effect variable, representing the first group, was 19.20, and the mean for this group on the dependent variable was 94.60 ($75.40 + 19.20$). The mean for the second group was 68.05 ($75.40 + [-7.35]$). If we want to find the mean score for the third group, we simply sum the two b 's ($19.20 + [-7.35] = 11.85$) and change the sign (-11.85). This is the deviation of the third group from the grand mean, so the mean of group 3 is 63.55 ($75.40 + [-11.85]$). The reason why is because the intercept is the grand mean, and each b represents each group's deviation from the mean. The three deviations from the mean must sum to zero, and thus the third deviation has the same absolute value as the sum of the other two deviations but with a reversed sign. This way the three deviations do sum to zero ($19.20 - 7.35 - 11.85 = 0$).

The t 's in this coding method represent the statistical significance of the difference between each group and the overall mean. That is, does each group differ at a statistically significant level from all other groups? This is an uncommon post hoc question, but it may be of interest in some applications. It is possible, of course, to calculate other post hoc comparisons using the group means (cf. Pedhazur, 1997, Chap. 11).

Recall that in earlier chapters I discussed ANOVA as a part of the general linear model, with formulas like $Y = \mu + \beta + e$. This formula may be stated as follows: any person's score on the dependent variable Y is the sum of the overall mean μ , plus (or minus) variation due to the effect of the treatment (β), plus (or minus) random variation due to the effect of error (e). We would interpret the regression equation ($Y = a + bX + e$) using effect coding in the exact same manner: a person's score on the dependent variable is the sum of the overall mean (a), plus (or minus) variation due to their group, plus (or minus) random variation due to error. One advantage of effect coding is that it illustrates nicely the general linear model in analysis of variance.

Criterion Scaling

Suppose you have a large number of categories for a categorical variable and are only interested in the *overall* effects of the categorical variable, not any subsequent post hoc comparisons. As an example, the controversial book *More Guns: Less Crime* (Lott, 2010) made extensive use of multiple regression. One categorical independent variable of interest was the 50 states, which could be represented by 49 dummy variables. There is an easier way to take the various states into account, however, through a method called *criterion scaling*.

With criterion scaling, a *single* new variable is created to replace the $g - 1$ dummy variables. For this single variable, each member of each group is coded with that group's mean score on the dependent variable. Thus, for the present example, using the group means displayed in Figure 6.3, all members of the Abused, PTSD group are assigned a score of 94.60 on this new variable, whereas members of the Abused, Non-PTSD group are assigned values of 68.05, and so on. Figure 6.9 shows a portion of the data following the creation of this criterion scaled variable (Crit_Var).

The dependent variable, percentage of false positive memories, was regressed on Crit_Var, and the results are shown in Figure 6.10. Note that the explained variance is identical to the previous printouts. Also note, however, that the F and its associated probability are different

FALSEPOS	GROUP	CRIT_VAR
55.00	1.00	94.60
93.00	1.00	94.60
89.00	1.00	94.60
98.00	1.00	94.60
96.00	1.00	94.60
100.00	2.00	68.05
100.00	2.00	68.05
16.00	2.00	68.05
7.00	2.00	68.05
73.00	2.00	68.05
61.00	2.00	68.05
100.00	3.00	63.55
61.00	3.00	63.55
27.00	3.00	63.55
10.00	3.00	63.55
96.00	3.00	63.55

Figure 6.9 Portions of the false memory data showing the group variable converted into a single criterion coded variable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.442 ^a	.196	.182	28.2589

a. Predictors: (Constant), CRIT_VAR

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11261.70	1	11261.700	14.102	.000 ^a
	Residual	46316.70	58	798.564		
	Total	57578.40	59			

a. Predictors: (Constant), CRIT_VAR

b. Dependent Variable: FALSEPOS percent of false positives

Figure 6.10 Multiple regression results using criterion coding. The ANOVA table needs to be corrected for the proper degrees of freedom.

(and incorrect). When criterion scaling is used, the degrees of freedom associated with the criterion scaled variable will be incorrect. Even though we have collapsed $g - 1$ dummy variables into a single criterion scaled variable, this variable still represents the g groups in the original categorical variable, and the df associated with it should be $g - 1$. In the present example, Crit_Var still represents three groups, and the df for the regression should still be 2 (and not 1). And because the df for the regression is incorrect, the df for the residual is incorrect, and the F is incorrect as well. The bottom line is that when you use criterion scaling you need to recalculate F using the printed sums of squares but the corrected degrees of freedom.

UNEQUAL GROUP SIZES

For the PTSD example used in this chapter, there were equal numbers of women in each of the three PTSD/Abuse groups. In the real world of research, however, there are often different numbers of participants in different levels of an independent variable; unequal n 's are especially common in nonexperimental research. Naturally occurring groups (ethnic group membership, religious affiliation) rarely conform to our research desire for equal numbers from each group (the variable sex is sometimes an exception, since this variable is close to evenly split at many ages). If we conduct our research by simply sampling from the population, our samples will reflect this difference in sample sizes across groups. Even in experimental research, where participants are assigned at random to different groups, we have participants who drop out of the research, and this participant mortality often varies by group. The result is unequal sample sizes by group.⁶

As you will see, having equal numbers in groups makes it easier to interpret the results of the regression. An example from NELS will illustrate the differences.

Family Structure and Substance Use

Does family structure affect adolescents' use of dangerous and illegal substances? Are adolescents from intact families less or more likely to use alcohol, tobacco, and drugs? To examine these questions, I analyzed the effect of Family Structure (coded 1 for students who lived with both parents, 2 for students who lived with one parent and one guardian or step-parent, and 3 for students who lived with a single parent) on Substance use, a composite of students' reports of their use of cigarettes, alcohol, and marijuana.⁷ The descriptive statistics for the two variables are shown in Figure 6.11. As you would expect, there were unequal numbers of students from households with two parents, a single parent, and so on. The Substance Use variable was a mean of z scores, with negative scores representing little use of substances and positive scores representing more common use of substances.

FAMSTRUC Family Structure

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Two-parent family	677	67.7	69.7	69.7
	2.00 One parent, one guardian	118	11.8	12.1	81.8
	3.00 Single-parent family	177	17.7	18.2	100.0
	Total	972	97.2	100.0	
Missing	System	28	2.8		
	Total	1000	100.0		

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
SUBSTANC Use of alcohol, drugs, tobacco	855	-.81	3.35	-.0008	.77200
Valid N (listwise)	855				

Figure 6.11 Descriptive information about the Family Structure and Substance Use variables created using the NELS data.

Descriptive Statistics

Dependent Variable: SUBSTANC Use of alcohol, drugs, tobacco

FAMSTRUC Family	Mean	Std. Deviation	N
1.00 Two-parent family	-.0585	.72621	597
2.00 One parent, one guardian	.1196	.76153	94
3.00 Single-parent family	.1918	.93617	142
Total	.0043	.77554	833

Tests of Between-Subjects Effects

Dependent Variable: SUBSTANC Use of alcohol, drugs, tobacco

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
FAMSTRUC	8.594	2	4.297	7.252	.001	.017
Error	491.824	830	.593			
Corrected Total	500.418	832				

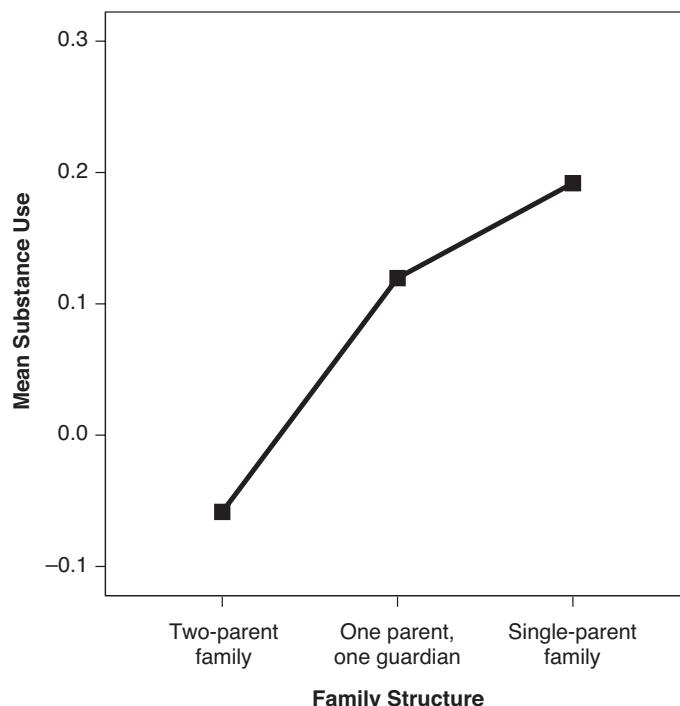


Figure 6.12 Analysis of Variance of the effects of family structure on adolescents' use of dangerous substances.

Figure 6.12 shows the results of an ANOVA using Substance Use as the dependent variable and Family Structure as the independent variable. As the figure shows, the effect of Family Structure was statistically significant ($F[2, 830] = 7.252, p = .001$), although the effect was small ($\eta^2 = .017$). The graph in the figure shows the mean levels of Substance Use by group: students from intact families are less likely, on average, to use substances than are those from families with one parent and one guardian, and students from families with one parent and one guardian are less likely to use substances than those from single-parent families.

Multiple Comparisons

Dependent Variable: SUBSTANC Use of alcohol, drugs, tobacco

	(I) Family Structure	(J) Family Structure	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1.00 Two-parent family	2.00 One parent, one guardian	-.1780*	.08542	.037	-.3457	-.0103
		3.00 Single-parent family	-.2503*	.07187	.001	-.3914	-.1092
		2.00 One parent, one guardian	.1780*	.08542	.037	.0103	.3457
	3.00 Single-parent family	1.00 Two-parent family	.2503*	.07187	.001	.1092	.3914
		2.00 One parent, one guardian	.0723	.10236	.480	-.2732	.1286
		1.00 Two-parent family	.2503*	.07187	.001	.1092	.3914
	Bonferroni	2.00 One parent, one guardian	-.1780	.08542	.112	-.3829	.0269
		3.00 Single-parent family	-.2503*	.07187	.002	-.4227	-.0779
		2.00 One parent, one guardian	.1780	.08542	.112	-.0269	.3829
Dunnett t (2-sided) ^a	2.00 One parent, one guardian	3.00 Single-parent family	-.0723	.10236	1.000	-.3178	.1733
		1.00 Two-parent family	.2503*	.07187	.002	.0779	.4227
		2.00 One parent, one guardian	.0723	.10236	1.000	-.1733	.3178
Dunnett t (2-sided) ^a	2.00 One parent, one guardian	1.00 Two-parent family	.1780	.08542	.073	-.0132	.3692
	3.00 Single-parent family	1.00 Two-parent family	.2503*	.07187	.001	.0894	.4112

Based on observed means.

*. The mean difference is significant at the .05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

Figure 6.13 Post hoc analyses of the effect of three types of family structures on substance use.

Post hoc tests (Fisher's LSD, Dunn–Bonferroni, and Dunnett's test) are shown in Figure 6.13. According to the LSD procedure, the differences between students from intact families and those from single-parent and parent–guardian families were both statistically significant. The difference between parent–guardian and single-parent families was not statistically significant. The comparison between students from two-parent and single-parent families was the only statistically significant difference according to the Dunn–Bonferroni post hoc comparison. For Dunnett's test, two-parent families were used as the reference (or “control”) group. Dunnett's test also suggested that students from single-parent homes use statistically significantly more substances than those from two-parent homes, but that the difference between two-parent and parent–guardian homes was not statistically significant.

Before reading further, take a minute to consider how you could analyze these data using MR. Consider how you would convert Family Structure into dummy variables (and how many dummy variables you would need). How would you convert Family Structure into effect variables? I present the results of such regressions only briefly but ask you to delve more deeply into the analyses in Exercise 2.

Dummy Variable Coding and Analysis

Two dummy variables are needed to capture the information contained in the three categories of the Family Structure variable. Table 6.4 shows my conversion of the Family Structure variable into two dummy variables. As in the Dunnett comparison, I used two-parent families as the reference group for comparison with other family structures. Thus two-parent families are coded zero on both dummy variables. It seems to me that our primary questions of interest in such an analysis will be whether other family structures are comparable to two-parent families. The first dummy variable (Step) contrasts students from parent–guardian families with those from two-parent families, and the second dummy variable (Single) contrasts students from single-parent families with those from two-parent families.

Table 6.4 Converting the Family Structure Variable into Two Dummy Variables

Group	Step	Single
1 Two-parent family	0	0
2 One parent, one guardian	1	0
3 Single-parent family	0	1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.131 ^a	.017	.015	.76978

a. Predictors: (Constant), SINGLE, STEP

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.594	2	4.297	7.252	^a .001 ^a
	Residual	491.824	830	.593		
	Total	500.418	832			

a. Predictors: (Constant), SINGLE, STEP

b. Dependent Variable: SUBSTANC Use of alcohol, drugs, tobacco

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	-.058	.032	-1.855	.064	-.120	.003
	STEP	.178	.085	.073	2.084	.037	.010
	SINGLE	.250	.072	.121	3.483	.001	.109

a. Dependent Variable: SUBSTANC Use of alcohol, drugs, tobacco

Figure 6.14 Analysis of the Substance Use data using multiple regression with dummy variables.

Figure 6.14 shows the results of a MR of Substance Use on these two dummy variables. The two Family Structure dummy variables accounted for 1.7% of the variance in Substance Use ($R^2 = \eta^2 = .017$). Given the large sample, this value was statistically significant ($F[2, 830] = 7.252, p = .001$). The values match those from the ANOVA.

Next, focus on the table of coefficients shown in Figure 6.14. The intercept (constant) is equal to the mean score on Substance Use for the contrast group, the group coded zero in both dummy variables, that is, students from two-parent families. As in the earlier example, the b 's represent the deviation for each group from the mean of the contrast group. So, for example, students from single-parent families had mean Substance Use scores of .1918 (Figure 6.12). From the table of regression coefficients in Figure 6.14, we can calculate the mean score on the dependent variable for students from single-parent families as $-.058 + .250 = .192$, the same value, within errors of rounding.

The interpretation of the t values and their statistical significance is the same as in the earlier example. We could look up the t values in a Dunnett's table and find that the Single dummy variable was statistically significant, whereas the Step dummy variable was not. These findings are also consistent with the findings from the ANOVA and suggest that students from single-parent families use statistically significantly more dangerous substances

than do students from two-parent families but that the difference between parent–guardian families and two-parent families is not statistically significant.

We can also use the t values and their associated statistical significance as the basis for a series of LSD or Dunn–Bonferroni post hoc comparisons. Using Fisher’s LSD, we would simply use the statistical significance of the t ’s as listed in the MR output. We would conclude that students from both single-parent and parent–guardian families are more likely to use dangerous substances than those from traditional two-parent families.

It would be easy to calculate the b associated with the third possible comparison, that between single-parent and parent–guardian families. We know the means for the two groups. If one of these groups was used as the comparison group (e.g., the Parent–Guardian group), the b for the other group (the Single-parent group) would be equal to the difference between these two means. That is, Single–(Parent–Guardian) = .1918 – .1196 = .0722. Unfortunately, with unequal sample sizes, the standard errors associated with each group are different (you can see this by comparing the standard errors associated with Single and Step in the table of coefficients in Figure 6.14). You can calculate the standard error for this comparison using the formula

$$SE_b = \sqrt{MS_r \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

The MS_r is the mean square for the residual from the ANOVA table in Figure 6.14, and the n ’s are the sample sizes for the two groups (the single-parent and parent–guardian groups). For this comparison, SE_b is equal to .102. The value for t ($t = b/SE_b$) is .705, and the probability of obtaining this t by chance alone is .481. Note that this is the same value for the standard error and significance shown in the LSD post hoc comparisons in Figure 6.13. This comparison shows no statistically significant differences in substance use for students from single-parent compared to parent–guardian families. You should perform these calculations yourself to make sure you get the same results that I did.

If you don’t trust yourself to make these calculations by hand, it is easy to get these same results by rerunning the MR. Simply create new dummy variables using the parent–guardian group as the reference group and conduct the regression using these dummy variables. The dummy variable associated with the comparison between the single-parent group and the parent–guardian group should provide the same standard error, t , and p as we calculated above. Whichever method you use, you can use these same t and p values to make post hoc comparisons using a Dunn–Bonferroni correction. For example, you can set the overall family-wise error rate at .05, meaning that each of three comparisons will need to have a probability of less than .0167 ($\alpha = \frac{.05}{3}$) to be considered statistically significant.

Effect Variable Coding and Analysis

With effect coding, one group is assigned values of -1 for all effect coded variables. As shown in Table 6.5, I chose to make two-parent families the group assigned -1 ’s because this is the group I am least interested in contrasting to the average. Figure 6.15 shows the results of the

Table 6.5 Converting the Family Structure Variable into Two Effect-Coded Variables

<i>Group</i>	<i>Step_eff</i>	<i>Single_eff</i>
Two-parent family	-1	-1
One parent, one guardian	1	0
Single-parent family	0	1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.131 ^a	.017	.015	.76978

a. Predictors: (Constant), single_eff, step_eff

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2	4.297	7.252	.001 ^a
	Residual	830	.593		
	Total	882			

a. Predictors: (Constant), single_eff, step_eff

b. Dependent Variable: substanc Use of alcohol, drugs, tobacoo

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.084	.036	2.362	.018	.014	.154
	step_eff	.035	.058			.544	-.079
	single_eff	.108	.052			.006	.149

a. Dependent Variable: substanc Use of alcohol, drugs, tobacoo

Figure 6.15 Multiple regression of Substance Use differences for students from three family types using effect-coded variables.

regression of Substance Use on the two effect coded variables, Single_eff (for single-parent families) and Step_eff (for parent-guardian families).

With unequal numbers in the three groups, the interpretation of the multiple regression is only slightly different from the interpretation with equal n 's. With equal sample sizes, the intercept is equal to the overall mean, across groups, on the dependent variable. With unequal sample sizes, the intercept is equal to the mean of means, or the unweighted means of the three groups. In other words, average the three means shown in Figure 6.12 without respect to the differences in the n 's of the three groups: $(-.0585 + .1196 + .1918)/3 = .0843$. As before, the b 's are the deviation from the mean for the group coded 1 in the effect variable. Thus, the mean on Substance Use for students from single-parent families is $.084 + .108 = .192$. Again, you will delve deeper into this analysis in Exercise 2.

ADDITIONAL METHODS AND ISSUES

There are still additional methods for coding simple or complex categorical variables. Like the methods illustrated here, the various methods produce the same overall results, such as R^2 and its statistical significance, but enable different contrasts among the different levels of the categorical variable. Orthogonal or contrast coding produces orthogonal contrasts among the levels of the categorical variable (usually an a priori rather than a post hoc test). Sequential coding can be used to compare categories that can be ranked in some way, nested coding can be used to compare categories within categories, and there are other possible coding schemes beyond these. In addition, we can have multiple categorical variables, as in a factorial design, and can test for possible interactions among these variables. We will discuss testing for interactions in the next chapter.

Which method of coding should you use? I expect that in most cases our interest in categorical variables will be to include such variables in a regression analysis along with other, continuous variables. Very often, these categorical variables will be “control” variables, which

we need to take into account in our regression but are not of central interest. Sex, region of the country, and ethnic origin often are used in regression analyses as such control variables. Under these circumstances, the simple methods of coding presented here are sufficient, and simple dummy coding will often work well. Dummy coding is also useful if you have an obvious contrast group (such as a control group) to which you wish to compare other groups.

Effect coding is useful when you wish to compare each group with the overall mean of all groups. Suppose, for example, you were interested in whether self-esteem differed across different religious groups. If you wanted to determine whether each religious group differed from the average, overall level of self-esteem, effect coding is a good choice for coding the religion variable. Criterion scaling is especially useful for categorical variables that have numerous categories. Other books may be consulted for further information about some of the more complex coding schemes mentioned (e.g., Cohen et al., 2003; Darlington, 1990; Pedhazur, 1997).

SUMMARY

This chapter introduced the analysis of categorical variables in multiple regression. Categorical, or nominal, variables are common in research, and one advantage of MR is that it can be used to analyze continuous, categorical, or a combination of continuous and categorical independent variables.

With dummy coding, a common method of dealing with categorical variables in MR analyses, the categorical variable is converted into as many dummy variables as there are group categories, minus one ($g - 1$). Thus, if the categorical variable includes four groups, three dummy variables are needed to capture the same information for analysis in MR. Each such dummy variable represents membership (coded 1) versus nonmembership (coded 0) in some category. The contrast group has a value of zero on all dummy variables. As a simple example, a Sex variable could be converted into a dummy variable in which girls are assigned 1 and boys assigned 0; thus the variable represents membership in the category girls. As shown in the chapter, the results of an analysis of the effects of a categorical independent variable (in the example used here, abuse and posttraumatic stress) on a continuous dependent variable are the same whether analyzed via ANOVA or MR. The F associated with the two procedures is the same, and the effect size η^2 from ANOVA is identical to the R^2 from MR. The table of coefficients from the MR may be used to perform post hoc comparisons using several different post hoc procedures.

Dummy coding is not the only method of dealing with categorical variables so that they can be analyzed in MR. With effect coding, one group, often the final group or a less interesting group, is assigned values of -1 on all effect coded variables; in contrast, with dummy coding this group is assigned all zeros. Effect coding contrasts each group's mean on the dependent variable with the grand mean. For criterion scaling, each group is assigned its mean value on the dependent variable as its value on a single criterion scale. So, for example, if boys achieved an average score of 50 on a reading test and girls a score of 53, a criterion scaled version of the Sex variable will assign all boys a value of 50 and all girls a value of 53 in a regression of Reading test scores on Sex. Criterion scaling is useful when there are many categories, because only one variable is needed, rather than $g - 1$ variables. When criterion scaling is used, however, you must correct the ANOVA table produced by the regression because the df will be incorrect (the df still equals $g - 1$). The interpretation of the intercept and regression coefficients for these three methods of coding is summarized in Table 6.6.

Table 6.6 Interpretation of Intercepts and Regression Coefficients Using Different Methods of Coding Categorical Variables

<i>Coding method</i>	<i>Intercept</i>	<i>b</i>
Dummy	Mean on the dependent variable of the reference group (the group coded zero on all dummy variables)	Deviation from the mean for the group coded 1
Effect	Unweighted mean, or mean of the means of the groups on the dependent variable	Deviation from the unweighted mean by the group coded 1
Criterion	Not of interest	Not of interest

Although it is possible to analyze the results of simple and complex experiments in which all independent variables are categorical using multiple regression, it is generally easier to do so via ANOVA. A more common use of categorical variables (and dummy and other coding) in MR analysis is when categorical variables are analyzed in combination with continuous variables in nonexperimental research. A researcher might want to control for Sex, for example, in an analysis of the effects of achievement on self-esteem. This analysis of both categorical and continuous variables in MR is the focus of the next chapter. Before analyzing both types of variables, however, it is necessary to understand how to analyze categorical variables in MR; the present chapter thus served as an introduction to this topic.

EXERCISES

1. The file “false memory data, 4 groups.sav” (or, .xls, or “false2.txt”), available on the Web site (www.tzkeith.com), includes the false memory simulated data analyzed in this chapter, plus data from a fourth group, men who were neither abused nor suffered from PTSD (the four groups from Bremner et al., 2000). For comparison purposes, analyze the data via ANOVA, with follow-up via Fisher’s LSD test, the Dunn–Bonferroni procedure, and Dunnett’s test (with men as the control group).
 - a. Convert the group variable into $g - 1$, or three, dummy variables and analyze the data with MR. Use the table of coefficients to conduct the three post hoc procedures. Compare the results with the ANOVA.
 - b. Convert the group variable into three effect coded variables and analyze the data with MR. Compare the results with the ANOVA and with the dummy coded solution.
 - c. Convert the group variable into a single criterion scaled variable and conduct the MR using it. Correct the ANOVA table from the MR for the proper degrees of freedom and compare the results with the other analyses of the same data.
2. Conduct the analyses of the effect of Family Structure on students’ Substance Use as outlined in this chapter using the NELS data. This is one of the more complex exercises you will do, because it requires the creation of several new variables. It is also probably one of the more realistic examples. I suggest you team up with a classmate as you work on it.
 - a. Create the Family Structure and Substance Use variables (see note 7). Examine descriptive statistics for each variable, and compute means and standard deviations of Substance Use by Family Structure.
 - b. Create dummy variables contrasting students from two-parent families with those from parent–guardian families and those from single-parent families. Regress

- Substance Use on these dummy variables. Interpret the overall regression. Use the table of coefficients to conduct post hoc testing. Make sure you compare single-parent families to parent-guardian families.
- c. Create effect variables with two-parent families as the group coded –1 on all variables. Regress Substance use on these effect variables and interpret the regression results.
 - d. Convert the Family Structure variable into a single criterion scaled variable and conduct the MR using it. Correct the ANOVA table from the MR for the correct degrees of freedom and compare the results with the other analyses of the same data.
 3. The file “homework experiment data.sav” (or “homework experiment.xls”) includes data from a simulated experiment in which children were given different types of homework. Sixth graders were randomly assigned to one of three groups (the Type variable in the data): group 1 was assigned drill sheet-derived homework in Social Studies at least three times per week. Group 2 was assigned practice homework (homework designed to practice the important concepts from that day’s lesson in Social Studies), also at least 3 times per week. Group 3 was assigned extension homework on the same schedule. Extension homework is designed to extend the lesson taught in school, often with additional content. At the end of six weeks students were administered a standardized measure of sixth-grade Social Studies achievement ($M=50$, $SD=10$).
 - a. Consider drill homework to be the norm. Analyze the results of this experiment using dummy coding, and comparing the other two types of homework with this norm.
 - b. Analyze the experiment using effect coding. Which group did you choose as the contrast group (the group with –1s for the effect-coded variables)? Explain why.

Notes

- 1 For now, we will postpone exploring what it means to say that sex affects self-esteem or religious affiliation influences voting behavior. We will address this issue in the next chapter.
- 2 Why the name “dummy” variables? Dummy means a stand-in, representation, a copy. Think of a store mannequin, rather than the derogatory slang usage of dummy.
- 3 The actual study included other measures of memory, an additional group (Nonabused, non-PTSD men) and unequal n ’s across groups. The data presented here are simulated, but are designed to mimic those in the original article (Bremner et al., 2000).
- 4 This discussion should make it obvious that our normal interpretation of the t ’s associated with regression coefficients does not make adjustments for the number of comparisons. Darlington (1990, p. 257) noted that normal multiple regression practice falls under the “Fisher Protected t ” method, whereby if the overall R^2 is statistically significant we can make all the individual comparisons represented by the t tests of each regression coefficient. Our discussion of post hoc tests and correcting for error rates is intended as a brief introduction only. Darlington is an excellent resource for more information about multiple comparison procedures in MR.
- 5 If there are different numbers of cases for each category, the standard errors of the b ’s will differ for each comparison.
- 6 Astute readers will recognize that the PTSD example is actually an example of nonexperimental research, since women were not assigned to the different groups but were sampled from preexisting groups. In the actual research, there were unequal numbers of participants in each category (Bremner et al., 2000).
- 7 Both variables were created from other NELS variables. Substance use (Substance) was the mean of variables F1S77 (How many cigarettes smoked per day), F1S78a (In lifetime, number of times had alcohol to drink), and F1S80Aa (In lifetime, number of times used marijuana). Because these variables used different scales, they were standardized (converted to z scores) prior to averaging. Family Structure (FamStruc) was created from the existing NELS variable FamComp (Adult composition of the household). FamComp was coded 1 = Mother & father, 2 = Mother and male guardian, 3 = Female guardian and father, 4 = Other two-adult families, 5 = Adult female only, and 6 = Adult male only. For Family Structure, category 1 was the same as for FamComp, categories 2 and 3 were combined, categories 5 and 6 were combined, and category 4 was set to a missing value.

Regression With Categorical and Continuous Variables

Sex, Achievement, and Self-Esteem	130
Interactions, AKA Moderation	132
<i>Testing Interactions in MR</i>	133
<i>Centering and Cross Products: Achievement and Sex</i>	133
<i>The MR Analysis</i>	134
<i>Interpretation</i>	135
A Statistically Significant Interaction	137
<i>Does Achievement Affect Self-Esteem? It Depends</i>	137
<i>Understanding an Interaction</i>	138
<i>Extensions and Other Examples</i>	140
<i>Testing Interactions in MR: Summary</i>	141
Specific Types of Interactions Between Categorical and Continuous Variables	141
<i>Test (and Other) Bias</i>	142
<i>Aptitude-Treatment Interactions</i>	150
<i>ANCOVA</i>	153
Caveats and Additional Information	154
<i>“Effects” of Categorical Subject Variables</i>	154
<i>Interactions and Cross Products</i>	155
<i>Further Probing and Figural Display of Statistically Significant Interactions</i>	155
Summary	158
Exercises	159
<i>Notes</i>	160

You should now have a firm grasp on how to analyze continuous variables using multiple regression, along with a new appreciation of how to analyze categorical variables in MR. In this chapter we will combine these two types of variables to analyze both categorical and continuous variables in a single MR. Our discussion starts with the straightforward analysis of both types of variables in a single multiple regression. We then turn to focus on the addition of *interactions* to such analyses. Specific types of interactions between continuous and categorical variables are often of particular interest to psychologists and other social science researchers: aptitude-treatment interactions and bias in the predictive validity of tests. We cover examples of such analyses. In the next chapter we will expand our

discussion of interactions to cover interactions of two continuous variables and the analysis of potential curvilinear effects.

SEX, ACHIEVEMENT, AND SELF-ESTEEM

Much has been written about differences in self-esteem among adolescents; research and conventional wisdom suggests that girls' self-esteem suffers, compared to boys, during adolescence (Kling, Hyde, Showers, & Buswell, 1999; Rentzsch, Wenzler, & Schütz, 2016). Will we find self-esteem differences between 10th-grade boys and girls in the NELS data? Will any differences persist once we take into account previous achievement?

To address these questions, I regressed 10th-grade self-esteem scores (F1Cncpt2) on Sex (Female) and Previous Achievement (ByTests). (Question: is it necessary to include ByTests for the regression to be valid?) Sex was converted into a dummy variable, Female, coded 0 for boys and 1 for girls. For this analysis, I also converted the existing Self-Esteem variable (which was a mean of z scores) into T scores ($M = 50$, $SD = 10$); the new variable is named S_Esteem in subsequent figures (it is not in the NELS data on the Web site, but you can easily create it¹).

Figure 7.1 shows the basic descriptive statistics for the variables in the regression. All statistics are consistent with the intended coding of the variables. Figure 7.2 shows some of the results of the simultaneous regression of Self-Esteem on Achievement and Female.

The interpretation of the regression is straightforward and consistent with our previous such interpretations. The two independent variables explained 2.6% of the variance in Self-Esteem, which, although small, is statistically significant ($F = 12.077$ [2, 907], $p < .001$). Achievement had a moderate, statistically significant, effect on Self-Esteem. The Female dummy variable was also statistically significant, and its negative sign ($b = -2.281$) means

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
S_ESTEEM	910	15.23	69.47	49.9602	10.01815
Female sex as dummy variable	910	.00	1.00	.4912	.50020
BYTESTS 8th-grade achievement tests (mean)	910	29.35	70.24	51.5758	8.76712
Valid N (listwise)	910				

Correlations

		S_ESTEEM	Female sex as dummy variable	BYTESTS 8th-grade achievement tests (mean)
Pearson Correlation	S_ESTEEM Female sex as dummy variable BYTESTS 8th-grade achievement tests (mean)	1.000 -.106 .114	-.106 1.000 .064	.114 .064 1.000

Figure 7.1 Descriptive statistics for the regression of Self-Esteem on Sex (female) and Previous Achievement.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.161 ^a	.026	.024	9.89826

a. Predictors: (Constant), BYTESTS 8th-grade achievement tests (mean), Female sex as dummy variable

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2366.444	2	1183.222	12.077	.000 ^b
	Residual	88863.783	907	97.976		
	Total	91230.227	909			

a. Dependent Variable: S_ESTEEM

b. Predictors: (Constant), BYTESTS 8th-grade achievement tests (mean), Female sex as dummy variable

Coefficients^a

Model		Unstandardized Coefficients		Beta	t	Sig.	95% Confidence Interval for B	
		B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	43.924	1.969		22.306	.000	40.059	47.788
	Female sex as dummy variable	-2.281	.658	-.114	-3.468	.001	-3.572	-.990
	BYTESTS 8th-grade achievement tests (mean)	.139	.038	.121	3.698	.000	.065	.212

a. Dependent Variable: S_ESTEEM

Figure 7.2 Simultaneous regression results: Self-esteem on Female and Previous Achievement.

that girls indeed scored lower than boys on the measure of Self-Esteem, even after controlling for Achievement (with the variable named Female it is easy to recall that the variable was coded so that boys = 0 and girls = 1). The value for the unstandardized regression for the Female variable suggests that girls scored, on average, 2.28 points lower than did boys (after Achievement is controlled). Concerning our question of interest, the findings suggest that 10th-grade girls do have slightly but statistically significantly lower self-esteem than do boys at the same grade level (although the findings do not help us understand why this difference exists).

In Chapter 6 we focused on the meaning of the intercept and the b's using dummy variables. For the present example, with one categorical and one continuous variable, the intercept was 43.924. As in previous examples, the intercept represents the predicted Self-Esteem for those with a value of zero on each predictor variable. Thus the intercept represents the predicted Self-Esteem score of boys (coded zero) with a score of zero on the Achievement test. The intercept is not particularly useful in this case, however, since the actual range of the achievement test was only approximately 29 to 70, with no scores of zero. The *b* for Female of -2.281 means, again, that girls scored an average of 2.28 points lower than boys on the Self-Esteem measure. This integration of categorical and continuous variables is straightforward.

Figure 7.3 displays these results in path form. Unstandardized coefficients are used because they are generally more useful with dummy variables. The -2.281 means that for each one unit increase (going from being a boy, coded 0, to being a girl, coded 1), Self-Esteem decreases by 2.281 points. The coefficient for the curved, double-headed arrow is the covariance, the unstandardized version of a correlation (see Chapter 1).

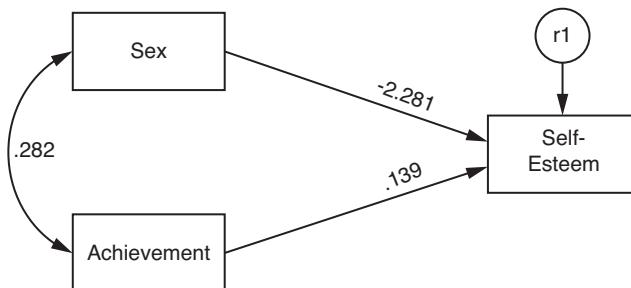


Figure 7.3 Figural (path) display of the effects of Sex and Achievement on the Self-Esteem of 10th-graders.

INTERACTIONS, AKA MODERATION

As you will discover in this section, it is also possible using MR to test for potential interactions between categorical and continuous variables (and also between several categorical variables or several continuous variables). Interactions are those instances when the effect of one variable depends on the value of another variable. In experimental research, we may find that the effect of a treatment depends on the sex of the participant; a cholesterol drug may be more effective in lowering the cholesterol of males than females, for example. To use an example from earlier chapters, we may find that homework is more effective for students with high levels of academic motivation compared to those with lower levels of motivation. In other words, the effect of one variable (homework) depends on the value of another variable (academic motivation).

What might such interactions look like? In our previous example, we found that 10th-grade girls had slightly, but statistically significantly, lower self-esteem than did boys in the same grade. Previous Achievement also had an apparent effect on 10th-grade Self-Esteem. Could it be, however, that there are different effects for achievement on self-esteem for boys as compared to girls? For example, perhaps girls' self-image is closely related to their school performance, with higher achievement leading to higher self-esteem. In contrast, it may be that achievement works differently for boys and that their self-esteem is unrelated to their school performance. Speaking as the father of three sons, I find this latter possibility fairly plausible.

This type of differential relation between achievement and self-esteem is illustrated (in an exaggerated way) in Figure 7.4. In the figure, the two lines represent the possible

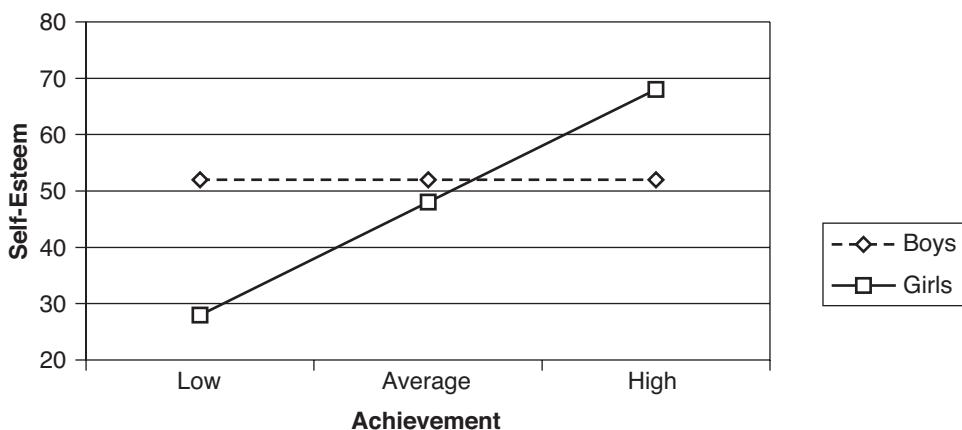


Figure 7.4 Graphic display of a possible interaction between Sex and Achievement on Self-Esteem.

regression lines from the regression of Self-Esteem on Achievement for boys and girls, respectively. Such graphs are an excellent way to understand interactions. The graph also illustrates that interactions occur when the slope of the regression line for one independent variable (Achievement) differs depending on the value of the other independent variable (Sex). Again, the graph illustrates a potential interaction between Sex and Achievement in their effects on Self-Esteem: for girls, Achievement has a strong effect on Self-Esteem, whereas it has no effect for boys. There are several ways to describe such an interaction. We could say that Sex and Achievement interact in their effect on Self-Esteem; that Achievement has differential effects on Self-Esteem depending on the Sex of the student; that Achievement has stronger effects on the Self-Esteem of girls than of boys; or that Sex *moderates* the effect of Achievement on Self-Esteem (cf. Baron & Kenny, 1986). A colleague of mine notes that most interactions can be described using the phrase “it depends.” Does achievement affect self-esteem? It depends; it depends on whether you are a boy or a girl.

Methodologists often distinguish between ordinal and disordinal interactions. Interactions in which the regression lines cross within the effective range of the independent variable are often referred to as disordinal interactions. Figure 7.4 illustrates a disordinal interaction; the two lines cross within the effective range of the independent variable (Achievement). With an ordinal interaction, the lines representing the effects do not cross within the effective range of the independent variable. Figure 8.3 in the next chapter illustrates an ordinal interaction.

Testing Interactions in MR

We test such interactions in multiple regression by creating cross-product variables and testing whether these cross-product terms are statistically significant when added to the regression equation. Cross-product terms are created by multiplying the two variables of interest (Cohen, 1978); in this case we would create the cross product by multiplying the Female variable and the Achievement variable.² Although this product term will work as an interaction term, there are interpretive advantages to first centering the continuous variables prior to multiplication (Aiken & West, 1991; Cohen et al., 2003; Hayes, 2018; Darlington & Hayes, 2017). These include the reduction in unnecessary collinearity, the creation of a zero point for the continuous scales that do not already have one, and the ease in interpreting all the resulting regression coefficients. Centering is most easily accomplished by subtracting the mean score of the variable from that variable (e.g., using a compute statement in SPSS), thus resulting in a new variable with a mean of zero and a standard deviation equal to the original standard deviation.

Centering and Cross Products: Achievement and Sex

For the current example, in which the interest was testing the possible interaction between sex (Female) and Achievement on Self-Esteem, I created two new variables. Ach_Cent was created as a centered version of the base year achievement tests by subtracting the mean for ByTests (51.57575; one more decimal than shown in Figure 7.1) from each person’s score on ByTests. Thus, students with an original achievement score of, for example, 30 will have a score of -21.57575 on Ach_Cent, whereas those with an original score of 70 will have a score of 18.42425 on Ach_Cent. Sex_Ach, the cross product, was created by multiplying Female (the sex dummy variable) and Ach_Cent.

Figure 7.5 shows the descriptive statistics for these new variables; as you can see, the mean of Ach_Cent is zero. Also shown in the figure are the correlations among the

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
S_ESTEEM	910	15.23	69.47	49.9602	10.01815
Female sex as dummy variable	910	.00	1.00	.4912	.50020
ACH_CENT BY achievement, centered	910	-22.23	18.66	.0000	8.76712
SEX_ACH Sex by Achieve interaction	910	-22.23	17.51	.2814	5.91969
Valid N (listwise)	910				

Correlations

		S_ESTEEM	Female sex as dummy variable	ACH_CENT BY achievement, centered	SEX_ACH Sex by Achieve interaction
Pearson Correlation	S_ESTEEM	1.000	-.106	.114	.103
	Female sex as dummy variable	-.106	1.000	.064	.048
	ACH_CENT BY achievement, centered	.114	.064	1.000	.677
	SEX_ACH Sex by Achieve interaction	.103	.048	.677	1.000

Figure 7.5 Descriptive statistics for the test of a possible interaction between Sex and Achievement in their effects on Self-Esteem.

variables. The centered variable has the same correlations with the other variables as does the original, uncentered variable (Figure 7.1). The interaction term (Sex_Ach), however, correlates at different levels with components (the Female and Achievement variables) than would an interaction created from a noncentered variable. For example, an interaction term built on the uncentered Achievement variable would have correlated .975 with Female, whereas the interaction term built on the centered Achievement variable correlated .048 with Female (Figure 7.5). The very high correlation between these two predictor variables is termed collinearity or multicollinearity. Collinearity can result in strange coefficients and large standard errors and make interpretation difficult. As noted by Cohen and colleagues, centering does not eliminate collinearity but reduces unnecessary collinearity (2003); see also Hayes (2018) for an in-depth discussion about centering. The topic and effects of multicollinearity will be discussed in more depth in Chapter 10.

The MR Analysis

To test the statistical significance of the interaction, Self-Esteem was first regressed on Female and Achievement (centered). These variables were entered using simultaneous multiple regression, a step similar to the first example from this chapter, but this was also the first step in a sequential multiple regression. As shown in Figure 7.6, these variables accounted for 2.6% of the variance in Self-Esteem (the same as in Figure 7.2). In the second block in this sequential regression, the cross product (Sex_Ach) was added to the equation. As shown, the addition of the cross product did not lead to a statistically significant increase in R^2 ($\Delta R^2 = .001$, $F[1, 906] = 1.218$, $p = .270$). This means that the interaction is not statistically significant: the interaction term does not help explain Self-Esteem beyond the explanation

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.161 ^a	.026	.026	12.077	2	907	.000
2	.165 ^b	.027	.001	1.218	1	906	.270

a. Predictors: (Constant), ACH_CENT BY achievement, centered, Female sex as dummy variable

b. Predictors: (Constant), ACH_CENT BY achievement, centered, Female sex as dummy variable, SEX_ACH Sex by Achieve interaction

Figure 7.6 Test of the interaction between Sex (Female) and Achievement in their effects of Self-Esteem.

provided by Female and Achievement. We cannot reject the null hypothesis that Achievement has the same effect for girls as for boys. Thus, contrary to our speculation, there appears to be no differential effect for Achievement on the Self-Esteem of girls as compared to boys; Achievement has the same magnitude of effect on boys' and girls' achievement.

How, you may wonder, does this method of testing an interaction relate to our graphical display of an interaction, such as that shown in Figure 7.4? The figure essentially shows separate regression lines for boys and girls, but we have not conducted separate regressions for the two groups. The method of testing an interaction illustrated here is equivalent to conducting separate regressions for boys and girls. We could, for example, regress Self-Esteem on Achievement separately for boys and girls and then compare the regression coefficients for Achievement for boys versus girls. If the interaction is statistically significant, these coefficients will be quite different for boys versus girls; in Figure 7.4, for example, the regression coefficient for girls will be large and statistically significant, whereas the coefficient for boys will be small and not statistically significant. The fact that the interaction in the current example is not statistically significant means that, in fact, the regression lines for boys and girls are close to parallel (not statistically significantly non-parallel). (The lines are parallel but not identical because the intercepts differ. This difference is not tested in the interaction term, which simply tests whether the lines are parallel.) The addition of an interaction term to the model is equivalent to testing separate models for different groups. The method of testing a cross product does this in one step, however, and also tests the statistical significance of the interaction.

Interpretation

Given that the interaction was not statistically significant, I would focus my interpretation on the coefficients from the first step of the multiple regression, prior to the addition of the interaction term. I would certainly report that the interaction was tested for statistical significance, as shown previously, and found lacking but then would turn my interpretation to the equation without the interaction term, as shown in Figure 7.7. With the centering of the Achievement variable, the intercept for the regression equation has changed. The intercept still represents the predicted Self-Esteem score for someone with zeros on each independent variable, but with centering a score of zero on the Achievement test represents the overall mean of Achievement for these students. Thus, the intercept now represents the predicted Self-Esteem score for boys who score at the mean (for the total sample) on the Achievement test (this ease of interpretation is an advantage of centering). The regression coefficients are

Model	Coefficients ^a							
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	51.081	.460	110.929	.000	50.177	51.984	
	Female sex as dummy variable	-2.281	.658	-.114	-.3468	.001	-3.572	-.990
	ACH_CENT BY achievement, centered	.139	.038	.121	3.698	.000	.065	.212

a. Dependent Variable: S_ESTEEM

Figure 7.7 Regression coefficients: effects of Female and Achievement on Self-Esteem.

Table 7.1 Effects of Sex and Achievement on the Self-Esteem of 10th Graders

Variable	β	$b (SE_b)$	p
Sex (Female)	-.114	-2.281 (.658)	.001
Achievement	.121	.139 (.038)	< .001

the same as shown in Figure 7.2, because centering does not change the standard deviations of the variables [and $b = \beta (SD_y)/(SD_x)$]. Here's a potential interpretation of these findings:

This research had two purposes. First, we were interested in the effect of sex on 10th-grade students' self-esteem. In particular, we tested whether girls have lower self-esteem in 10th-grade than do boys, after controlling for prior achievement. Previous research has suggested that achievement has differential effects on the self-esteem of boys versus girls (not really, I just made this up)—that sex moderates the effect of achievement such that achievement influences the self-esteem of girls but not boys. The second purpose of this research was to test for this possible differential effect.

Self-Esteem was regressed on Sex and prior Achievement to address the first purpose of this research. A cross-product term (Sex \times Achievement) was added next to the model to test the possible interaction between Sex and Achievement in their effects on Self-Esteem (cf. Aiken & West, 1991; Cohen, 1978); the Achievement variable was centered.

Sex and prior Achievement together accounted for 2.6% of the variance in 10th-grade Self-Esteem ($F[2, 907] = 12.077, p < .001$). The interaction was not statistically significant, however ($\Delta R^2 = .001, F[1, 906] = 1.218, p = .270$), suggesting that Achievement has the same effect on the Self-Esteem of both boys and girls.

The regression coefficients in Table 7.1 show the extent of the influence of Sex (Female, with girls = 1 and boys = 0) and Achievement on Self-Esteem. The effect of Sex on Self-Esteem was indeed statistically significant, and girls scored, on average, 2.28 points lower on the Self-Esteem scale than did boys, even after prior Achievement was controlled statistically. This effect of Sex can be considered a small to moderate effect. Although these results show that adolescent girls have lower self-esteem than do boys, the results do not illuminate why this may be so or what aspect of being a girl as opposed to a boy leads to lower self-esteem. Achievement also had a moderate and statistically significant effect on subsequent Self-Esteem. Thus, Achievement appears to have the same effect on the Self-Esteem of boys and girls and this effect is of moderate magnitude and statistically significant for both groups.

This example illustrates the basic method of testing interactions (moderation) in multiple regression. And, although I have introduced the method in the context of an interaction between a categorical and a continuous variable, the method is the same for testing interactions between continuous variables (as illustrated in the next chapter) or between categorical variables. The example also illustrates the simple fact that such interactions are not very common in nonexperimental research, especially with small to medium sample sizes. There are several reasons for the infrequent finding of interactions in nonexperimental research. First, the nature of testing for interactions focuses on the unique effects attributable to the interaction after the variation due to the original variables has been statistically removed (e.g., in a sequential regression). Second, it is also the case that measurement error (unreliability and invalidity) reduces the statistical power to detect interactions in MR (Aiken & West, 1991); the unreliability of the interaction term is a product of the unreliability of both its components. As a result, tests of interactions in MR are simply less sensitive than tests to detect main effects. Third, simulation research has shown that when the assumption of homogeneity of error variances across groups is violated (such assumptions will be discussed in more detail in Chapter 10), the power to detect interactions can vary considerably. This variability can be especially problematic when sample sizes vary across groups (Alexander & DeShon, 1994). Finally, “It may be that substantial interaction effects just rarely exist in the real world” (Darlington & Hayes, 2017, p. 429). For these reasons, I recommend testing for interactions primarily when testing specific hypotheses. That is, I do not recommend testing all possible interactions but instead testing those suggested in previous research or those designed to answer specific research questions. So, for example, in their examination of the effects of parent involvement, homework, and TV viewing on achievement, Keith, Reimers, Fehrman, Pottebaum, and Aubey (1986) tested the interaction between TV viewing and ability because previous research had suggested the presence of such an interaction (this example is discussed in more detail in the next chapter). Adequate to large sample sizes are also needed (Alexander & DeShon, 1994).

A STATISTICALLY SIGNIFICANT INTERACTION

Another example will illustrate a statistically significant interaction. Based on theory, previous research, or even persuasive argument, we might suspect that achievement interacts with students’ ethnic background (rather than sex) in its effect on subsequent self-esteem. Just as we speculated for boys versus girls, we might speculate that achievement has a positive effect on the self-esteem of white youth but that achievement is relatively unimportant for the self-esteem of youth from various ethnic minority groups. To test this hypothesis, the Race variable in the NELS data was recoded into a new minority-white (Minority) variable with white non-Hispanic students coded 0 and members of all other ethnic groups coded 1. Minority and a centered version of the Achievement variable (Ach_Cen2, BYTests-51.589590) were multiplied to create an Ethnic background by previous Achievement cross-product term.³

Does Achievement Affect Self-Esteem? It Depends

Figure 7.8 shows some of the results of the sequential regression to test the significance of the interaction. In the first block, Self-Esteem was regressed on Minority and previous Achievement (Ach_Cen2); in the second step, the Minority–Achievement interaction term was entered (Eth_Ach). The addition of the interaction term resulted in a statistically significant increase in variance explained ($\Delta R^2 = .008$, $F[1, 896] = 7.642$, $p = .006$); in other words, the interaction is statistically significant.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.148 ^a	.022	.020	9.94243	.022	10.089	2	897	.000
2	.174 ^b	.030	.027	9.90582	.008	7.642	1	896	.006

- a. Predictors: (Constant), Minority Minority vs white, ach_cen2 Achievement, centered
 b. Predictors: (Constant), Minority Minority vs white, ach_cen2 Achievement, centered, eth_ach
 Ethnicity achievement interaction

Figure 7.8 Test of the interaction between Ethnic background and Achievement in their effects on Self-Esteem using sequential regression.

Understanding an Interaction

Given a statistically significant interaction, what does it mean? Probably the easiest way to understand an interaction (in multiple regression or ANOVA) is to graph it. Recall that an interaction between a categorical and continuous variable in multiple regression represents different regression coefficients for the two (or more) groups. When we say that Achievement and Minority interact in their effect on Self-Esteem, we mean that Achievement has different effects on the Self-Esteem for (in this example) ethnic minority members and non-minority White adolescents. These different effects mean the regression coefficients (*b*'s) associated with Achievement will be different for the two groups if we conduct separate regressions (Self-Esteem regressed on Achievement) for minority and White youth. Different regression coefficients further mean that the slopes of the regression lines will be different for the two groups (because the *b* is the slope of the regression line). What we need, then, is a graph of the regression lines of Self-Esteem on Achievement for minority youth and white youth separately.

Fortunately, it is relatively easy to produce such graphs using standard statistical analysis programs. Figure 7.9 shows such separate regression lines for Ethnic minority and white youth, (created using the SPSS Scatterplot command, followed by some touch-up). As shown in the graph, it indeed appears that Achievement has a positive effect on the Self-Esteem of white non-Hispanic youth, but that it has little effect, or perhaps a small negative effect, on the Self-Esteem of ethnic minority group youth. If these findings are correct, they suggest that improving the achievement of white adolescents will result in increased self-esteem. For ethnic minority youth, however, it appears that increased achievement will result in no increases in self-esteem.

Further Analysis

We are now faced with additional questions; although we know that the interaction is statistically significant, and we have an understanding of the nature of the interaction, we don't know whether the effects of Achievement on Self-Esteem are statistically significant for the two groups. This lack of clarity is especially obvious when looking at the regression line for minority students; does Achievement have no statistically significant effect on Self-Esteem for minority students, or does it actually have a negative effect? To investigate further, we can easily conduct the two separate regression analyses represented by the regression lines in Figure 7.9. (This is similar to conducting a test of simple main effects in ANOVA to probe a statistically significant interaction. The problem with this approach is that the standard errors of the *b*'s will be slightly off, and thus a coefficient may seem statistically significant when it is not or vice-versa. We will return to this problem in the next example. Also see the supplemental material on tzkeith.com for Chapter 7.) Nevertheless, if we conduct separate

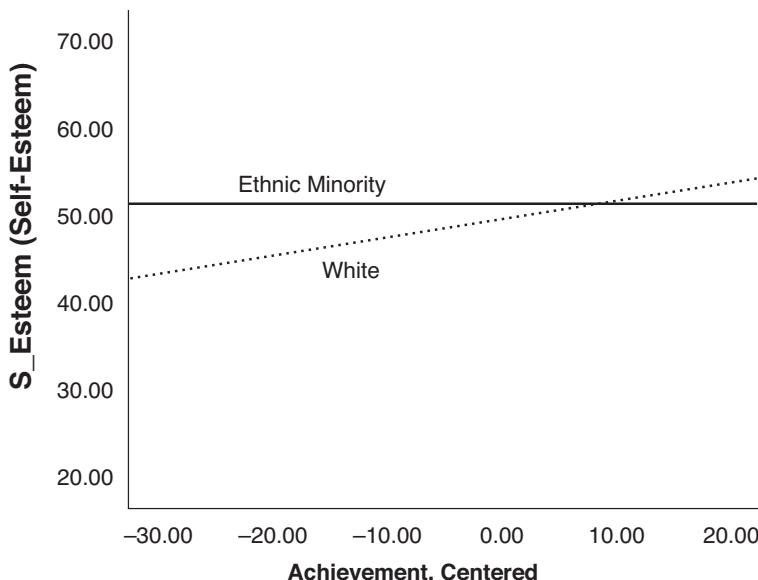


Figure 7.9 Regression lines illustrating the interaction of Ethnic background and Achievement in their effects on Self-Esteem.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
2	(Constant)	49.250	.392	125.555	.000	48.480	50.020
	Minority Minority vs white	1.627	.786	.072	.2069	.039	.084
	ach_cen2 Achievement, centered	.230	.046	.201	4.967	.000	.139
	eth_ach Ethnicity achievement interaction	-.237	.086	-.114	-2.764	.006	-.405
							-.069

a. Dependent Variable: s_esteem

Figure 7.10 Regression coefficients: effects of Ethnic background, Achievement, and their interaction on Self-Esteem.

regressions, we find that the regression of Self-Esteem on Achievement was statistically significant for white adolescents but not statistically significant for minority youth.

It is also possible to calculate the coefficients for these separate regression equations from the coefficients from the overall regression including the interaction term (Figure 7.10). This figure shows the lower half of the table of coefficients from the multiple regression (once the cross-product has been entered into the regression). The intercept for the regression with the interaction term included is equivalent to the intercept for a separate regression of Self-Esteem on Achievement *for the group coded zero*. Thus, the intercept for the regression of Self-Esteem on Achievement for white students is 49.250. The intercept for the group coded 1 is equal to the overall intercept plus the coefficient for the minority categorical (dummy-coded) variable ($49.250 + 1.627 = 50.877$). So, if we were to conduct separate regressions, the intercept for minority students would be 50.877 and the intercept for white students would be 49.250. Likewise, the *b* for Achievement in a separate regression for the group coded zero is equal to the *b* for Achievement for the overall regression with the interaction term. Thus, the *b* for white students is the same as the *b* for Achievement (Ach_cen2) in Figure 7.10 (.230). The *b* for minority students (the group coded 1), in turn, is equal to the *b* for Achievement from Figure 7.9 plus

the b associated with the interaction term ($.230 + [-.237] = -.007$). Thus, if we were to conduct separate regressions for white and minority youth, the equation for ethnic minority youth would be $\text{Self-Esteem}_{\text{predicted}} = 50.877 - .007 \text{ Achievement}$, and the equation for white students would be $\text{Self-Esteem}_{\text{predicted}} = 49.250 + .230 \text{ Achievement}$. These interpretations of the various coefficients are summarized in Table 7.2. Try conducting separate regressions of Self-Esteem on Achievement by ethnic background to see if your results match these. It is also possible to calculate the statistical significance of these separate regression coefficients from the overall regression output, but these calculations are more involved (see Aiken & West, 1991, for more information). (For more information, caveats, and a method of coding that also tests the statistical significance of the separate regression equations, see the website www.tzkeith.com. Note also that the PROCESS macro discussed in more detail in Chapter 9 is also useful for such moderation analyses. It also tests the statistical significance for the separate regressions. See Hayes, 2018.) One final thing to note before moving to additional examples: All of the interpretations of the regression coefficients in these analyses have focused on the unstandardized coefficients; the standardized coefficients are generally not interpreted in interaction/moderation analyses.

Table 7.2 Using the Regression Coefficients from the Overall Regression to Develop Separate Regression Equations by Group when there is a Significant Interaction between a Categorical and Continuous Variable. All Coefficients are from the Block that Includes the Cross-Product

Coefficient	Interpretation
Intercept	Intercept for the group coded zero
Regression coefficient for the dummy variable	Difference in intercept for the other group
Regression coefficient for the continuous variable	Regression coefficient (slope) for the group coded zero
Regression coefficient for the cross-product	Difference in the regression coefficient (slope) for the other group

Extensions and Other Examples

Note that I have illustrated the simplest of examples, one with a single continuous variable and a single categorical variable; the categorical variable also included only two categories. Extensions are straightforward. We could easily have included several other variables in the analysis, such as students' SES, or their sex as well as their ethnic background. We could have included interaction terms with these variables as well (e.g., $\text{Achievement} \times \text{SES}$ or $\text{Sex} \times \text{Ethnic Background}$). Recall, however, that I earlier recommended that you not conduct a "fishing expedition" for interactions but only include those terms that you are specifically interested in testing and have some reason to test (e.g., to test a specific hypothesis). Likewise, we could have left the ethnic background variable as a mult categorial variable (Asian-Pacific Islander, Hispanic, Black not Hispanic, White not Hispanic, and American Indian-Alaskan Native), in which case we would have needed to create four dummy variables and four interaction terms (each dummy variable multiplied times the centered achievement variable). These four interaction terms would then have been added in one block in the second step of a sequential regression to test the statistical significance of the interaction. The essentials of the analysis would be the same in these cases.

You should be aware that social scientists may use different terms (other than interaction) for this process. The most common term you will likely encounter is testing for *moderation*.

Thus in the present example, we could say that we tested whether Ethnic background moderated the effect of Achievement on Self-Esteem. Other terms are also possible. For example, Krivo and Peterson (2000) investigated whether the variables that affect violence (homicide rate) have the same magnitude of effect for African Americans and whites. In other words, the authors tested whether ethnic background interacted with a series of influences in their effects on violence, but they did not label these as tests of potential interactions. Yet anytime researchers suggest a difference in magnitude on effects (b 's) across groups, they are, in fact, suggesting a potential interaction between a categorical and continuous variable(s). This example is interesting in a number of other ways as well. To test the primary questions, the authors conducted separate regressions across groups, rather than using a series of interaction terms, presumably because they were interested in the potential interaction of all variables with ethnic background. The authors (correctly) used the *unstandardized* coefficients to compare the influences across groups. Within the separate models for African American participants, however, the authors added several interaction (cross-product) terms and labeled them as such.

Testing Interactions in MR: Summary

As a review, these are the steps involved in testing for an interaction between a categorical and a continuous variable in multiple regression:

1. Center the continuous variable expected to interact with a categorical variable by creating a new variable in which the mean of this variable is subtracted from each person's score on the variable.
2. Multiply the centered variable by the dummy variable(s) to create cross-product (interaction). Other types of coding, such as effect coding, can also be used, although the interpretation will be different (see tzkeith.com for an example). It is also possible to center the dummy/categorical variable, although we have not done so here.
3. Regress the outcome variable on the independent variables of interest using simultaneous regression. Use the centered versions of relevant variables, but exclude the interaction terms.
4. Add, in a sequential fashion, the interaction term(s). Check the statistical significance of the ΔR^2 to determine whether the interaction is statistically significant. If the ΔR^2 is statistically significant, graph the interaction. Follow up by calculating the separate regression equations for each group, or by conducting separate regressions for each level of the categorical variable. Interpret the unstandardized coefficients.
5. If the ΔR^2 is not statistically significant, interpret the findings from the first portion of the multiple regression (before the addition of the interaction term).

SPECIFIC TYPES OF INTERACTIONS BETWEEN CATEGORICAL AND CONTINUOUS VARIABLES

Several specific types of interactions between categorical and continuous variables are often of interest in psychology, education, public policy, and other social sciences. A psychologist may be interested in whether a psychological test is biased against ethnic minority students in predicting various outcomes. More broadly, a policy maker may be interested in whether women are underpaid compared to men with the same level of experience and productivity. An educator may be interested in whether an intervention is more effective for teaching children who have high aptitude in some area versus those with lower aptitude in the same area. Each example can be examined via multiple regression by testing for an interaction between a categorical and a continuous variable.

Test (and Other) Bias

Psychological, educational, and other tests should be unbiased and fair for all people who take them. One type of bias to be avoided is bias in predictive validity; in other words, if a test is designed to predict some related outcome, it should predict that outcome equally well for all groups to which it may be given. The Scholastic Aptitude Test (SAT), for example, is designed, in large part, to determine which students will do well in college and which will not, and thus colleges use the SAT to select students based on its predictive power. If the SAT were a better predictor of college GPA for girls than for boys, then a potential student would have a differential chance of being selected for a given college based on sex, which should be irrelevant. In this case, we would be justified in saying that the SAT was biased. Likewise, an intelligence test may be used to select children for participation (or nonparticipation) in a program for gifted students. If the intelligence test is a better predictor for White than for ethnic minority students, the test is biased.⁴

Psychometric researchers can evaluate this type of bias using multiple regression. In essence, what we are saying is that a biased test has different regression lines for the groups (males and females, Ethnic minority and non-minority); we can therefore conceive of bias in predictive validity as a problem of the possible interaction of a categorical (e.g., male versus female) and a continuous variable (e.g., the SAT) in their effect on some outcome (e.g., college GPA). Let's flesh out this example a little more fully, after which we will turn to a research example of predictive bias.

Predictive Bias

Assume you are in charge of admissions for a selective college and that one type of information you use to select students is their scores on the SAT. Figure 7.11 shows your likely, albeit exaggerated, expectation of the relation between the SAT and college GPA. Based perhaps on data collected during a period of open admissions, you know that students with low SAT scores generally perform poorly in your college, whereas those with high SAT scores generally go on to perform well, with high grades in most courses. In addition, the graph shows that this ability of the SAT to predict future GPA is equal for males and females. If you decide to use a cutoff, for example, of 1000, you will be equally fair (or unfair) to both males and females. The females you accept with a SAT of, for example, 1200, will likely perform at the same level as your college as males with a score of 1200.

Figure 7.12, however, shows a different possibility. In this example, the regression lines are parallel, but the line for females is higher than that for males. What this means is that if you, as the admissions officer, use the common regression line (not taking into account sex) you will in essence treat males and females differently. Using a cutoff of 1000 for admissions,

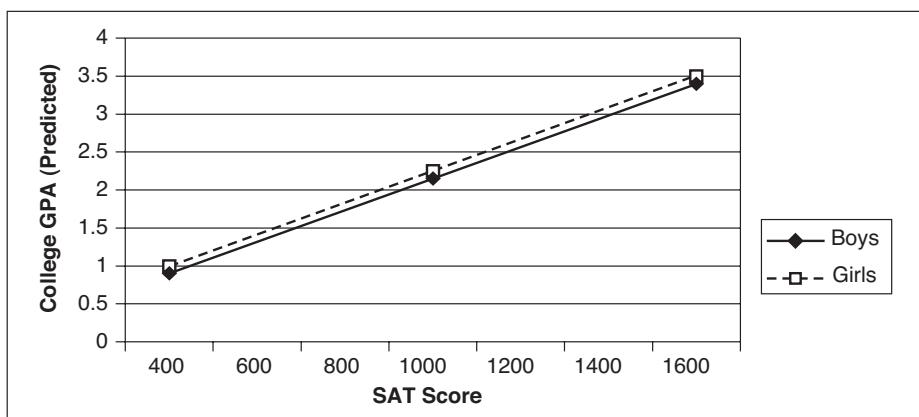


Figure 7.11 Possible regression lines: the SAT predicts College GPA equally well for boys and girls.

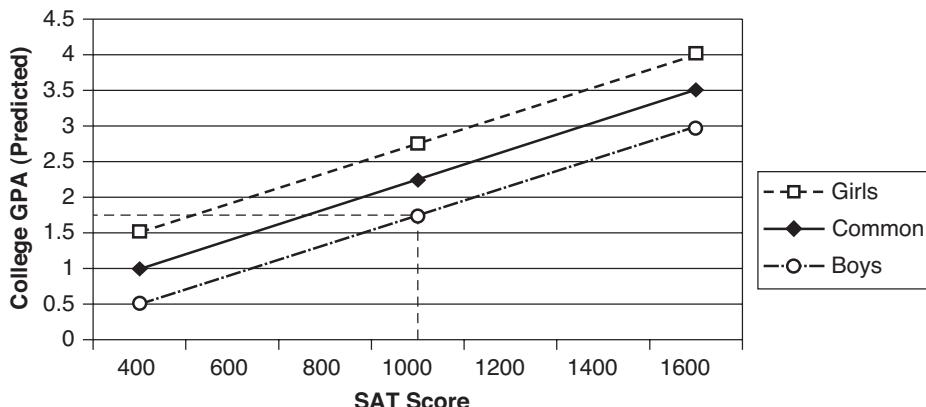


Figure 7.12 Possible regression lines: intercept bias in the use of the SAT to predict College GPA.

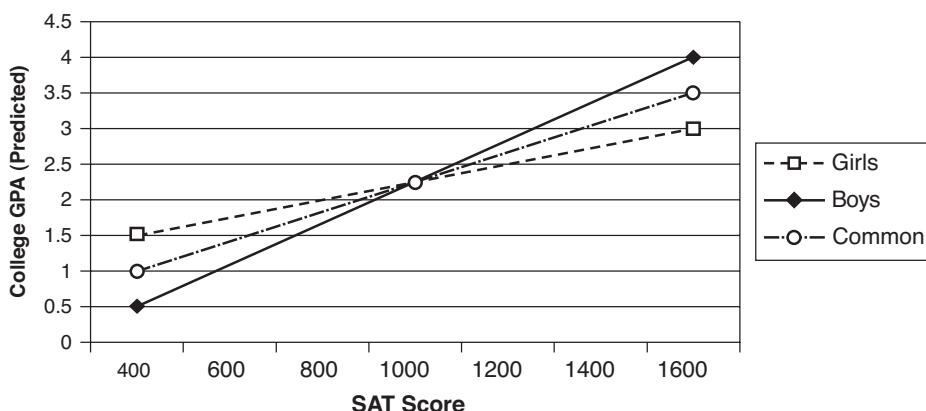


Figure 7.13 Possible regression lines: slope bias in the SAT in its prediction of College GPA for boys and girls.

you will end up selecting a group of males who will likely perform at a lower level than the females, while rejecting females who will likely perform as well or better than the males. Follow the dotted line vertically from the point on the X-axis representing a SAT score of 1000 up to the regression line for males and then horizontally across to the Y-axis. As you can see, a cutoff of 1000 on the SAT means you will be admitting males for whom their predicted college GPA is about 1.75. Yet, in this made-up example, you could admit females scoring around 500 on the SAT who will likely achieve at that same level in college (a GPA of 1.75). If you use the common regression line (instead of separate regression lines) to make admissions decisions, you have discriminated against females who scored above 500 but below 1000. If this is the case, the SAT will be biased when used for such purposes. This type of bias is termed *intercept bias* because the intercepts for the two groups are substantially different.

Figure 7.13 shows yet another possibility, in which the slopes for the regression line for males and females are different. As shown, the SAT has a steeper slope in predicting college GPA for males than for females. This example illustrates *slope bias*. In this example, the use of the common regression line will be biased against either males or females, depending on where we drew the SAT cutoff for admission. At a value of 800, our admissions will be biased against females, because we would not admit some qualified females (those expected to perform as well as some of the males admitted). If the cutoff were 1200, however, the use of the common regression line (instead of separate regression lines) will be biased against males, because some of the males rejected would likely perform as well or better than some of the females accepted. In some sense, slope bias is more problematic than intercept bias.

With intercept bias we can have faith that our selection is fair if we use separate regression lines for the two groups. With slope bias, however, even if we use separate regression lines, our prediction is often simply better for one group than another.

Research Example: Investigating Test Bias

One common duty of school psychologists is to assess children who are having learning or behavioral problems in school, with the assessment results being used, along with other information, to develop interventions to ameliorate these problems. One possible outcome of such assessment is placement in a special education program. Curriculum-based measurement (CBM), is a method of assessment in which a student's curriculum materials are used in assessing the student. For a reading CBM, for example, the psychologist might have a student read passages from his or her reading textbook and count the number of words read correctly in 2 minutes. One advantage of CBM is that the measures are short and can be repeated frequently, even several times a week. CBMs, therefore, are especially useful for determining whether an academic intervention is working.

Although there is ample evidence that CBMs can be reliable and valid (VanDerHeyden, Witt, Naquin, & Noell, 2001), there is little research addressing potential bias in CBMs. Kranzler, Miller, and Jordan (1999) examined a set of reading CBMs for potential racial–ethnic and sex bias in predictive validity. Their research included children in grades 2 through 5, and used reading CBMs to predict Reading Comprehension scores on the California Achievement Test (CAT). Their results suggested possible intercept bias (for race–ethnicity) at grade 4, and both intercept (for sex and race–ethnicity) and slope bias (for sex) at grade 5.

The data set “Kranzler et al simulated .sav” or “Kranzler.txt” includes data designed to simulate those reported in Kranzler and colleagues (1999) for boys and girls in grade 5. We will use these simulated data to go through the steps needed to test for predictive bias. Figure 7.14 shows the summary statistics for the total sample and for boys and girls in the sample.

Report			
Girl Sex, girls=1		CBM CBM Reading	CAT California Achievement Test, Reading Comprehension
.00 Boys	Mean	105.8800	674.4656
	N	50	50
	Std. Deviation	59.44871	35.74755
	Minimum	2.00	610.31
	Maximum	224.00	743.40
1.00 Girls	Mean	123.0000	660.2107
	N	50	50
	Std. Deviation	54.01889	44.02849
	Minimum	19.00	565.75
	Maximum	244.00	773.53
Total	Mean	114.4400	667.3382
	N	100	100
	Std. Deviation	57.16224	40.53723
	Minimum	2.00	565.75
	Maximum	244.00	773.53

Figure 7.14 Descriptive data for the Kranzler et al. (1999) simulated data.

The multiple regression to test for bias in predictive validity is similar to the more generic test for interactions between categorical and continuous variables. In the first step, CAT scores were regressed on the sex variable, labelled Girl (and coded 1 for girls and 0 for boys) and the predictor variable, centered Reading CBM scores. In the second step, a Sex by CBM (centered) cross product was added to the regression equation to test for a possible interaction (i.e., slope bias) between Girl and CBM. The basic results of the multiple regression are shown in Figure 7.15.

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.475 ^a	.226	.226	14.163	2	97	.000
2	.556 ^b	.309	.083	11.484	1	96	.001

a. Predictors: (Constant), Girl Sex, girls=1, cbm_cen

b. Predictors: (Constant), Girl Sex, girls=1, cbm_cen, sex_cbm

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.	95.0% Confidence Interval for B	
							Lower Bound	Upper Bound
1	Regression	36770.411	2	18385.206	14.163	.000 ^b	667.005	687.347
	Residual	125912.997	97	1298.072			.189	.444
	Total	162683.408	99				-34.141	-5.209
2	Regression	50223.551	3	16741.184	14.291	.000 ^c	665.862	685.280
	Residual	112459.857	96	1171.457			-.034	.292
	Total	162683.408	99				-33.760	-6.268

a. Dependent Variable: CAT California Achievement Test, Reading Comprehension

b. Predictors: (Constant), Girl Sex, girls=1, cbm_cen

c. Predictors: (Constant), Girl Sex, girls=1, cbm_cen, sex_cbm

Coefficients^a

Model		Unstandardized Coefficients		Beta	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	677.176	5.125	.446	132.140	.000	667.005	687.347
	cbm_cen	.317	.064		4.941	.000	.189	.444
	Girl Sex, girls=1	-19.675	7.289		-2.699	.008	-34.141	-5.209
2	(Constant)	675.571	4.891	.182	138.117	.000	665.862	685.280
	cbm_cen	.129	.082		1.570	.120	-.034	.292
	Girl Sex, girls=1	-20.014	6.925		-2.890	.005	-33.760	-6.268
	sex_cbm	.414	.122		.3.389	.001	.172	.657

a. Dependent Variable: CAT California Achievement Test, Reading Comprehension

Figure 7.15 Regression results for the simulated Kranzler et al. (1999) predictive bias study.

The regression of CAT Reading Comprehension on Reading CBM, Sex, and the interaction term was statistically significant ($R^2 = .309$, $F [3, 96] = 14.291$, $p < .001$). Furthermore, the addition of the Sex by CBM cross product led to a statistically significant increase in explained variance ($\Delta R^2 = .083$, $F [1, 96] = 11.484$, $p = .001$), meaning that the interaction between Girls and CBM scores was statistically significant. This statistically significant interaction, in turn, suggests that Reading CBMs (in these simulated data) may indeed show

sex-related slope bias for 5th-graders when predicting Reading Comprehension. (Note that the same story is told by the statistically significant b for the cross product in the second half of the table of coefficients: $b = .414$, $t [96] = 3.389$, $p < .001$.)

For the next step, given the statistically significant interaction, I graphed the interaction to understand it more completely. Additional follow-up is conducted to determine the separate regression lines for both groups. Although we know that the slopes and the b 's are different for the two groups, it might be that CBMs are significant predictors for both groups but are simply better for one group compared to the other. The graph is shown in Figure 7.16, and it suggests that Reading CBMs are strongly related to Reading Comprehension for 5th-grade girls but are not as good predictors for 5th-grade boys. This interpretation of the graph is confirmed by the results of the separate regressions of Reading Comprehension on Reading CBMs for girls and boys, partial results of which are shown in Figure 7.17. The results thus

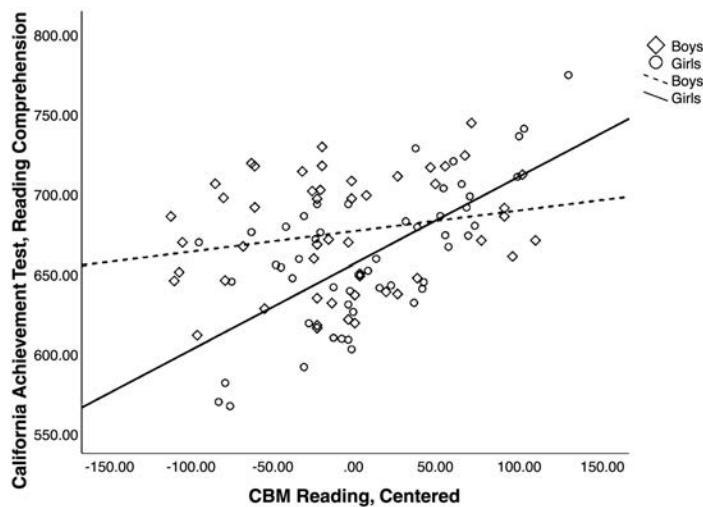


Figure 7.16 Plot of the regression lines for boys and girls illustrating slope bias in CBMs.

Model	Coefficients ^{a,b}						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	655.557	4.747		138.097	.000	646.013	665.102
cbm_cen	.544	.088	.667	6.202	.000	.367	.720

a. Dependent Variable: CAT California Achievement Test, Reading Comprehension

b. Selecting only cases for which Girl Sex, girls=1 = 1.00 Girls

Model	Coefficients ^{a,b}						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	675.571	5.041		134.012	.000	665.435	685.707
cbm_cen	.129	.085	.215	1.524	.134	-.041	.300

a. Dependent Variable: CAT California Achievement Test, Reading Comprehension

b. Selecting only cases for which Girl Sex, girls=1 = .00 Boys

Figure 7.17 Separate regressions of the California Achievement Test on CBMs for boys and girls.

suggest that Reading CBMs are excellent predictors of Reading Comprehension for girls ($r = .667$, $r^2 = .445$), but poor predictors for boys ($R^2 = .046$). In other words, Reading CBMs appear valid for girls, but not for boys at this age (remember these are simulated data).

Return to Figure 7.15 to review what all the coefficients in the table of coefficients mean. For the bottom portion of the table (the portion that includes the interaction term):

1. The constant, or intercept, represents the value, on the predicted dependent variable, for the group coded 0, for a value of zero on the continuous (centered) independent variable. On the graph, this represents the predicted CAT score (677.176) for boys who have a score of zero on the centered CBM score. The centered CBM score, in turn, represents the overall mean of the original CBM variable.
2. The unstandardized coefficient for Girl (sex) represents the *change* in intercept for the group scored 1 on the dummy variable. Thus, the predicted CAT score for girls scoring at the mean on the CBM variable is 20.014 points lower than for boys.
3. The coefficient for the continuous variable represents the slope of the regression line for those with a score of 0 on the dummy variable. Thus, the slope of the regression of CAT on centered CBM for boys is .129.
4. The coefficient for the interaction term is the change for the regression line for the group scored 1 on the dummy variable. Thus, a separate regression line for girls will have a slope of .543. (.129 + .414).

Thus from the lower half of the table of coefficients it is possible to generate the separate regression equations for girls and for boys:

$$\text{Boys: } \text{CAT}_{\text{predicted}} = 675.571 + .129\text{CBM}, \text{ and}$$

$$\text{Girls: } \text{CAT}_{\text{predicted}} = 655.557 + .543\text{CBM}$$

Compare these values with the values shown for the intercepts and slopes using separate regressions (Figure 7.17); they match within errors of rounding.

It is tempting to look at the results for the separate regressions in Figure 7.17 and conclude that CBM is a statistically significant predictor for girls but that it is not statistically significant for boys. Unfortunately, these are a little off. Note the values shown in the Boys regression in Figure 7.17 for the CBM centered variable ($b = .129$, $SE_b = .085$, $t = 1.54$, and $p = .134$). Now note the values associated with CBM centered in Figure 7.15, and recall that this row in the output represents the value for the group coded zero in the analysis of cross products, that is, for Boys. Here, the values are $b = .129$, $SE_b = .082$, $t = 1.57$, and $p = .120$. The b s are the same, but the standard error and the resulting t and probability values differ. Quite simply, the standard error value shown in Figure 7.17 is not correct, and the resulting t and p values are also incorrect. They do not differ by much, but by enough so that we could come to an incorrect conclusion about the statistical significance of our predictor for boys (and also for girls). Thus if we use these to determine whether CBM scores are statistically significant for each group separately then we could be misled.

Fortunately, the standard error shown in Figure 7.15 is the correct value. The value shown in Figure 7.17 is incorrect because the degrees of freedom are incorrect. When we conducted separate regressions we told SPSS only to analyze the data for girls, and then only analyze the data for boys. Thus for each follow-up regression only half the data were used, leading to misleading standard errors. This does not really matter for the boys' regression, because we have the correct SE shown in Figure 7.15. But the value for the girls' regression shown in

Figure 7.17 is also incorrect, and our original regression did not provide the correct standard error for girls.

This is a longish explanation of a problem in using the separate regression approach to determine whether the continuous variable predictor is statistically significant for all values of the categorical predictor. If that information is desired as a part of the follow-up, a different approach will be needed. You could calculate this by hand, but those calculations are beyond the scope of this text. Perhaps the easiest way to get this information is to re-do the regression shown in Figure 7.15 with the dummy variable re-ordered so that girls are now coded zero and boys coded 1 (making it a “Boy” variable), and the cross product re-created using this new dummy variable (cf. Aiken & West, 1991). The line for the CBM centered variable will then show the b and the correct standard error, t , and p for girls. For a more elegant approach, see the supplemental material for Chapter 7 on tzkeith.com. Finally, as noted earlier, the PROCESS macro discussed in more detail in Chapter 9 will also produce the correct standard error, error, t , and p for both boys and girls (Hayes, 2018).

If we had found no slope bias in this example (a statistically nonsignificant interaction), we would have focused on the first step of the regression equation, without the interaction term, to determine whether there was a difference in intercepts for the two groups (the top half of the table of coefficients). This would have been evidence for intercept bias.

Suppose our bias research focused on more than two groups; we would then have more than one dummy variable. We can determine whether intercept bias exists in this case by adding the dummy variables in a block, in a combination simultaneous and sequential regression. For example, in the first block we enter the continuous variable; in the second block, the dummy variables representing the categorical variables; and in the third block, the set of interaction terms.

Predictive Bias: Steps

Let me summarize the steps for investigating predictive bias using multiple regression (summarized, with modifications, from Pedhazur, 1997, chap. 14):

1. Determine whether the variance accounted for by the regression including all three terms (the categorical variable, the continuous variable, and the interaction) is statistically significant and meaningful. If not, it makes little sense to proceed. If R^2 is meaningful, go to step 2.
2. Determine whether the interaction is statistically significant. The most general method for doing so is to conduct a simultaneous regression using the categorical and continuous variable and then sequentially add the cross-product (interaction) term(s). If ΔR^2 for the cross product is statistically significant, then the interaction between the categorical and continuous variable is statistically significant; in the context of predictive bias, this suggests the presence of slope bias. If the interaction is statistically significant, go to step 3. If the interaction is not statistically significant (suggesting a lack of slope bias), go to step 4.
3. Graph the interaction and conduct follow-up regressions of the outcome variable on the continuous variable. These steps will help you determine the nature of the interaction and the slope bias. Interpret the unstandardized coefficients. Stop. (And again, see [www.tzkeith.com](http://tzkeith.com) for more depth.)

4. Determine whether the continuous variable is statistically significant across groups (without the cross-product term in the equation). You can do this in two ways. You can regress the outcome on the categorical variable and then add the continuous variable to the regression equation, focusing on ΔR^2 and the associated test of statistical significance. Alternatively, you can focus on the statistical significance of the b associated with the continuous variable with the categorical variable in the equation (in the present example, the b associated with CBM in the top of the table of coefficients in Figure 7.15). If the continuous variable is statistically significant, meaning that the test is a valid predictor of the outcome across groups, go to step 5. If not, meaning a lack of predictive validity across groups, go to step 6.
5. Determine whether the intercepts are different for the groups with the continuous variable in the equation. Most generally, you could regress the outcome on the continuous variable, sequentially adding the categorical variable(s) and focus on the ΔR^2 and its statistical significance. In the present example, with only two categories and one dummy variable, we could garner the same information by focusing on the statistical significance of the b associated with the categorical variable (Sex) in the top half of the table of coefficients shown in Figure 7.15. A difference in intercepts suggests intercept bias, whereas no difference suggests a lack of intercept bias. If there is no intercept or slope bias, then a single regression equation functions equally well for all groups; go to step 6.
6. Determine whether the groups differ without the continuous variable in the equation. Regress the outcome on the categorical variable alone and check for statistical significance. If the categorical variable is statistically significant, this means that the groups have different means, which does not constitute bias.

Before proceeding to the next section, I should note that findings of slope bias (like findings of more general interactions of categorical and continuous variables) are relatively uncommon. I also know of no other evidence suggesting slope bias for CBMs, although, as already noted, little research has been conducted concerning bias in curriculum-based assessment; I know of one other study suggesting no such bias (Hintze, Callahan, Matthews, Williams, & Tobin, 2002). The current example was chosen because it does illustrate this special type of interaction and because it was a well-executed, well-written study.

Although I have discussed the narrow issue of bias in predictive validity here, it is also worth noting that this methodology extends to other types of bias beyond *test* bias. Suppose, for example, you were interested in the existence and nature of pay disparities between male and female college professors. Whether such disparities represent bias is also addressable through multiple regression using continuous and categorical variables. You could, for example, regress Salaries on variables representing experience, productivity, and sex, as well as cross-product terms (Sex by Experience, Sex by Productivity). Differences in slopes and differences in intercepts across groups would suggest inequities in salaries (cf. Birnbaum, 1979). Finally, it is worth noting that such testing for predictive bias provides an incomplete answer to the question of whether a test (or other measure) is a biased predictor of some outcome. A more complete test would involve first establishing measurement invariance (see Chapter 20) and then testing for invariance in prediction within that model. It is possible to have real bias without it showing up in tests such as conducted here, or the reverse (Borsboom, 2006; Millsap, 2007; Wicherts & Millsap, 2009).

Aptitude–Treatment Interactions

Psychologists and educators often develop interventions and treatments with the belief that the effectiveness of these interventions depends, in part, on the characteristics of those receiving the interventions. Children may be placed in different reading groups (high, medium, low) based on their prior reading achievement with the belief that one teaching method is more effective with one group, whereas another method is more effective with another. A psychologist may use one type of therapy with clients who are depressed, but believe a different approach is more effective for those without depression. These are examples of potential Aptitude–Treatment Interactions (ATIs), also known as Attribute–Treatment Interactions (ATIs) or Trait–Treatment Interactions (TTIs). Whatever the terminology, ATIs are interactions between some characteristic of the individual with a treatment or intervention so that the treatments have differential effectiveness depending on the characteristics, attributes, traits, or aptitudes of the person. These attributes can generally be measured on a continuous scale (e.g., reading skill, depression), whereas the treatments are often categorical variables (e.g., two different reading approaches, two types of therapy).

ATIs are, then, generally an interaction between a categorical and continuous variable. Thus, they may be tested using multiple regression in the same way we test for potential predictive bias, by testing the statistical significance of a cross-product term. An example will illustrate.

Verbal Skills and Memory Strategies

Do children with lower verbal reasoning skills profit more from learning different memorization methods than do children with good verbal reasoning skills? For example, will children with lower verbal reasoning skills be more accurate in memorization if they use a visual mapping strategy as a memory aid (as opposed to a verbal rehearsal strategy)? In contrast, will children with higher verbal reasoning skills show greater accuracy using a verbal rehearsal memorization strategy? To answer these questions, you could develop an experiment to test for this possible attribute–treatment interaction. You could, for example, assess children's verbal reasoning skills, rank ordering the children based on their scores on the verbal reasoning measure. Take the first pair of students (the one with highest score and the one with the second highest score) and assign one (at random) to the verbal rehearsal group and one to the visual matching group. Continue with each pair of children, down through the lowest scoring child and the second lowest scoring child, assigned at random to one group or the other. Children in the verbal rehearsal group are taught to memorize things (e.g., words, lists, colors) using a memory strategy based on verbal rehearsal, while those in the visual mapping group are taught a memory strategy in which they memorize by visualizing the placement of the objects to be memorized in stops in a map.

"ATI Data.sav" is a data set designed to simulate the possible results of such an experiment (the data are loosely based on Brady & Richman, 1994). If our speculation is correct, the verbal rehearsal strategy should be effective for children with high verbal skills, and the visual mapping strategy should be more effective for children with lower verbal skills. The data are plotted in Figure 7.18; it certainly appears that our speculation is correct: there is an interaction between the attribute (Verbal Reasoning) and the treatment (type of Memory Strategy) in their effect on Visual Memory skills. Let's test the statistical significance of the interaction (I displayed the graph prior to the testing of the interaction to give you a sense of the data).

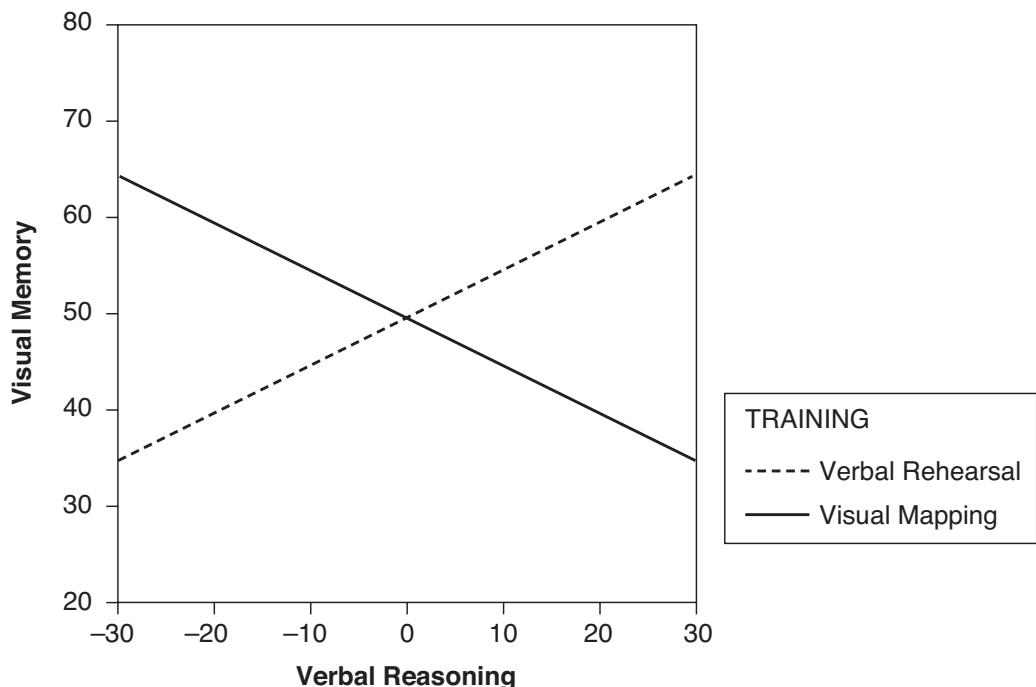


Figure 7.18 Plot of regression lines illustrating an aptitude treatment interaction. Verbal Reasoning is the aptitude, and type of Memory Strategy taught is the treatment.

The process of testing for an ATI is the same as testing for predictive bias. Visual Memory (measured by having the child recall an increasing number of color chips, expressed as a T score) was regressed on Memory Strategy (verbal rehearsal, coded 0, or visual mapping, coded 1) and Verbal Reasoning (T score, centered) in a simultaneous regression. In a second step, the cross product (Memory Strategy multiplied by Verbal Reasoning) was added sequentially to the regression to test for the statistical significance of the interaction. Relevant portions of the printout are shown in Figure 7.19. We can use the same basic steps for evaluating the results that we used for our analyses of bias, although the interpretation is slightly different.

Steps for Testing for ATIs

1. Is the overall regression meaningful? R^2 is indeed meaningful and statistically significant ($R^2 = .280, F[3, 96] = 12.454, p < .001$). We go to step 2.
2. Is the interaction term statistically significant? The addition of the cross-product term to the regression equation resulted in a statistically significant increase to $\Delta R^2 (.280, F[1, 96] = 37.332, p < .001)$, suggesting the statistical significance of the interaction. In the context of ATIs, this suggests that the Attribute-Treatment Interaction is statistically significant. In the current example, the finding of an interaction suggests that the two memory strategies are indeed differentially effective depending on the verbal skills of the children (go to step 3).

3. Follow up the statistically significant interaction. The interaction is already graphed in Figure 7.19. Follow-up regressions (not shown) showed that Visual Memory regressed on Verbal Reasoning was statistically significant for both groups (both treatments). For those trained in verbal rehearsal as a memory strategy, the slope (b) of the regression line shown in the figure was .514 ($b = .514$, $t [48] = 4.199$, $p < .001$). For those in the visual mapping group, the slope was negative ($b = -.544$, $t [48] = 4.442$, $p < .001$). What do these findings mean (assuming they represent real, rather than simulated, data)? One way to interpret the findings is that for children taught to use verbal rehearsal to memorize lists verbal reasoning skills are useful, but for those taught to use visual maps as a memory aid verbal reasoning ability is a hindrance to effective memorization. I don't find this interpretation particularly helpful. Instead, a more useful interpretation is that visual mapping strategies are more effective for children with difficulties in verbal reasoning, whereas verbal rehearsal memory strategies appear more useful for students with good verbal reasoning skills. For students with average-level verbal reasoning skills, the approaches appear equally effective. Stop.
4. Had the interaction not been statistically significant, we would have gone to step 4 in the previous (Steps: predictive bias) list (the statistical significance of the continuous variable), followed by step 5 (the statistical significance of the categorical variable). In the context of ATIs, the statistical significance of the continuous variable and the categorical variable are analogous to tests of the main effects in ANOVA.

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.015 ^a	.000	.000	.011	2	97	.989
2	.529 ^b	.280	.280	37.332	1	96	.000

a. Predictors: (Constant), TRAINING Type of Memory Strategy, VERBAL Verbal Reasoning

b. Predictors: (Constant), TRAINING Type of Memory Strategy, VERBAL Verbal Reasoning, V_TRAIN Training by Verbal crossproduct

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	50.000	1.421	-.015	35.178	.000	47.179	52.821
	VERBAL Verbal Reasoning	-1.49E-02	.102		-.147	.884	-.216	.187
	TRAINING Type of Memory Strategy	.000	2.010		.000	1.000	-3.989	3.989
2	(Constant)	50.000	1.212	.514	41.244	.000	47.594	52.406
	VERBAL Verbal Reasoning	.514	.122		.514	4.199	.000	.271
	TRAINING Type of Memory Strategy	.000	1.714		.000	1.000	-3.403	3.403
	V_TRAIN Training by Verbal crossproduct	-1.058	.173		-.748	-6.110	.000	-.714

a. Dependent Variable: VIS_MEM Visual Memory

Figure 7.19 Regression of Visual Memory on an aptitude (Verbal Reasoning), a treatment (Memory Strategy), and their interaction.

Although multiple regression is ideal for the analysis of ATIs, its use is much too uncommon. Faced with the example above, many researchers try to fit the data into a classic ANOVA design by categorizing the continuous variable. That is, the researcher not familiar with this better analysis of ATIs might place anyone scoring below the median on the verbal reasoning scale in a “low verbal” group and anyone above the median in a “high verbal” group, analyzing the data with a 2 by 2 Analysis of Variance. This approach, at minimum, ignores and discards variation in the continuous variable, thus reducing the power of the statistical analysis. Unfortunately, in my experience this improper approach is more common than is the more proper, more powerful MR approach outlined here. Be warned: you now know better.

The search for ATIs is most common in psychology and education; indeed, much of special education is predicated on the assumption that ATIs are important. Children with learning problems are sometimes placed in different classes (e.g., classes for children with mild intellectual disabilities versus classes for children with learning disabilities) based in part on the assumption that different teaching methods should be used with the two groups. But these designs are applicable to other research areas as well. Are two different types of psychotherapy (treatment) differentially effective for depressed versus nondepressed (attribute) clients? Is one management style (treatment) more effective with less productive (attribute) employees, with a different style being more effective with more productive employees? ATI designs have wide applicability. For more information, Cronbach and Snow (1977) is the classic source on ATIs and their analysis.

ANCOVA

Suppose you are interested in the effectiveness of Internet-based instruction in research methodology. Is, for example, an Internet-based research course as effective as traditional face to face instruction? One way of studying this problem would be via a classic randomized pretest–posttest control group design. That is, you might assign, at random, students entering a course in research methodology to an online course versus a traditional classroom course. Because you believe the effectiveness of the coursework may depend, in part, on participants’ prior knowledge, you give participants a pretest on research methodology knowledge. After course completion, participants are given another measure of knowledge of research methodology. One straightforward method of analysis of the results of this experiment would be through analysis of covariance (ANCOVA), where the pretest serves as the covariate, and assignment to the Internet versus regular coursework is the independent variable of interest. ANCOVA is used to examine the effects of course type on research knowledge, controlling for participants’ prior knowledge of research methodology. ANCOVA serves to reduce error variance by controlling for participants’ individual differences and thus provides a more sensitive statistical test than does a simple ANOVA.

I hope that it is obvious that ANCOVA can also be conceived as a multiple regression analysis with a continuous and a categorical variable. MR subsumes ANCOVA; if you analyze these same data using a simultaneous multiple regression, your results will be the same as those from the ANCOVA. There is, however, an advantage to analysis via MR. One assumption underlying analysis of covariance is that the regression lines of the dependent variable on the covariate are parallel for the different groups (e.g., Internet versus traditional course). In other words, ANCOVA assumes but does not generally test for, the nonexistence of an interaction between the independent (or categorical) variable and the covariate (continuous variable). It might well be that Internet-based instruction is more effective for students

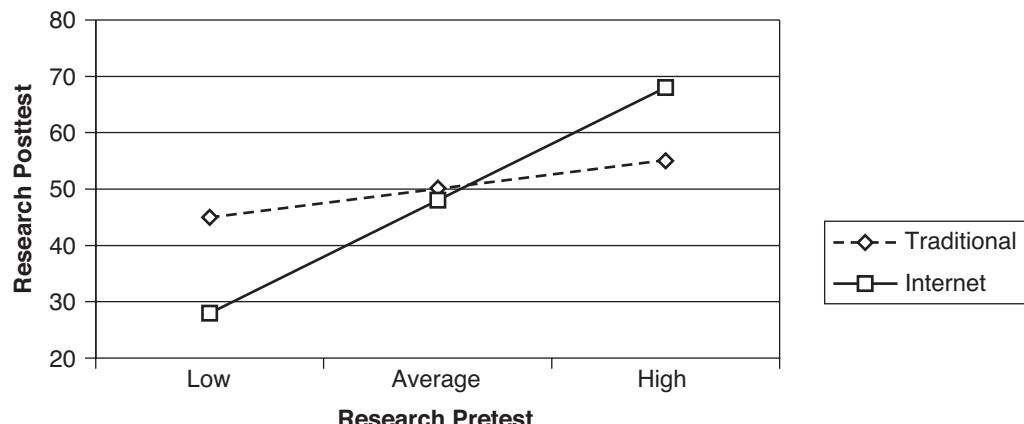


Figure 7.20 A potential interaction in a pretest–posttest control group design.

with strong prior knowledge but less effective for students whose prior research knowledge is weak. If this is the case, a graph of your findings might look something like Figure 7.20, which is simply one more illustration of an interaction between a categorical and continuous variable. Obviously, you can test this assumption using multiple regression, using the same method explained throughout this chapter, whereas most software packages ignore the interaction in ANCOVA.⁵

One way of thinking about ATIs and ANCOVAs is this: If the interaction is not statistically significant in an ATI design, you can think of it as being a simple ANCOVA analysis. If in a pretest–posttest design the pretest (covariate) interacts with the treatment, you can consider it an ATI design and analyze it accordingly.

CAVEATS AND ADDITIONAL INFORMATION

“Effects” of Categorical Subject Variables

In this chapter and elsewhere, I have discussed the effects of variables such as Sex and Ethnic background on outcomes such as Self-Esteem. Yet I hope it is obvious that these types of variables and others (e.g., rural and urban, region of the country, and religious affiliation) are very broad categories and can mean many different things. If we say that Sex affects Self-Esteem, what does this mean? That the biological differences between boys and girls result in different levels of self-esteem? Or that the way boys and girls are socialized results in differences in self-esteem? Or some other of the myriad of differences that are associated with being a boy or a girl? Would the results differ if we asked about subjects’ self-identified gender as opposed to biological sex? We just don’t know (although when we discuss testing for indirect effects in structural equation modeling, you will have a tool that you can use to investigate some of the possibilities). All it means, really, is that *something* about being a boy versus being a girl results in differences in self-esteem. Similarly, if we say that sex and achievement interacted in their effects on self-esteem or that achievement had different effects on self-esteem for boys versus girls, we will be left to speculate about the many possible reasons that such an interaction might happen and what it might mean. “Big” categorical variables like Sex and Ethnic background carry a lot of baggage, and sometimes, when we discover an interaction

between them and some other variable, we are confronted with many new questions about meaning.

Some methodologists see this as a major problem. I don't. I think it's okay to say that sex affects self-esteem, as long as you know that this statement means "there is something—we don't know what—about being a boy versus being a girl that results in differences in self-esteem." Likewise, I think it is fine to say that Ethnic background and Achievement interact in their effects on Self-Esteem, as long as you know that what this means is for some reason—unknown at this point—achievement has a different effect on self-esteem for adolescents of one group versus another. Understand the meaning behind such statements and then maybe the next step can be hypothesizing and testing why such effects come about.

Interactions and Cross Products

In an earlier footnote, I discussed the distinction between cross-product terms and interactions. Strictly speaking, the *partialed* cross product (controlling for the two variables used in the cross product) is an interaction term. Of course, these variables are controlled when all are entered into a multiple regression equation, either simultaneously or sequentially (the cross product entered last), so many researchers use the terms interchangeably.

Further Probing and Figural Display of Statistically Significant Interactions

Suppose, as in several of the examples in this chapter, you find a statistically significant interaction between a categorical variable and a continuous variable. How can you explore that interaction in more depth? Here I have suggested graphing the interaction and then producing separate regression equations across the different categories of the categorical variable. Yet further exploration is possible. You might be interested in knowing whether the regression lines are statistically significantly different for a specific value of the continuous variable. In our ATI example, you might wonder for a student with a verbal score of 10 whether the two approaches are really different or not. You may also be interested in the regions of significance; in other words, the point at which the two lines become statistically significantly different.

These are worthwhile topics, but they are beyond the scope of this text. Some references given throughout this chapter provide additional detail for how to probe a significant interaction in more depth than is discussed here (Aiken & West, 1991; Cohen et al., 2003; Cronbach & Snow, 1977; Darlington & Hayes, 2017; Hayes, 2018; Pedhazur, 1997). Some of the procedures are relatively complex. If you are faced with an interaction that requires more complex probing, I recommend these sources.

It is relatively easy, however, to develop a less formal sense of answers to these kinds of questions, using the graphing features of common statistical programs. Figure 7.21, for example, shows another version of the graph of the ATI example originally shown in Figure 7.18. In this version, however, I asked for the 95% confidence interval around the two regression lines, which provides at least a general sense of where, at what points, the lines become significantly different from one another.

Throughout this text I have illustrated regressions in figural form using path models. How, you might wonder, should we illustrate the types of interactions covered in this chapter, that is, interactions between categorical and continuous variables in their effect on some

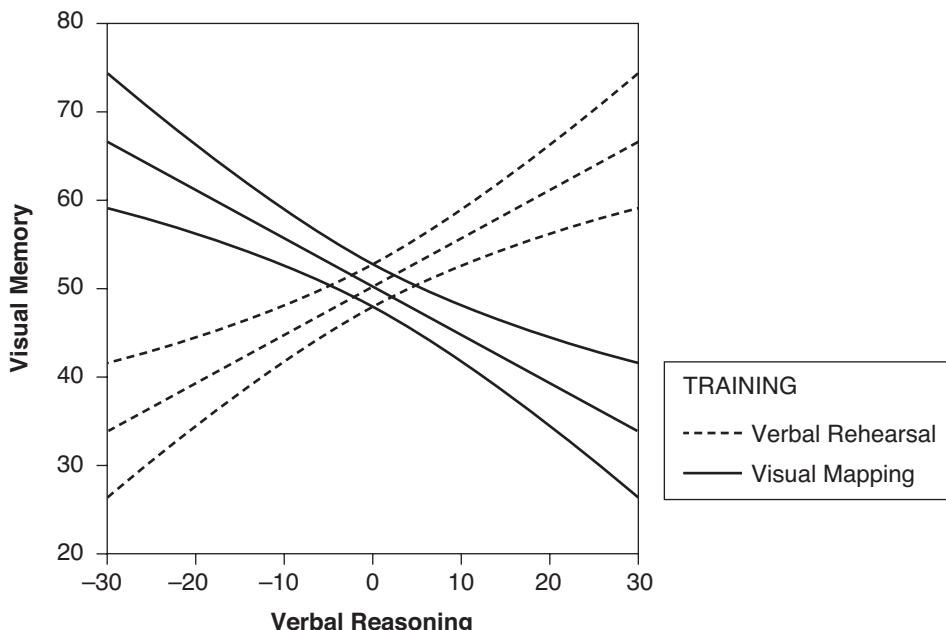


Figure 7.21 Regression lines for the ATI analysis, with 95% confidence intervals.

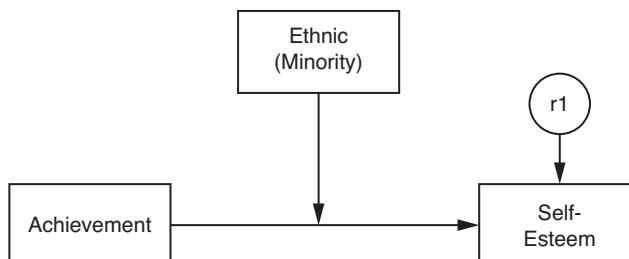


Figure 7.22 One way of illustrating an interaction between Ethnic Background and Achievement in their effect on Self-Esteem.

outcome? One common way of illustrating interactions, or moderation, is via a display like that shown in Figure 7.22. This model, a figural representation of the test of the interaction between Ethnic Background and Achievement on Self-Esteem (our first statistically significant interaction), has a path drawn from Ethnic background to the path from Achievement to Self-Esteem. This conceptual model suggests that Ethnic background influences (or moderates) the effect of Achievement on Self-Esteem. Such a model conveys the essence of an interaction (does Achievement influence Self-Esteem? it depends on one's Ethnic Background) but generally are not used to display results, that is, they generally don't have numbers attached.

Another way to display the findings in path form is shown in Figure 7.23. This model shows the coefficients from the final block of the regression (Figure 7.10). These are the unstandardized coefficients. A final method of displaying the results of a statistically

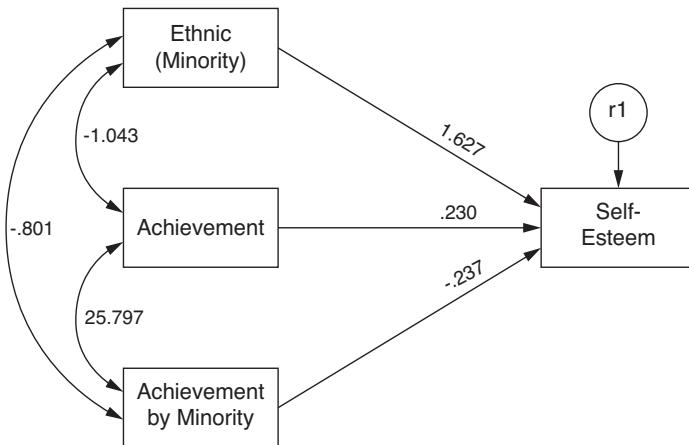
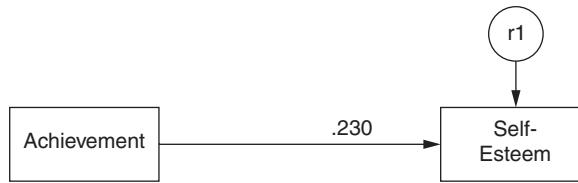


Figure 7.23 Path illustration of regression results for an interaction between a categorical and a continuous variable.



Effect of Achievement on Self-Esteem for Ethnic Minority Students



Effect of Achievement on Self-Esteem for White Students

Figure 7.24 Another method of displaying moderation (interaction) results in path format.

significant interaction in path form is shown in Figure 7.24. This method (standardized coefficients this time), suggests no effect for Achievement on Self-Esteem for minority students but substantial effects for white students. This method of display is analogous to the post-hoc follow-up we have been conducting following the finding of a statistically significant interaction.

The web site for this book (www.tzkeith.com) includes additional resources for the topic of testing for interactions for categorical and continuous variables, including an illustration of the use of effect coding for the bias example. Also illustrated is a method for testing the statistical significance of the separate regression equations (post-hoc probing) in a single regression.

I have noted several times in this chapter that we commonly refer to interactions, especially in nonexperimental research, as moderation. Figure 7.24 can be described as demonstrating the degree to which research class type moderates the effect of pretest scores on research knowledge. The topic of moderation will be discussed in more depth in Chapter 9: Mediation, Moderation, and Common Cause. Multigroup SEM, an important topic in Part 2 of this book, is a method of testing for moderation between categorical and continuous variables in structural equation modelling.

SUMMARY

In this chapter we focused on the analysis of categorical and continuous variables in the same multiple regression. Our first example examined the effect of Sex and Achievement on adolescents' Self-Esteem. As the example illustrated, analyses including both categorical and continuous variables are analytically and conceptually little different from those including only continuous variables. When the categorical variable is a single dummy variable, the b associated with it represents the difference, on the dependent variable, between the two groups, controlling for the other variables in the equation.

It is possible to test for interactions between variables by multiplying the two variables that may interact and entering this cross-product term in the regression equation along with the two original variables. It is desirable to center any continuous variables used to create such a cross product by subtracting the mean of that variable from each person's score on the variable. The ΔR^2 (if sequential regression is used) or the t associated with the cross product term (if simultaneous regression is used with a single cross product) is used to test the statistical significance of the interaction. The same procedure works to test the interaction between two categorical variables, two continuous variables, or a categorical variable and a continuous variable. This chapter illustrated several examples of interactions between categorical and continuous variables. We found no statistically significant interaction between Sex and Achievement in their effect on Self-Esteem but did find a statistically significant interaction between Ethnic background and Achievement on Self-Esteem. Graphs and separate regression lines across groups were used to probe the statistically significant interaction. It appears that Achievement affects Self-Esteem for White adolescents, but not adolescents from various Ethnic minority backgrounds. In answer to the question "Does Achievement affect Self-Esteem?" we would need to answer "It depends. . ." The phrase "it depends" is generally a clue that we are describing the presence of an interaction. Interactions are not common in nonexperimental research.

A few specific research questions are best conceived of as interactions between categorical and continuous variables. These include investigations of predictive bias and attribute or aptitude treatment interactions (ATIs). Examples were given of each, using simulated data designed to mimic previous research. Analysis of covariance (ANCOVA) can also be considered as a multiple regression analysis involving both continuous (the covariate) and categorical (the treatment or independent variable) variables. One potential advantage of using MR to analyze ANCOVAs is that it is simple to test for an interaction between the covariate and the treatment, whereas this is simply assumed for most ANCOVAs.

I noted that it is loose usage to discuss the "effects" of broad, existing categorical variables, such as Sex, on various outcomes, because of all the things that may be subsumed under the meaning of such categorical variables. My belief is that such usage is acceptable if you are clear as to the meaning. Likewise, a cross product is not strictly an interaction term, even though it is used to test for an interaction, but many people use these terms interchangeably. Finally, I discussed several additional sources for more detail on testing for interactions in MR, including subsequent chapters in Part 1 and Part 2 of this book.

EXERCISES

1. Conduct the first three examples used in this chapter that used the NELS data: the regression of Self-Esteem on Sex and Achievement, the same analysis with the addition of an Achievement by Sex cross product, and the regression of Self-Esteem on Ethnic background, Achievement, and an Ethnic by Achievement cross product. Make sure your results match those presented in the chapter.
2. Use the “Kranzler et al simulated .sav” (or “Kranzler et al simulated.xls”) data set found on the Web site (www.tzkeith.com). Center the CBM scores and create a Sex by CBM cross product using the centered variable. Conduct an analysis for predictive bias using the centered data. Do the results match those presented here? Try conducting the analysis using the uncentered data (and cross product based on uncentered data), as was done in Kranzler et al. (1999). Compare the correlations, coefficients, and graphs from the two analyses. Would your interpretation be the same? Compare the two printouts; see if you can develop a sense of why the differences occur. It is also worth focusing on the correlation matrices and then the standard errors of the regression coefficients. Reread the section discussing the advantages of centering.
3. Is the NELS math test biased against girls? Conduct an analysis of predictive bias using the base year test (ByTxMStd) and Sex, with 10th-grade Math GPA as the outcome (F1S39a). Make sure you convert Sex into a dummy variable and center the Math test score.
4. The file “ATI Data b.sav” (or the Excel or plain text versions of these data) includes another, perhaps more realistic, simulated data set for the attribute–treatment interaction problem illustrated in the chapter. Perform an ATI analysis and interpret the results.
5. The file “ancova exercise.sav” includes simulated data for the ANCOVA example presented in the chapter (see also the Excel or plain text versions of this file). This was a pretest–posttest two-group design in which 60 students registered for a course in research methodology were assigned, at random, to a traditional version of the class or an Internet version of the class. All students attended an orientation session in which they were given a pretest of their research knowledge. The posttest scores are students’ grades for the class. Analyze the results of the experiment using multiple regression analysis. Test for the presence of an interaction between the pretest and the treatment (type of class). Conduct any needed follow-up. Conduct an ANCOVA and compare the results of this analysis with those of the multiple regression.
6. Carter, Greenberg, and Walker (2017) reported a randomized control study investigating the effects of allowing versus prohibiting laptops and tablets in class on West Point students’ performance in an Economics class. The authors reported a lot of analyses and fascinating results. One interesting follow-up suggested that ACT scores (a college entrance exam) may moderate the effects of computer usage on final exam score. The file “Carter et al computers.sav” shows simulated data loosely designed to mirror this finding. Analyze the data using multiple regression and testing for an interaction between ACT and experimental treatment; also conduct needed follow-up analyses. What do your analyses show? Are laptops (and tablets) helpful or harmful? Do ACT scores predict Economics class performance? Are laptops equally helpful/harmful for well-prepared (based on ACT score) versus unprepared students? Focus on the final table of coefficients from your overall regression. Which line in the output shows the intercept for the laptop group? Which line in the output shows the difference in intercept for the no-laptop group?
7. Kristen Alexander and colleagues were interested in whether the impact of a traumatic event (child sexual abuse) predicts (explains) one’s subsequent memory of that event (Alexander et al., 2005). In that study the researchers were interested in whether the severity of posttraumatic stress disorder symptoms helped explain how accurate abuse victims were in recalling the details of their abuse 12–21 years later. The file “Alexander et al abuse.sav” includes data designed to simulate some of the important variables

in that study. The outcome variable Ncorrect is the number of details of the abuse recalled correctly. Sex was coded 0 for male victims, 1 for female; Support was the presence of maternal support for abuse disclosure (0=no, 1=yes); MTE was whether their sexual abuse was the most traumatic event they every experienced (0=no, 1=yes); NPTSD was the number of criteria for posttraumatic stress disorder currently met. The NPTSD scale ranged from 0 to 9 in the simulated data, and included criteria such as re-experiencing events and impairment in daily life. Use multiple regression to determine whether these variables are important in explaining memory accuracy (Ncorrect). Test for an interaction between MTE and NPTSD in their effect on Ncorrect. Conduct needed followup analyses (e.g., graphing and follow-up regressions). Explain your findings, in English; that is, what do these results mean?

8. Do fathers' aspirations influence children? Do fathers' aspirations have differential effects for boys versus girls? Using the NELS data and multiple regression, test whether father's educational aspirations for their children (ByS48a) have any effect on their eighth grade children's GPAs (ByGrads). Also test whether sex (Sex) moderates the effect of father's aspirations on Grades. You should also control for family background characteristics (BySES) in this analysis. Note: you should treat the aspiration variable as a continuous variable; no modifications are needed to it. Treat boys as the reference group. Conduct any needed pre-analyses and any needed follow-up analyses. What do your analyses show? Are fathers' aspirations important? Are they important for both boys and for girls? Are they more important for one sex versus the other? Focus on the final table of coefficients from your overall regression. Which line in the output shows the intercept for the boys? Which line in the output shows the difference in intercept for girls?

Notes

- 1 For example, in SPSS, you could create this variable from a z-score version of the original scale by: 1. Convert F1Cncpt2 into a z score, e.g., using a DESCRIPTIVES command, and 2. Creating a T-score version of this new composite using a compute statement: COMPUTE S_Esteem=((ZF1Cncpt2*10)+50). Note also that for Figure 7.1 I have used listwise deletion of missing data. This is accomplished using syntax, rather than point-and-click, in SPSS, with the addition of a /missing=listwise subcommand.
- 2 These cross products are often referred to as interaction terms. Strictly speaking, this multiplication of the two variables should be referred to as a cross-product term, rather than an interaction term. To create a pure interaction term, we need to remove the variance attributable to the categorical and continuous variables from the cross product (e.g., regress the cross product on the categorical and continuous variables and save the residuals as an interaction variable). The testing process is identical, however, and I will generally use the terms cross product and interaction interchangeably. Of course, the effects of these variables are also removed in the MR regression analysis when they are entered simultaneously or prior to the cross product.
- 3 If our primary purpose were really to investigate the interaction of ethnic/racial group and achievement on their effects on self-esteem, it would be better to convert the NELS Race variable into multiple categorical variables representing the multiple ethnic groups in the sample. For example, we could create three categorical (dummy or effect-coded) variables coded 1 for Asian, Hispanic, and Black students, respectively, with White being the contrast group (coded 0) for each. The primary purpose, however, is to illustrate a statistically significant interaction/moderation with a fairly simple example. See, however, tzkeith.com for additional analysis.
- 4 Note that predictive bias is only one of several types of potential bias. It is also referred to as the regression model of bias, or Cleary's definition of test bias, after the late T. Anne Cleary who explicated the nature of bias in prediction (Cleary, 1968). My purpose here is not the exhaustive discussion of test bias but to illustrate one instance of the wide applicability of testing interactions in regression. As noted at the end of this section, this type of analysis is by no means an exhaustive test of bias.
- 5 Depending on the software you use, it is possible to test for interactions of an independent variable with a covariate in an ANCOVA analysis (it is possible in SPSS, for example). Testing for such interactions is not very common in my experience, however.

8

Testing for Interactions and Curves With Continuous Variables

Interactions Between Continuous Variables	161
<i>Effect of TV Time on Achievement</i>	161
Curvilinear Regression	168
<i>Curvilinear Effects of Homework on GPA</i>	169
Summary	175
Exercises	175
<i>Note</i>	176

As noted in Chapter 7, it is possible to have interactions between two or more continuous variables in their effect on some outcome. This chapter will discuss such interactions, as well as regression in which there is a curve in the regression line. As we will see, such curves can be considered cases in which a variable interacts with itself in its effect on some outcome variable.

INTERACTIONS BETWEEN CONTINUOUS VARIABLES

Conceptually, there is little difference between testing an interaction between two continuous variables and testing an interaction between a categorical and continuous variable. Although the probing of a statistically significant interaction is slightly more complex when both variables are continuous, the basic steps are the same. With two continuous variables, both variables are centered, and then the centered variables are multiplied to create a cross-product term. The outcome variable is regressed on the two centered continuous variables (plus any other variables you wish to take into account) in a simultaneous regression. In a second, sequential step, the cross-product (interaction) term is entered into the regression. If the addition of the cross product leads to a statistically significant increase in R^2 , the interaction is statistically significant. An example will illustrate the process.

Effects of TV Time on Achievement

In Chapter 7 I mentioned research testing for an interaction between TV viewing and ability in their effects on achievement (Keith et al., 1986). The primary purpose of this study was to assess and compare the effects of parent involvement, homework, and TV

viewing on achievement. Previous research, however, had suggested that TV viewing may interact with Ability in their effects on Achievement (Williams, Haertel, Haertel, & Walberg, 1982). TV viewing appears to have a negative effect on achievement, but the extent of the effect may depend on the ability level of the student watching TV (remember, “it depends” often signals an interaction). Specifically, TV viewing may be especially detrimental for high-ability youth and less detrimental for low-ability youth (Williams et al.). Keith et al. (1986) sought to test the possible interaction between hours spent watching TV and intellectual Ability on adolescents’ academic Achievement. Another common way of phrasing our interest would be to ask whether ability *moderates* the effect of TV viewing on Achievement.

The Data: Centering and Cross Products

The data sets “tv ability interact2.sav,” “tv ability interact2.xls,” and “tv_abil.txt” include 500 cases of data designed to simulate the results of Keith et al. (1986). Variables in the data set include Ability (a composite of six verbal and non-verbal tests, each with a mean of 100 and a *SD* of 15), TV (average time per day, in hours, spent watching TV), and Achieve (an Achievement composite of Reading and Math, expressed as a *T* score). Also included is the background variable SES (in *z*-score format: a combination of parents’ educational attainment, parents’ occupational status, family income, and possessions in the home). From these data, I created centered versions of the two continuous independent variables of interest (TV_Cen and Abil_Cen) and the cross product of the centered TV and Ability variables (TV×Abil). The descriptive statistics for these variables are shown in Figure 8.1.

You may wonder why I did not create and use cross products reflecting interactions between TV viewing and SES, or between SES and Ability, and so on. Recall that in Chapter 7 I suggested that you should test only specific interactions, those designed to test specific hypotheses of interest in research, rather than wholesale testing of all possible interactions.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
SES Family Background	500	-2.84	3.12	.1099	1.01568
ABILITY Ability	500	75.00	130.00	100.4040	9.47504
TV TV Time, weekdays	500	0	8	4.01	1.754
ABIL_CEN Ability (centered)	500	-25.40	29.60	.0000	9.47504
TV_CEN TV Time, weekdays (centered)	500	-4.01	3.99	.0000	1.75445
TVXABIL TV by Ability crossproduct	500	-74.53	58.37	-2.9192	16.46830
ACHIEVE Achievement Test Score	500	29.00	75.00	50.0960	8.71290
Valid N (listwise)	500				

Figure 8.1 Descriptive statistics for the “tv ability interact.sav” data.

The current example exemplifies this approach by testing only the interaction of interest and suggested by previous research.

The Regression

Achievement was regressed on SES, Ability (centered), and TV viewing (centered) in a simultaneous regression, with the Ability–TV cross product in a second, sequential step. Some of the regression results are shown in Figure 8.2. As shown in the model summary, the initial three independent variables accounted for 51% of the variance in Achievement ($F[3, 496] = 172.274, p < .001$), and the addition of the Ability–TV cross product explained an additional 4.4% of the variance in Achievement, a statistically significant increase ($F[1, 495] = 49.143, p < .001$). The interaction between ability and time spent watching TV is statistically significant.

The table of coefficients, also shown in Figure 8.2, provides additional information about the effects of TV viewing on Achievement. As shown in the top portion of the table, prior to consideration of the cross product, each independent variable had a statistically significant effect on Achievement. Indeed, Ability had a large effect on Achievement (more able students achieve at a higher level), SES had a moderate effect (more advantaged students achieve at a higher level), and TV viewing had a small to moderate negative effect on Achievement. Other things being equal, the more time adolescents spend watching TV, the lower their academic achievement. The lower portion of the table again shows the statistical significance of the interaction.

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.714 ^a	.510	.510	172.274	3	496	.000
2	.745 ^b	.555	.044	49.143	1	495	.000

a. Predictors: (Constant), TV_CEN TV Time, weekdays (centered), SES Family Background, ABIL_CEN Ability (centered)

b. Predictors: (Constant), TV_CEN TV Time, weekdays (centered), SES Family Background, ABIL_CEN Ability (centered), TVXABIL TV by Ability crossproduct

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	49.937	.275		181.324	.000	49.396	50.479
	SES Family Background	1.442	.294	.168	4.909	.000	.865	2.020
	ABIL_CEN Ability (centered)	.561	.032	.610	17.794	.000	.499	.623
	TV_CEN TV Time, weekdays (centered)	-.423	.159	-.085	-2.655	.008	-.737	-.110
2	(Constant)	49.616	.267		185.892	.000	49.092	50.140
	SES Family Background	1.373	.281	.160	4.892	.000	.822	1.925
	ABIL_CEN Ability (centered)	.555	.030	.604	18.427	.000	.496	.614
	TV_CEN TV Time, weekdays (centered)	-.278	.154	-.056	-1.806	.072	-.580	.024
	TVXABIL TV by Ability crossproduct	-.113	.016	-.213	-7.010	.000	-.144	-.081

a. Dependent Variable: ACHIEVE Achievement Test Score

Figure 8.2 Regression results testing for an interaction between time spent watching TV and Ability in their effects on Achievement.

Probing an Interaction between Continuous Variables

With interactions between categorical and continuous variables, it is relatively easy to probe the interaction through graphing, because one variable already represents a limited number of categories. These categories can thus be plotted as separate lines in a graph of the dependent variable on the other (continuous) independent variable, and separate regressions can be conducted across the different categories. It is slightly more complex to investigate further a statistically significant interaction between two continuous variables. I will outline several methods by which you can get a sense of the nature of interactions between continuous variables and will briefly mention methods for more complete post hoc probing of such interactions.

One relatively easy method of getting a sense for such interactions involves converting (for the purposes of follow-up) one continuous variable into a limited number of ordered categories and conducting the same sorts of analyses that we used when one variable was categorical. For the current example, I converted the Ability variable into a new, trichotomized Ability variable (*Abil_3*, which is also included in the data set). On this new *Abil_3* variable, a value of 1 included approximately the lowest 33% of participants (on the Ability variable). The middle third of participants on the Ability variable was coded 2 on *Abil_3*. The top third of those on Ability was assigned a value of 3 on the *Abil_3* variable. Thus scores of 1, 2, and 3 on the *Abil_3* variable represent low, middle, and high ability, respectively. We can then use this trichotomized version of the Ability variable to graph the interaction and to conduct separate regressions.

Figure 8.3 shows three separate regression lines for the regression of Achievement on TV time for these three levels of Ability (SES is not taken into account in the graph). The graph clearly shows the nature of the interaction. It appears that TV viewing is considerably more detrimental for the achievement of high-ability youth than for other youth, in that each additional hour spent viewing TV appears to result in considerably lowered achievement for high-ability youth. In contrast, for students of average or lower ability, TV viewing seems to have little effect on their achievement. The results are consistent with previous research on

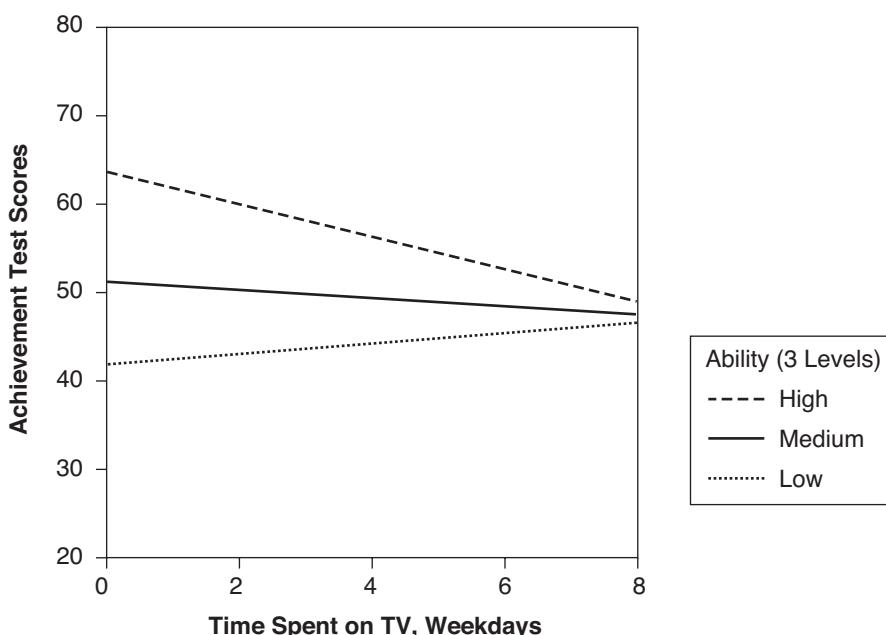


Figure 8.3 One method of exploring an interaction between continuous variables: regression of Achievement on TV for three levels of Ability.

the effects of TV viewing (e.g., Williams et al., 1982). Although not shown here, we could continue our post-hoc probing by using the trichotomized version of the Ability variable in follow-up regressions. In regressions of Achievement on SES and TV time, TV time was statistically significant, large, and negative for high ability students ($\beta = -.371$), not statistically significant for those of average ability ($-.081$), and statistically significant and positive for low ability youth (.165). This latter finding suggests that TV viewing may have a slightly positive effect on the achievement of low ability youth!

It is also possible to plot mean Achievement scores by levels of time spent on TV and (trichotomized) Ability (cf. Keith et al., 1986), as was done in Figure 8.4, to get a general sense of the nature of the interaction. This method of graphing is a potential alternative to the use of separate regression lines. Although this approach has some advantages—the variation in the lines is interesting—these lines represent means, not regression lines, and thus the nature of the difference in regression lines is less obvious. Also note that this procedure will only work if there are a limited number of possibilities for the independent variable being plotted on the X-axis, or the samples are large, or both. The present example fulfills these requirements, because there are only nine levels of the TV viewing variable.

Using the overall regression equation for follow-up. It is also possible to calculate the regression equation for any given value of Ability using the overall regression equation by substituting the desired values of Ability in the equation. The regression equation is

$$\text{Achieve}_{\text{predicted}} = 49.616 + 1.373\text{SES} + .555\text{Ability} - .278\text{TV} - .113\text{TV} \times \text{Ability}.$$

What values should be substituted? Common values are $-1 SD$, the mean, and $+1 SD$ on the continuous independent variable (Aiken & West, 1991; Cohen & Cohen, 1983). In the present analysis, these would be values of approximately -9 , 0 , and 9 on the centered Ability variable (the mean and SD are shown in Figure 8.1). Other values are also possible, including clinically relevant values or commonly used cutoffs. So, for example, if you are especially interested in

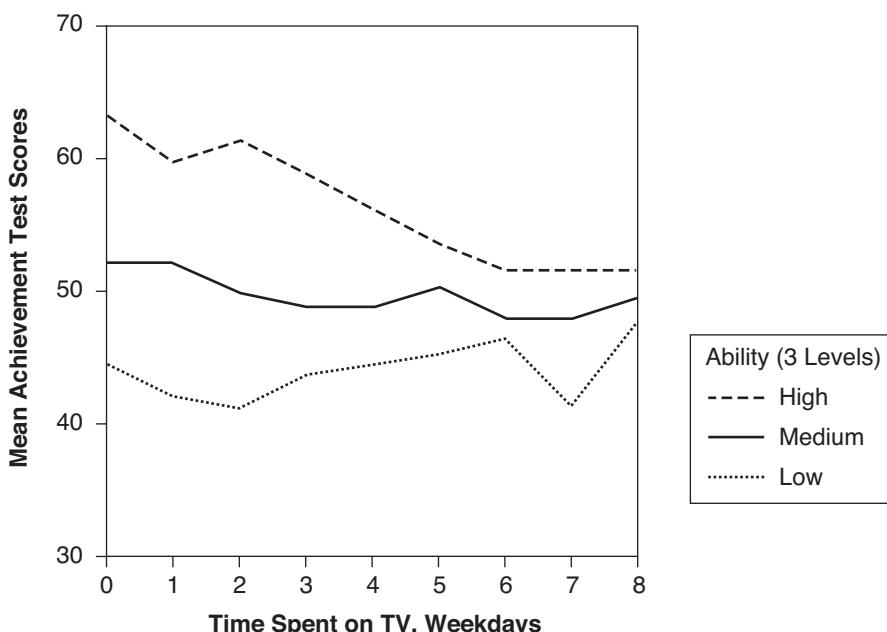


Figure 8.4 Mean levels of Achievement for different levels of TV viewing and Ability.

the implications of this research for low-ability students, you will probably want to calculate a regression equation for students whose ability is a standard deviation or more below the mean.

If you substitute values of 9, 0, and -9 for Ability (and in the Ability by TV interaction) in the above equation, you generate three new equations. The equation for high-ability youth is

$$\text{Achieve}_{\text{predicted}} = 49.616 + 1.373(0) + .555(9) - .278\text{TV} - .113\text{TV}(9).$$

Again, for this equation, +9 was substituted for Ability wherever it occurred in the overall regression equation. I also substituted a value of zero for SES (the population mean) to simplify the equations. The equation is simplified as

$$\begin{aligned}\text{Achieve}_{\text{predicted}} &= 49.616 + 4.995 - .278\text{TV} - 1.01\text{TV} \\ &= 54.611 - 1.29\text{TV}.\end{aligned}$$

For middle-ability youth, a value of zero is substituted for Ability. The regression equation is

$$\begin{aligned}\text{Achieve}_{\text{predicted}} &= 49.616 + 1.373(0) + .555(0) - .278\text{TV} - .113\text{TV}(0) \\ &= 49.616 - .278\text{TV}\end{aligned}$$

For low-ability youth, a value of -9 is substituted, resulting in this regression equation:

$$\begin{aligned}\text{Achieve}_{\text{predicted}} &= 49.616 + 1.373(0) + .555(-9) - .278\text{TV} - .113\text{TV}(-9) \\ &= 49.616 - 4.995 - .278\text{TV} + 1.017\text{TV} \\ &= 44.621 + .739\text{TV}\end{aligned}$$

These, then, are the regression equations for the regression of Achievement on TV time for high-, middle-, and low-ability youth. These equations can then be plotted (Figure 8.5) to

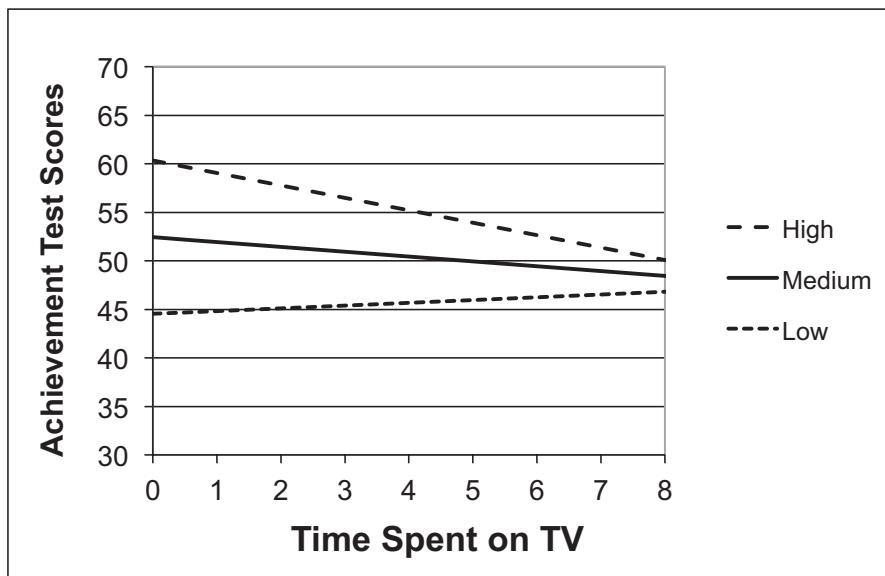


Figure 8.5 Using the overall regression equation to plot the effects of TV on Achievement for different levels of Ability.

demonstrate the nature of the interaction. (The graph was created in Excel. Because the homework variable is a meaningful metric I converted back to that metric for this display rather than using the centered version.) Although slightly more complex than the other methods outlined, this method has the advantage of being based on the original regression equation (rather than three new equations). In addition, it is possible to test the statistical significance of the slopes of the calculated regression equations. This topic is beyond the scope of this book but is presented in detail in Aiken and West (1991), Cohen and colleagues (2003), and Hayes (2018). Hayes's PROCESS program, designed to test for moderation, mediation, and a combination of the two, will calculate the new equations and test their statistical significance. For more detail concerning PROCESS, see Chapter 9.

To reiterate, there are several possibilities for exploring, through graphing, a statistically significant interaction with two continuous variables. They are (in reverse order):

1. Use the original regression equation to graph lines for different levels of one continuous variable. You can substitute, for example, values representing $+1\text{ SD}$, the mean, and -1 SD for one interacting variable to develop three regression equations representing participants who have high, medium, and low values on this variable.
2. Divide the sample into categories (e.g., lowest, middle, and high categories) on one of the interacting variables. Alternatively, you could make this division at $+1\text{ SD}$, the mean, and -1 SD . Plot a line for each category showing the mean level of the dependent variable for each level of the other interacting variable. This procedure requires large samples and a limited number of levels of the other interacting variable.
3. Divide the sample into categories (e.g., lowest, middle, and high categories) on one interacting variable. Alternatively, you could make this division at $+1\text{ SD}$, the mean, and -1 SD , or other clinically significant values. Plot a regression line, with the outcome variable regressed on the other interacting variable, for each category.

Points to Consider

Several aspects of these examples are worth noting. First, you should consider which continuous variable to categorize. We could have focused just as easily on high, medium, and low levels of TV viewing instead of high, medium, and low ability. In this case, our graph would have shown the regression of Achievement on Ability for low, medium, and high levels of TV viewing. Such a presentation strikes me as considerably less useful than the one presented; it would, for example, provide little illumination for parents wondering whether they should worry about their children's TV consumption. Basically, the way you choose to graph and analyze such interactions should depend on the questions you are interested in addressing. Different presentations answer different questions, so you should be clear about the questions you want to address and set up your graphs and additional analyses appropriately. You can often get a clue as to the variable to categorize by our original interest. Here we wondered whether ability moderated the effect of TV viewing on Achievement. Thus it makes sense to plot Achievement against TV viewing, with separate lines for the moderating variable, Ability.

Second, note that for the graphs I used the original metric of TV viewing, rather than the centered metric. Either will work, but since the metric of TV viewing is meaningful (hours per day), I didn't want to waste the interpretive advantages of this metric. If I were graphing a variable without such a meaningful metric, say self-esteem, I would probably choose the centered version of the variable.

Third, you may wonder if I have here adopted a practice that I previously criticized: the mind-set that sometimes leads researchers to categorize continuous variables so that they may be analyzed by ANOVA. Here I seem to be advocating such categorization. Note,

however, that I did not categorize Ability prior to the test of the interaction. The continuous variable was only converted to categories after a statistically significant interaction was found and as an aid to probing the nature of this interaction. My harsh criticism of categorizing continuous variables prior to analysis still stands.

CURVILINEAR REGRESSION

All the regression lines we have encountered so far have been straight lines. Indeed, as you will see in Chapter 10, linearity is one of the basic assumptions of regression. But it is also possible for a regression line to have curves in it. As an example, think of the likely relation of anxiety to test performance. If you have no anxiety at all about an upcoming exam, you likely will not study for it nor take it very seriously while it is being administered; the likely result is that you will not perform particularly well on the exam. At the other end of the anxiety spectrum, if you are extremely anxious about the same exam, your high anxiety will also likely inhibit your performance. Some middle level of anxiety should be most beneficial: enough anxiety to motivate you to study and perform well, but not so high as to interfere with your performance. If this expectation about anxiety and test performance is accurate (Teigen, 1995), the proper graph of test performance on anxiety might look something like that shown in Figure 8.6.

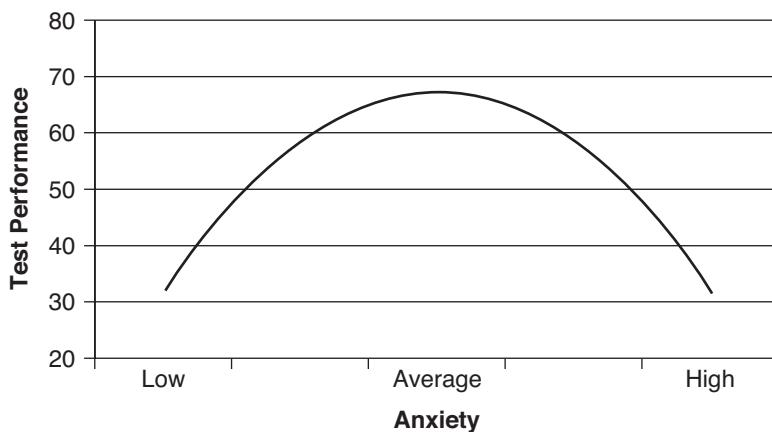


Figure 8.6 Curvilinear effect: test performance as a function of Anxiety.

Using normal linear regression, we would likely explain none of the variance in test performance based on anxiety; the regression line would be straight and flat. But it *is* possible to take into account possible curves in the regression line. How? Recall how we described the results of interactions by saying “it depends.” If asked to describe the effect of anxiety on test performance, we need to use this same language. What type of effect does anxiety have on test performance? It depends; it depends on the level of anxiety. For low levels of anxiety, anxiety has a positive effect on test performance, whereas for high levels of anxiety, anxiety has a negative effect on test performance. If the use of the term “it depends” signals a possible interaction, then in essence what we are saying is that anxiety interacts with *itself* in its effects on test performance. And if a curve in a regression line can be described as the interaction of a variable with itself on another variable, then the method of analysis also becomes clear: multiply the two variables that interact—in this case, multiply anxiety times anxiety—and enter the cross product in the regression equation following the original variable. Let’s turn to real data to illustrate the method.

Curvilinear Effects of Homework on GPA

We have examined in several ways the effect of homework on achievement and grades. But doesn't it seem likely that homework's effect on learning should be curvilinear? Certainly, homework improves learning, but don't you think that there will be diminishing returns for each additional hour of homework completed? In other words, shouldn't the payoff for learning be greater when going from zero to 1 hour per week than when going from, say, 10 to 11 hours per week? In fact, research on homework suggests exactly this type of curvilinear relation: there are diminishing returns for the effects of homework on learning (cf. Cooper, 1989; Fredrick & Walberg, 1980).

The Data: Homework and Homework Squared

This expectation for diminishing returns for homework is in fact built into the NELS data, at least to some extent. Look at the values of the homework variable, shown in Figure 8.7. Note that for lower values of homework the increment is 1 hour or less (e.g., from zero hours to 1 hour or less), whereas for later values the increment is greater (e.g., a value of 6 is used to describe 13, 14, or 15 hours of homework). This compression of the homework scale at the upper end may take some of the likely curvilinear effect of homework on learning into account. We'll see here if the effect is still curvilinear.

Let's be a little more explicit: we will test the effect of time spent on out of school Homework in grade 10 on students' 10th-grade GPA. We are interested in testing for possible curvilinear effects for Homework, so we will use both the Homework variable and a Homework-squared variable in the regression. Just as in our tests for interactions, we will first center the continuous Homework variable prior to squaring it and will use centered Homework and centered Homework squared in the regression. We will also control for students' family background, or Socioeconomic Status, and Previous Achievement, with the thinking that SES and Previous Achievement may affect both Homework and subsequent Grades.

Figure 8.8 shows the descriptive statistics and correlations for the variables used in the analysis. All these variables are included in your version of the NELS data, except two: HW_Cen and HW_Sq. HW_Cen is the centered version of the Homework

F1S36A2 TIME SPENT ON HOMEWORK OUT OF SCHOOL

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 NONE	63	6.3	6.7	6.7
	1 1 HOUR OR LESS	232	23.2	24.6	31.3
	2 2-3 HOURS	264	26.4	28.0	59.3
	3 4-6 HOURS	168	16.8	17.8	77.1
	4 7-9 HOURS	80	8.0	8.5	85.6
	5 10-12 HOURS	66	6.6	7.0	92.6
	6 13-15 HOURS	31	3.1	3.3	95.9
	7 OVER 15 HOURS	39	3.9	4.1	100.0
Missing	Total	943	94.3	100.0	
	96 MULTIPLE RESPONSE	7	.7		
	98 MISSING	17	1.7		
	System	33	3.3		
Total	Total	57	5.7		
		1000	100.0		

Figure 8.7 The scale of the Homework time variable in NELS.

Descriptive Statistics

	Mean	Std. Deviation	N
FFUGRAD ffu grades	5.6866	1.4726	896
BYSES SOCIO-ECONOMIC STATUS COMPOSITE	2.17E-02	.77097	896
BYTESTS 8th-grade achievement tests (mean)	51.8150	8.7000	896
HW_CEN Homework out of school, centered	-1.5E-13	1.7110	896
HW_SQ Homework centered, squared	2.9243	4.3862	896

Correlations

	FFUGRAD ffu grades	BYSES SOCIO-ECONOMIC STATUS COMPOSITE	BYTESTS Eighth grade achievement tests (mean)	HW_CEN Homework out of school, centered	HW_SQ Homework centered, squared
Pearson Correlation					
	1.000	.311	.494	.325	.097
	.311	1.000	.467	.285	.134
	.494	.467	1.000	.304	.138
	.325	.285	.304	1.000	.582
	.097	.134	.138	.582	1.000

Figure 8.8 Descriptive statistics for the variables used in the curvilinear regression example.

variable, created by subtracting the mean of F1S36A2 from F1S36A2 [e.g., compute Hw_Cen=(F1S36A2-2.544642857143)]. HW_Sq was created by squaring HW_Cen. Note the correlation between HW_Sq and HW_Cen: .582. Had we not centered the Homework variable prior to squaring it, the correlation between Homework and Homework squared would have been .953.

Tenth-grade GPA was regressed on SES, Previous Achievement, and HW_Cen in one block, and HW_Sq was sequentially added in a second block in the regression. Note that we could just as easily have added all variables in a single block to determine the statistical significance of the curve in the regression line (using the *t* test of the HW_Sq regression coefficient).¹

The Regression

Figure 8.9 shows the results of the multiple regression. As shown in the Model Summary, the addition of the HW_Sq term to the regression resulted in a statistically significant increase in the variance explained by the regression ($\Delta R^2 = .008$, $F = 10.366$ [1, 891] $p = .001$). There is a statistically significant curve in the regression line. Of course, the statistical significance of the HW_Sq variable in the lower half of the table of coefficients leads to the same conclusion.

Graphing the Curve

The curved regression line is shown in Figure 8.10 (created by specifying a quadratic fit line as chart option in SPSS's scatterplot command; SES and Previous Achievement are not controlled in this graph). Our findings are consistent with previous research, and it appears our speculation was correct: for lower levels of homework, grades improve fairly quickly for each

Model Summary^c

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.531 ^a	.282	.282	116.543	3	892	.000
2	.538 ^b	.290	.008	10.366	1	891	.001

a. Predictors: (Constant), HW_CEN Homework out of school, centered, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYTESTS 8th-grade achievement tests (mean)

b. Predictors: (Constant), HW_CEN Homework out of school, centered, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYTESTS 8th-grade achievement tests (mean), HW_SQ Homework centered, squared

c. Dependent Variable: FFUGRAD ffu grades

Coefficients^a

Model	Unstandardized Coefficients			t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Standardized Coefficients			Lower Bound	Upper Bound
	Beta						
1	(Constant)	2.115	.290	7.296	.000	1.546	2.683
	BYSES	.133	.062	.069	2.132	.033	.011 .255
	SOCIO-ECONOMIC STATUS COMPOSITE						
	BYTESTS 8th-grade achievement tests (mean)	6.89E-02	.006	.407	12.421	.000	.058 .080
	HW_CEN Homework out of school, centered	.156	.026	.181	5.993	.000	.105 .207
2	(Constant)	2.258	.292	7.741	.000	1.686	2.831
	BYSES	.128	.062	.067	2.074	.038	.007 .250
	SOCIO-ECONOMIC STATUS COMPOSITE						
	BYTESTS 8th-grade achievement tests (mean)	6.82E-02	.006	.403	12.359	.000	.057 .079
	HW_CEN Homework out of school, centered	.214	.031	.248	6.786	.000	.152 .275
	HW_SQ Homework centered, squared	-3.8E-02	.012	-.112	-3.220	.001	-.060 -.015

a. Dependent Variable: FFUGRAD ffu grades

Figure 8.9 Regression results testing for a curvilinear effect of Homework on GPA.

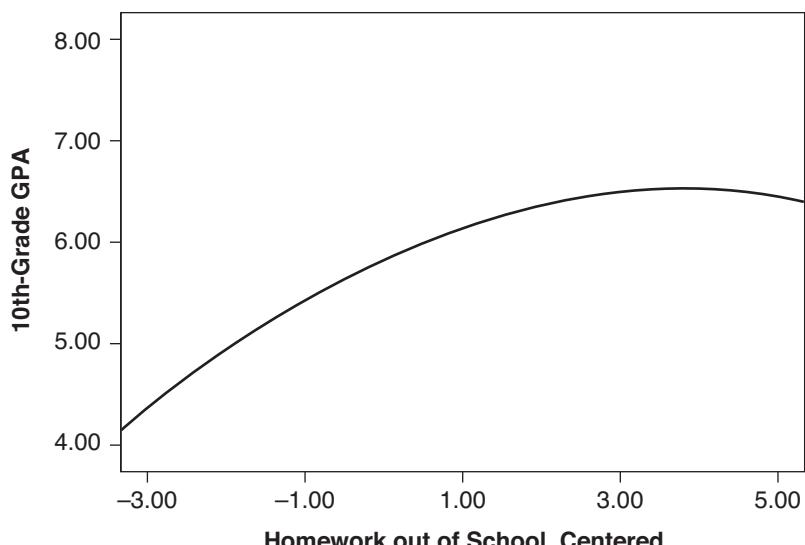


Figure 8.10 Plot of the curvilinear effect of Homework on 10th-grade GPA, using the centered Homework variable.

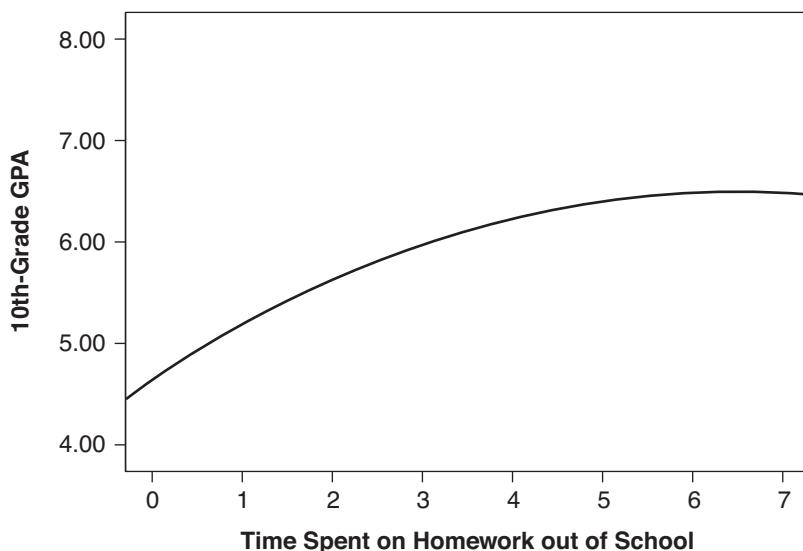


Figure 8.11 Another plot of the curvilinear effect of Homework on 10th-grade GPA, using the original Homework variable.

Table 8.1 Relation between Regression Coefficients in a Curvilinear Regression and the Trend and Shape of the Regression Line

Coefficient Associated with:	What it Describes	Is the Value	
		Positive	Negative
Unsquared variable	Trend of the regression line	Upward trend	Downward trend
Squared variable (curve component)	Shape of the regression line	Concave shape	Convex shape

unit increase in homework, but these increases quickly flatten out; so for students already completing substantial amounts of homework, a unit increase in homework has little or no effect on GPA. This initial graph uses the centered homework variable, but the regression line using the uncentered data is shown in Figure 8.11. Note that the two graphs are essentially the same, with the only difference being the scale of the X-axis.

Note the shape of the regression line: primarily upward, with a convex shape. This shape is also revealed by the regression coefficients in the bottom half of the table of coefficients in Figure 8.9. The *positive* coefficient for HW_Cen suggests the general upward trend of the regression line, whereas the negative coefficient for the curve component (HW_Sq) suggests the gradually flattening, convex shape. In contrast, if there were a negative coefficient for the independent variable that would suggest a generally downward trend to the regression line, and a positive coefficient for the squared independent variable would suggest a concave shape. These relations between regression coefficients and the regression line are summarized in Table 8.1. Given this description, what do you think the coefficients associated with Figure 8.6 might be? The coefficient for Anxiety would be zero, and the coefficient for anxiety squared would be negative.

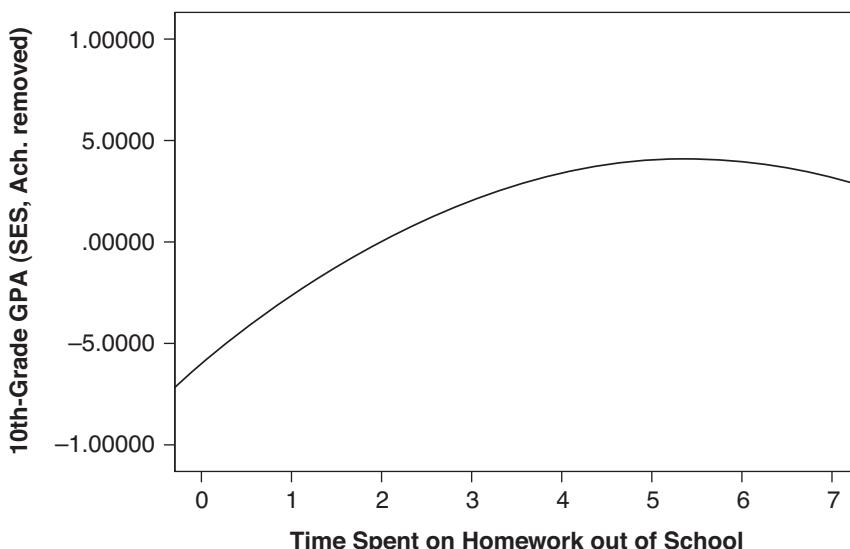


Figure 8.12 Plot of the curvilinear effect of Homework on 10th-grade GPA; SES and Previous Achievement are also controlled.

Controlling for Other Variables

In the multiple regression, we controlled for SES and Previous Achievement when examining the linear and curvilinear effect of homework on GPA, but SES and Previous Achievement were not considered in the graphs. It is also possible to take SES and Previous Achievement into account in these graphs. Recall in our discussion of residuals we found that residuals could be considered as the dependent variable with the effects of the independent variables removed. In the present case we are interested in plotting the effects of homework on GPA, with the effects of Previous Achievement and SES removed from GPA. Thus, we can easily regress GPA on SES and Previous Achievement, saving the residuals, which now represent GPA with SES and Previous Achievement taken into account. Figure 8.17 shows the curvilinear regression line for GPA, with SES and Previous Achievement removed, on Homework. The variable now labeled 10th-Grade GPA (SES, Ach removed) is, in turn, the saved residuals from the regression of GPA on SES and Previous Achievement.

Testing Additional Curves

Is it possible to have more than one curve in the regression line? Yes; for example, consider the possible effects of student employment during the school year on achievement. It may be that working a few hours a week is actually beneficial to student achievement, but that as students work beyond these few hours, their achievement suffers (this describes one curve in the regression line). Beyond a certain number of hours, however, additional hours may have no effect, and therefore the line would flatten out (where the slope changes from negative to flat describes another curve). Figure 8.13 illustrates such a possibility (cf. Quirk, Keith, & Quirk, 2001 see Neyt, Omey, Baert, & Verhaest, in press for a review of student employment literature.).

To test for additional curves, we simply test additional powers of the independent variable. To test for one curve in the regression line, we add the centered independent variable squared (a quadratic term) to the regression equation. To test for two curves, we additionally add a cubed version of the centered independent variable to the equation; to test for three curves, we add the independent variable to the fourth power, and so on. Figure 8.14 shows some of the results from the regression of GPA on the control

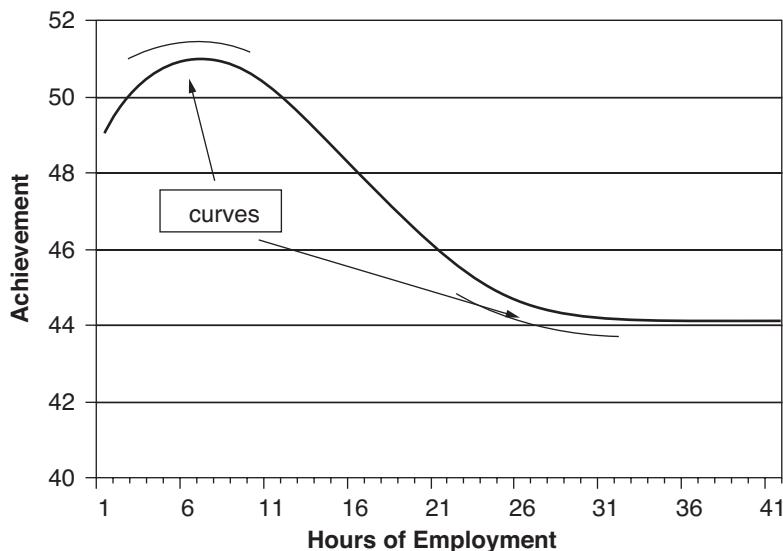


Figure 8.13 Graph of a regression line with two curves. These curves can be tested by adding variables representing Employment-squared and Employment-cubed to the regression equation.

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.531 ^a	.282	.282	116.543	3	892	.000
2	.538 ^b	.290	.008	10.366	1	891	.001
3	.539 ^c	.290	.000	.364	1	890	.547

- a. Predictors: (Constant), HW_CEN Homework out of school, centered, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYTESTS 8th-grade achievement tests (mean)
- b. Predictors: (Constant), HW_CEN Homework out of school, centered, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYTESTS 8th-grade achievement tests (mean), HW_SQ Homework centered, squared
- c. Predictors: (Constant), HW_CEN Homework out of school, centered, BYSES SOCIO-ECONOMIC STATUS COMPOSITE, BYTESTS 8th-grade achievement tests (mean), HW_SQ Homework centered, squared, HW_CUBE Homework centered, cubed

Figure 8.14 Testing for two curves in the regression equation.

variables, Homework, Homework squared, and Homework cubed. As shown, the cubic term entered at the third step was not statistically significant. There is only one curve in the regression line, and the shape of the regression line was adequately graphed in previous figures. It may be worthwhile to test such higher-order terms until statistical nonsignificance is found.

There are other methods of transforming data beyond the power transformations (X -squared, X -cubed, etc.) discussed in this chapter. For example, logarithmic transformations are possible, as are square root transformations. According to Cohen and colleagues (2003, p. 221), one major reason for such transformations is to simplify the relation between the predictor and outcome variables. For example, it is common to use a logarithmic transformation

of income in regression rather than income, per se. Other reasons involve dealing with threats to regression assumptions: homoscedasticity and normal distributions of residuals (these topics are discussed in Chapter 10). Finally, for complex nonlinear models, there is the method of nonlinear regression that can go well beyond the simple modeling of curves in a regression line that we are able to accomplish with ordinary multiple regression.

As with interactions in multiple regression, curves in regression lines are relatively rare, especially regression lines with more than one curve. As with interactions, it may be that curvilinear effects are rare or that straight lines are simply reasonably good approximations in most cases. It is also the case, however, that these tests are less often statistically significant because of their lower power. In addition, unusual data points can sometimes trick you into thinking you have a curve in your regression line. You should always inspect your data for such anomalies. These outliers will be discussed in more detail in Chapter 10.

SUMMARY

This chapter extended our discussion of testing for interaction in multiple regression to interactions involving continuous variables. Simulating data from published research, we regressed Achievement on Ability, time spent in leisure TV viewing, and a cross product of TV and Ability to determine whether TV viewing interacts with Ability in its effect on Achievement. The findings indeed suggested the presence of an interaction. We discussed several methods for exploring the nature of such interactions. First, we divided the Ability variable into three categories and graphed regression lines of Achievement on TV viewing for these three levels of Ability. Second, we plotted mean levels of Achievement by each level of TV and Ability. Third, we used the overall regression equation and substituted values representing low, medium, and high ability (-1 SD , the mean, and $+1\text{ SD}$) into the equation to generate three regression equations. These three equations were also plotted to explore the nature of the interaction. These methods should help you understand and describe the nature of any interaction that you do find.

We introduced regression lines including curves in this chapter, as well, and conceptualized such curve components as an interaction of a variable with itself in its influence on some outcome. Returning to an earlier example, we showed that Homework may, in fact, have a curvilinear effect on GPA, such that each additional hour spent on homework has a smaller effect on GPA than did the previous hour. We uncovered this curvilinear effect by adding a Homework-squared variable to the regression equation and testing its statistical significance. Higher-order terms (e.g., Homework cubed) could be added to test for additional curves in the regression line. Again, graphs were used to understand the nature of the curvilinear effect.

EXERCISES

1. If you have not done so already, conduct the multiple regression testing the interaction of TV and Ability on Achievement conducted earlier in this chapter. Compare your results to mine. Make sure you are able to correctly center the variables and create the interaction term. Try the different methods for graphing the interaction. The data are on the Web site (tzkeith.com) ("tv ability interact2.sav," "tvability interact2.xls," and "tv_abil.txt").
2. Conduct a similar analysis using the NELS data. Try using F1S45A as the measure of time spent watching TV and a mean of the 10th-grade test scores (F1TxRStd, F1TxM-Std, F1TxSStd, F1TxHStd) as the outcome. Because NELS did not include measures of ability, test the interaction of TV and Previous Achievement (ByTests). Also control for base year SES (BySES). Is the interaction statistically significant? Graph the interaction

(or the lack of an interaction). How do you account for the differences between these findings and those from Exercise 1?

3. Conduct the multiple regression testing the curvilinear effect of Homework on Grades conducted earlier in this chapter. Compare your results to mine. Make sure you are able to correctly center the variables and create the Homework squared term. Graph the curved regression line.
4. Does TV viewing have a curvilinear effect on Grades? Spend a few minutes thinking about this question. If you believe TV viewing has such an effect, what do you think will be the shape of the regression line: negative and concave; negative and convex? Use NELS to test this question. Use F1S45A as a measure of TV viewing and FFUGrad as a measure of 10th-grade GPA. Also control for SES and Previous Achievement (BySES and ByTests).

Note

- 1 Technically, because there are additional independent variables, we are not testing for a curve in a line but rather a curve in a regression plane.

9

Mediation, Moderation, and Common Cause

Moderation	177
Mediation	179
<i>Baron and Kenny (Causal Steps)</i>	180
<i>Joint Significance</i>	181
<i>Sobel Test</i>	181
<i>Bootstrapping and PROCESS</i>	181
<i>Mediation Example</i>	182
Common Cause	187
A Further Comment on Language	192
Summary	193
Exercises	194

The previous two chapters focused on interactions in multiple regression. As noted, interactions can go by another label, moderation. In other chapters we briefly discussed the concept of mediation. I find that many people, including seasoned researchers, confuse these two concepts. In this chapter I want to discuss these two concepts together and draw a distinction between them. In addition, we will differentiate both moderation and mediation from common cause, a topic we have touched on at several points in this text (see especially Chapter 4). I hope these three concepts are clear and distinct in your mind, but, in my experience, students often confuse the nature of these three concepts, especially with regard to how they show up in multiple regression. Perhaps what is confusing is that these concepts all involve the influence of one variable on another, and how that effect is changed by a third variable. Here we will also explore briefly the combination of mediation and moderation (mediated moderation and moderated mediation). As in previous chapters, we will use path diagrams to illustrate important concepts.

MODERATION

Moderation means the same thing as interaction. When we say that ability moderates the effect of TV viewing on achievement, this is the same as saying that ability and TV viewing interact in their effect on achievement. Likewise, it is equivalent to saying that the effect of TV viewing differs for different levels of ability or that TV viewing has different effects for those of high ability versus those of low ability. Said differently, the magnitude of the effect

of TV viewing on achievement differs for different levels of ability. Because regression coefficients represent the slope of the regression line, moderation is often described as differences in slopes across groups. Interactions, or moderation, can often be described using the statement “it depends.” If you found, for example, that sex moderated the effect of motivation on achievement, and someone were to ask you about the effect of motivation on achievement, you would need to use the words “it depends . . . it depends on whether you are a boy or a girl.” When you hear the term “moderated regression,” it generally means to test for moderation (interaction) using the regression procedures outlined in the previous two chapters. Another way of describing moderation is to think of the phrase “different slopes for different folks,” a phrase that may be easy to remember if you like classic rock and R&B.

As noted in Chapter 7, path diagrams can be used to illustrate moderation. Figure 9.1 shows that the variable Group membership moderates the effect of Influence on Outcome, in that the magnitude of the effect of Influence on Outcome changes depending on Group membership.

Figure 9.2 illustrates another method of displaying moderation. Here, the diagram illustrates that Group membership affects—moderates—the extent of the influence of Influence on Outcome. Figure 9.3 illustrates this same moderation example via regression lines drawn for the two groups. Like Figure 9.1, this graph illustrates that Influence has a large effect on Outcome for Group 1 but a small effect for Group 2.

As we have illustrated in the last two chapters, the primary method of testing for moderation in multiple regression is to create a cross product of the two variables that are suspected of interacting and add that cross product sequentially to the regression analysis. A statistically significant increase in ΔR^2 suggests that the interaction is statistically significant, that the effect of one variable is indeed moderated by the other. This methodology suggests yet

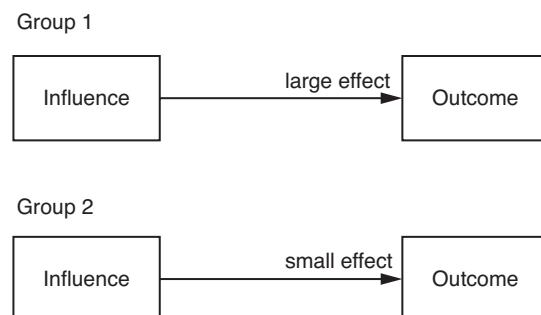


Figure 9.1 One method of illustrating moderation. Group membership moderates the effect of Influence on Outcome.

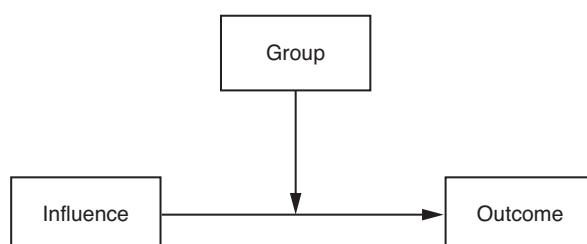


Figure 9.2 A second method of illustrating moderation. Group membership moderates (affects) the effect of Influence on Outcome.

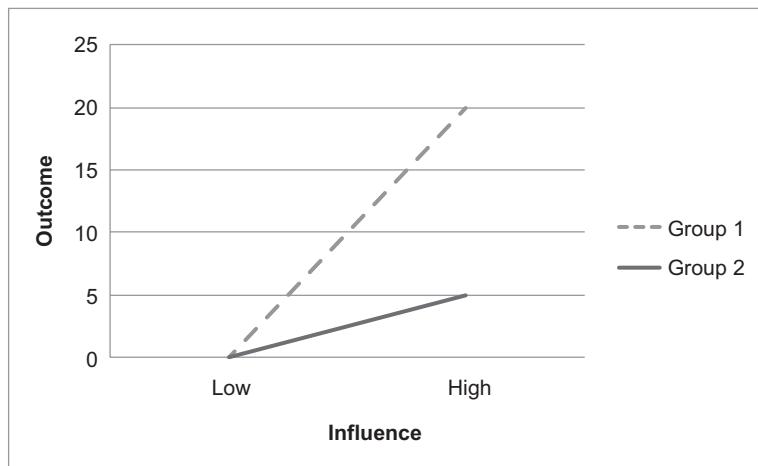


Figure 9.3 Moderation illustrated via regression lines for Group 1 and Group 2.

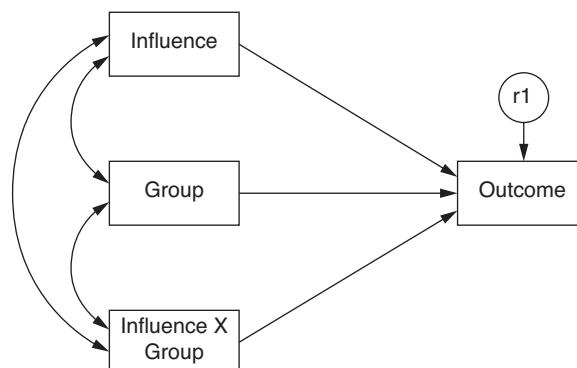


Figure 9.4 Moderation illustrated via the addition of a cross product in regression. This is the method we have used to test for interactions (moderation).

another way for illustrating moderation figurally. Figure 9.4 shows the Outcome variable regressed on Influence, Group, and an Influence-by-Group cross product.

This method of testing for moderation in multiple regression is the same for categorical and continuous variables, although follow-up analyses differ. Graphing, as in Figure 9.3, is a useful follow-up for either.

MEDIATION

The term mediation means the same thing as an indirect effect. When we say that motivation affects achievement through homework, this is the same as saying that motivation has an indirect effect on achievement through homework or that homework mediates the effect of motivation on achievement. We can describe this relation by explaining that more motivated students complete more homework and that homework, in turn, increases their achievement. Thus mediation is useful for understanding *how* an effect comes about. If I tell you that academic motivation influences achievement, you may wonder how. Mediators are attempts to explain how motivation affects achievement by affecting time spent on

homework; more motivated students complete more homework, and that homework, in turn, improves achievement.

Mediation, or indirect effect, is nicely illustrated via path diagrams such as the one shown in Figure 9.5. Although we have discussed mediation or indirect effects (see, for example, Chapters 4 and 5), we have not really discussed how to *test* for mediation using multiple regression. There are several possible ways.

Baron and Kenny (Causal Steps)

In a classic article on this topic, Baron and Kenny (1986) showed that mediation can be assumed to exist under the following conditions (see also Judd & Kenny, 1981):

1. In a regression of Outcome on Influence (using the labels from Figures 9.5 and 9.6), the effect of Influence on Outcome is statistically significant. This regression does not include the mediating variable. It is illustrated as path d in Figure 9.6.
2. The regression of the Mediator on the Influence results in a statistically significant effect (path a in Figure 9.5).
3. The regression of Outcome on both Influence and the Mediator results in a statistically significant effect for the Mediator on the Outcome (path b, Figure 9.5), controlling for the Influence.
4. The regression of Outcome on both Influence and the Mediator results in a reduction in the effect of Influence from step 1 (path d, Figure 9.6). In other words, the effect represented by path c in Figure 9.5 is smaller than was the effect without the Mediator in the regression (path d, Figure 9.6). In a strict version of step 4, complete mediation exists when the addition of the Mediator to the regression reduces the coefficient c (Figure 9.5) to zero. Partial mediation exists when the effect is simply reduced.

These conditions have been debated and adjusted over the years by Kenny and others, but this presentation will do for our purposes (see Kenny's excellent web resources for more detail about this and other methods: www.davidakenny.net).

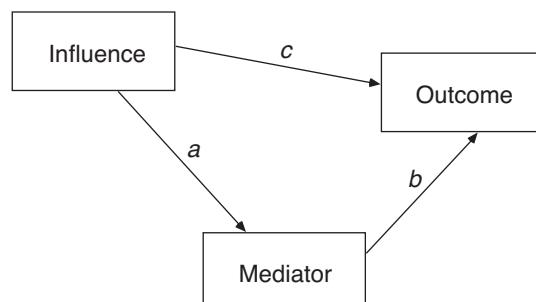


Figure 9.5 Mediation illustrated. Influence affects Outcome, in part through the Mediator.

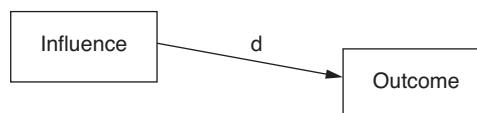


Figure 9.6 Baron and Kenny (1986) test of mediation, step 1.

Joint Significance

The strictest version of the Baron and Kenny causal steps approach requires that the value for d (Figure 9.6) goes from statistically significant to zero when the mediator is added (path c in Figure 9.5). As shown in Fritz and MacKinnon (2007), this is a very low-power test, unlikely to find statistical significance without large effects or large sample sizes. Less strict versions (partial mediation) have more power. In addition, the Baron and Kenny approach may be better thought of as a set of conditions necessary for mediation to occur rather than a series of tests for mediation. MacKinnon and colleagues (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002) showed that focusing on steps 2 and 3, however, can provide a useful test of mediation. That is, if paths a and b (Figure 9.5) are both statistically significant, then one can conclude that mediation occurs. Paths a and b may be estimated by first regressing Outcome on Influence and Mediator (for path b) and then regressing Mediator on Influence (a). Simulation work has shown this approach to provide better power than a strict Baron and Kenny approach or the Sobel test (next) (Fritz & MacKinnon, 2007).

Sobel Test

What is really being evaluated in a test of mediation is the magnitude (and statistical significance) of the indirect effect of Influence on Outcome *through* Mediator. This indirect effect is easily calculated as the product of paths a and b in Figure 9.5. Thus, via multiple regression:

1. Regress Outcome on Influence and Mediator. Note the regression coefficients for Influence (path c) and Mediator (path b).
2. Regress Mediator on Influence. Note the regression coefficient (path a).
3. Calculate the indirect effect by multiplying the coefficient associated with path a times that associated with path b.

The tricky part of the estimation of mediation is determining whether this indirect effect is statistically significant. It is worth noting that all structural equation modeling programs can easily calculate such indirect effects and their statistical significance, a topic we will cover in Part 2 of this book. Let's continue to discuss how to test for such effects within the context of multiple regression, however.

One traditional method to determine the statistical significance of an indirect effect is to calculate the standard error of the indirect effect based on the standard errors of the unstandardized regression coefficients associated with paths a and b. A common method is the Sobel test (Sobel, 1982). The indirect effect is then compared to its standard error, just like we first did for regression coefficients in Chapter 1. The standard error can also be used to create confidence intervals around the indirect effect. There are several such tests, although I think most people use the term "Sobel test" fairly generically; these are also referred to as the "normal theory approach."

As of this writing, Kris Preacher at Vanderbilt University has an excellent web page that will perform these calculations for you interactively (www.quantpsy.org/sobel/sobel.htm). All you need to do is to type in the regression coefficients and their standard errors (or the *t* values). The web site provides a great discussion of mediation (including reasons why the Sobel test is generally not the best option for testing mediation).

Bootstrapping and PROCESS

As previously noted, the Sobel test and variations generally are not the best option for testing for mediation. This method of calculating standard errors of indirect effects assumes that

the underlying distribution is normal, a questionable assumption without large samples. The test is also relatively low powered, meaning that it will often suggest a non-statistically significant indirect effect when it should not (Fritz & MacKinnon, 2007; Hayes, 2018). Based on a considerable body of simulation work, methodologists now generally recommend the use of bootstrapped standard errors, and this is, in fact, the method generally used by SEM programs to calculate the statistical significance of indirect effects. With bootstrapping, standard errors are estimated by taking repeated random samples from the existing data, and thus requiring fewer assumptions about the normality of the distribution. As noted, this can be accomplished via SEM programs. Several of the web sites mentioned (www.davidakenny.net, www.quantpsy.org/sobel/sobel.htm) have or point to SPSS, SAS, or R macros that test for mediation in multiple regression via bootstrapping.

The most complete and up-to-date series of such macros I know of are Andrew Hayes's PROCESS macros for SPSS and SAS (www.processmacro.org). PROCESS conducts mediation analysis via bootstrapping (and, if requested, the Sobel test), moderation analysis, and conditional process analysis (a combination of mediation and moderation, often referred to as moderated mediation). For those interested in exploring these topics in more depth I strongly recommend Hayes's book on mediation, moderation, and the PROCESS approach (Hayes, 2018).

Mediation Example

I illustrated how mediation might work earlier in the chapter with an example involving academic motivation, homework, and achievement. Suppose you were to conduct research examining the possible effect of academic motivation (working hard in school, believing school learning to be important, etc.) on subsequent achievement. Assuming you found such an effect you might next wonder how such an effect comes about. By what mechanism does motivation affect or improve achievement? You might, in turn, reason that students who are more motivated are likely to spend more time on homework, and that increased homework, in turn, should improve achievement. Another way of saying this is that you suspect homework to mediate the effect of motivation on achievement.

Figure 9.7 shows a model used to simulate data for this example. In the model, Homework is indeed a possible mediator between Motivation and Achievement. There are also

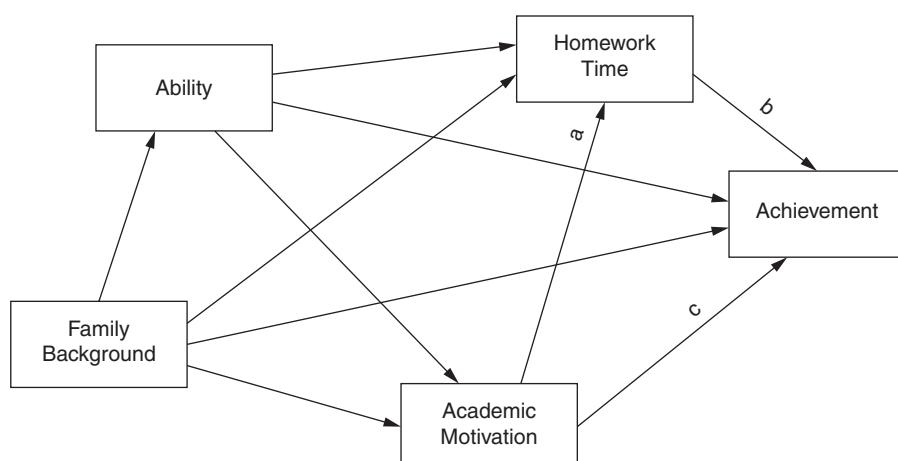


Figure 9.7 Mediation model: Motivation, Homework, and Achievement.

	Descriptive Statistics				
	N	Minimum	Maximum	Mean	Std. Deviation
FamBack	579	-3.267687	2.970156	-.02129082	1.006706031
Ability	579	60.00	147.00	102.2021	15.17894
Motivate	579	22.00	78.00	49.1088	10.36047
HWork	579	.00	7.00	2.4991	1.52644
Achieve	579	19.00	85.00	50.8463	11.22168
Valid N (listwise)	579				

Figure 9.8 Descriptive statistics for the Motivation-Homework mediation example data.

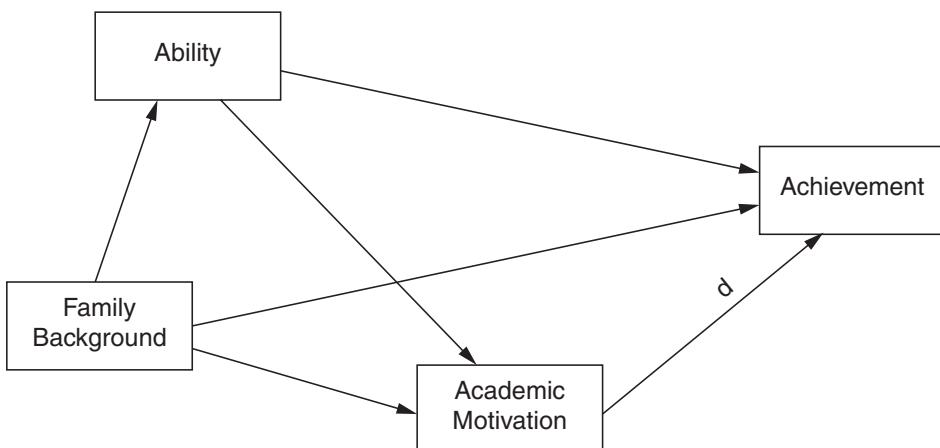


Figure 9.9 Step 1, Baron and Kenny test for mediation.

two possible background variables/covariates: Family Background (SES) and Cognitive Ability. Note that this example is not based on a particular research study. I suspect these relations are plausible, and are likely smaller than those I have posited here, but the purpose here is to illustrate testing for mediation. Figure 9.8 shows descriptive statistics for the simulated data. Motivation and Achievement are set up as T -scores ($M = 50$, $SD = 10$); Ability is on a deviation IQ scale ($M = 100$, $SD = 15$); Family Background is a mean of z -scores; and Homework is in hours per week. All descriptive statistics look reasonable.

Consider the series of regressions needed to test for mediation using the Baron and Kenny approach. If the background variables are to be included—and there are no reasons not to—you will need to regress Achievement on Family Background, Ability, and Motivation. This will provide the regression coefficient/path labeled d in Figure 9.9 needed for step 1 of this approach. Next, regress Homework Time on Academic Motivation (controlling for Family Background and Ability). This provides an estimate for path a in Figure 9.7, as needed for step 2 of the Baron and Kenny approach. Finally, regress Achievement on both Academic Motivation and Homework Time, also controlling for Family Background and Ability. This regression provides the regression coefficients b and c in Figure 9.7, and provides information for step 3 and 4 in this approach. When actually conducting the regressions it might make more sense to do

them in some other order, but let's stick to the 1, 2, 3, and 4 approach for the sake of illustration.

Baron and Kenny Causal Steps

The tables of coefficients for these regressions are shown in Figure 9.10, with the coefficients of interest bolded. The coefficients are also labeled (a, b, c, d) to correspond to the paths in Figures 9.7 and 9.8. The Baron and Kenny steps:

1. In a regression of Outcome on Influence (using the labels from Figures 9.5 and 9.6), the effect of Influence on Outcome is statistically significant. For this example, in a regression of Achievement on Motivation, the effect of Motivation on Achievement is statistically significant. As shown in the table of coefficients, this regression coefficient is indeed statistically significant and the β is moderate ($b = .123$, $SE_b = .035$, $p = .001$, $\beta = .114$).
2. The regression of the Mediator (Homework) on the Influence (Motivation) results in a statistically significant effect. Step 2 is also fulfilled ($b = .047$, $SE_b = .005$, $p < .001$, $\beta = .320$).
3. The regression of Outcome (Achievement) on both Influence (Motivation) and the Mediator (Homework) results in a statistically significant effect for the Mediator (Homework) on the Outcome (Achievement). Again, this condition is fulfilled: $b = 1.566$, $SE_b = .299$, $p < .001$, $\beta = .213$.
4. The regression of Outcome (Achievement) on both Influence (Motivation) and the Mediator (Homework) results in a reduction in the effect of Influence (Motivation) from step 1. The regression weight for Motivation in step 1 was of moderate magnitude and statistically significant ($b = .123$, $SE_b = .035$, $p = .001$, $\beta = .114$). In step 3 that coefficient is smaller and is no longer statistically significant ($b = .049$, $SE_b = .037$, $p = .188$, $\beta = .045$).

Thus, according to this series of steps, Homework does indeed mediate the effect of Motivation on Achievement. Does Motivation affect Achievement? Yes, it does. *How* does Motivation affect Achievement? It appears to work indirectly through Homework. That is, more motivated students complete more homework, and that homework, in turn, produces higher achievement.

A few things to note. First, the fact that the path from Motivation to Achievement reduced from a significant to a nonsignificant effect in step 1 compared to step 4 suggests that Homework completely mediates the effect of Motivation on Achievement (see step 4 at davidakenny.net/cm/mediate.htm). Second, as we will see in Chapter 12, and as was hinted at in Chapter 5, the coefficient for Motivation on Achievement in step 1 may be considered the *total effect* for Motivation on Achievement (combination of direct and indirect effect).

Joint Significance

The joint significance criterion for mediation focuses on the statistical significance of paths a and b in Figure 9.7; these coefficients, standard errors, and so on, are shown in the output for steps 2 and 3–4 in Figure 9.10. Both are statistically significant (for path a, the

Coefficients: Baron and Kenny Step 1^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.554	3.035	.182	.855	-5.408	6.515
	FamBack	.616	.399			-.168	1.399
	Ability	.433	.027			.381	.485
	Motivate (d)	.123	.035			.053	.193

a. Dependent Variable: Achieve

Coefficients: Step 2^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3.721	.413	-9.001	.000	-4.534	-2.909
	FamBack	.280	.054			.173	.387
	Ability	.038	.004			.031	.045
	Motivate (a)	.047	.005			.038	.057

a. Dependent Variable: HWork

Coefficients: Step 3^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.381	3.170	2.013	.045	.154	12.607
	FamBack	.177	.399			-.606	.961
	Ability	.373	.028			.317	.429
	Motivate (c)	.049	.037			-.024	.123
	HWork (b)	1.566	.299			.978	2.154

a. Dependent Variable: Achieve

Figure 9.10 Regression results for Baron and Kenny mediation test.

path from Motivation to Homework, $b = .047$, $SE_b = .005$, $p < .001$, $\beta = .320$; for path b, the path from Homework to Achievement, $b = 1.566$, $SE_b = .299$, $p < .001$, $\beta = .213$). According to this criterion, Homework mediates the effect of Motivation on Achievement.

Sobel Test

To conduct the Sobel test for the Motivation/Homework/Achievement example we need the bolded information from steps 2 and 3–4 in Figure 9.10; that is, the regression coefficients (b) and their standard errors. The results from the online calculator (www.quantpsy.org/sobel/sobel.htm) are shown in Figure 9.11. The Sobel test and two alternatives are shown, also explained on the website. All suggest that the indirect effect, although small ($a^*b = .07$ standardized) is statistically significant.

Input:	Test statistic:	Std. Error:	p-value:
a .047	Sobel test: 4.5752105	0.01608713	0.00000476
b 1.566	Aroian test: 4.55558121	0.01615645	0.00000522
s _a .005	Goodman test: 4.59509573	0.01601751	0.00000433
s _b .299	Reset all	Calculate	

Figure 9.11 Results of the Sobel test of the statistical significance of the indirect effect for the Motivation, Homework, and Achievement example. From the online calculator at <http://quantpsy.org/sobel/sobel.htm>.

Total effect of X on Y								
Effect	se	t	p	LLCI	ULCI	c_ps	c_cs	
.1230	.0354	3.4732	.0006	.0534	.1925	.0110	.1135	
Direct effect of X on Y								
Effect	se	t	p	LLCI	ULCI	c'_ps	c'_cs	
.0492	.0374	1.3171	.1883	-.0242	.1226	.0044	.0455	
Indirect effect(s) of X on Y:								
Effect	BootSE	BootLLCI	BootULCI					
HWork	.0737	.0160	.0438		.1064			
Normal theory test for indirect effect(s):								
Effect	se	Z	p					
HWork	.0737	.0161	4.5921		.0000			
Partially standardized indirect effect(s) of X on Y:								
Effect	BootSE	BootLLCI	BootULCI					
HWork	.0066	.0014	.0039		.0094			
Completely standardized indirect effect(s) of X on Y:								
Effect	BootSE	BootLLCI	BootULCI					
HWork	.0681	.0146	.0404		.0979			

Figure 9.12 Portion of the PROCESS bootstrapped analysis of the Motivation, Homework, Achievement model.

PROCESS

The PROCESS V3.0 SPSS macro (Hayes, 2018) was used to analyze these same data within SPSS. Ten thousand bootstrapped samples were used and 95% confidence intervals specified. Motivation was specified as X, Achievement as Y, and Homework as the mediator. Family Background and Ability were included in the model as covariates.

A portion of the PROCESS output is shown in Figure 9.12. For the rows showing the Total and Direct effects the unstandardized effects (b 's) are listed, with their standard errors, the t values, probabilities, the lower and upper limits of the confidences intervals for these effects (LLCI and ULCI, respectively), and the partially and completely standardized effects (e.g., c_{ps} and c_{cs}). Of more direct interest are the lines associated with the “Indirect effect(s) of X on Y.” These show the indirect effects (a^*b), the bootstrapped standard errors, and the bootstrapped confidence intervals. For this analysis, the 95% CI for the indirect effect of Motivation on Achievement was .044–.106. The CI does not include a value of zero, meaning that the indirect effect—the mediation of Motivation on Achievement by Homework—was statistically significant. The standardized indirect effect is shown at the bottom of the figure (.0681). Note that PROCESS will also conduct the Sobel test, and this is shown in the lines labeled “Normal theory test for indirect effect(s).”

In this example, all four methods used suggest Homework mediates the effect of Motivation on Achievement. In other words, if these data were real, the results would suggest that students with higher levels of academic motivation spend more time on homework, and that homework time, in turn, improves achievement. The different methods agree in this example, but they do not always do so. For this example, even though the indirect effect was relatively small (standardized indirect and total effects = .07 and .11, respectively), it was statistically significant across methods. This finding is, in part, a result of the relatively large sample size ($N = 579$) for a test of mediation. Fritz and MacKinnon (2007) reported a median sample size of 187 in their survey of 166 articles testing for mediation in two applied psychology journals 2000–2003. As already noted, the different methods vary in their power to detect mediation, with the bias-corrected bootstrap method used by PROCESS most likely to detect a true effect (Fritz & MacKinnon, 2007).

The example used here was fairly simple, with a single mediator. It is possible to test for multiple mediators or to add another level of mediators. For example, we might speculate that screen time also mediates the effect of motivation on achievement, and thus have two potential mediators. We might wonder how, in more detail, motivation affects homework. Perhaps we might speculate that more motivated students take a more academic mix of courses in high school, and those courses, in turn, assign more homework. If we had measures of coursework we could then test a model that had the mediation chain of Motivation → Coursework → Homework → Achievement. As already noted, it is also possible to have both mediation and moderation. Perhaps, for example, Homework time mediates the effect of Motivation on Achievement for girls, but not for boys. The PROCESS macro can test all these possibilities, and such an analysis is well explained in Hayes's book (2018). The PROCESS macro can also test for simple and complex moderation. For a review of recent developments in mediation analysis, see Preacher (2015).

I began this section by noting that mediation and indirect effect mean the same thing, and I often use the terms interchangeably. Not everyone agrees; some suggest that the term indirect effect is broader, and that the term mediation should be used only with longitudinal designs (e.g., Kline, 2016, pp. 134–135). If you subscribe to this notion, mediation is a subset of indirect effects. I am less sure. As we will discuss in more depth in Part 2, longitudinal data can certainly boost our confidence in time precedence, a requirement for a valid inference of causality. It is also the case that cross-sectional mediational designs often over- or underestimate true effects (Maxwell & Cole, 2007). Nevertheless, it seems to me that we go through the same process to establish indirect effects and mediation, and we generally interpret them in the same way. In addition, establishing causal ordering should be, in my opinion, based more on theory, logic, and previous research than on statistics (which instead tell us the implications of our causal inferences) (cf. Hayes, 2018, chap. 1). Thus I am looser in my verbiage. But we will discuss the issue of causal inference and its requirements in more detail in Part 2. As noted by Hayes, "causality is the cinnamon bun of social science"; it is desired (whether we admit it or not) and delicious, but quite sticky (2018, p. 18).

COMMON CAUSE

A common cause is a variable that affects both a presumed influence and its presumed outcome. If both the coefficients represented by paths a and c in Figure 9.5 are substantial and statistically significant, the variable Influence is a common cause of the Mediator and the Outcome. Common causes may also be referred to as confounding variables, or the "third variable problem." As noted in Chapter 4, important common causes *must* be included in a regression for the regression coefficient to provide valid estimates of the effect of one variable on another. If an important common cause is neglected in an analysis, the regression

Common cause 1

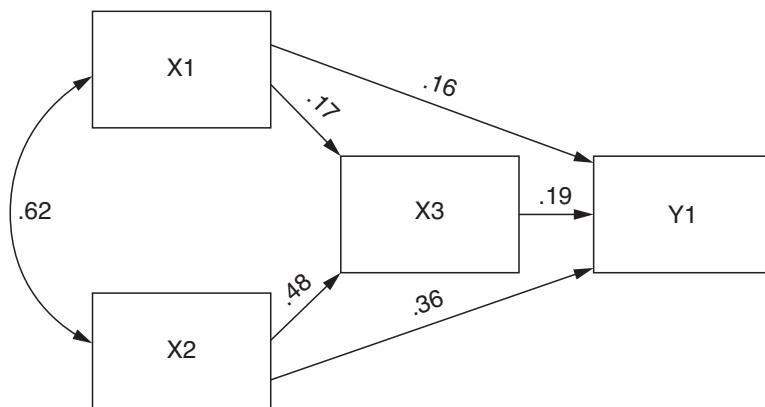


Figure 9.13 Common cause example 1. Variable X2 is a common cause of X3 and Y1.

coefficients will be misleading estimates of the effects of one variable on another. Analyses in which such common causes are not included are often referred to as misspecified analyses or models.

Figure 9.13 shows a model with three influences on an outcome. Some of the data used for this example were originally presented in Exercise 6 in Chapter 4. Consider this the “true” model; that is, variables X1, X2, and X3 have the effects on Y1 as shown in this figure. As can be seen in the figure, variable X2 is indeed a common cause of variable X3 and Y1; it has meaningful and, as we shall see, statistically significant effects on both X3 and Y1. Note that variable X1 is also a common cause of X3 and Y1, but subsequent examples will focus on changes to variable X2 and how these changes affect the estimate from X3 to Y1, so we will do the same here. Our primary focus will be on what happens to the β (standardized path coefficient) from X3 to Y1 as variable X2 is or is not a common cause of variables X3 and Y1.

Figure 9.14 shows the regression results for this first model. The first regression, at the top of the figure, shows the regression of Y1 on X1, X2, and X3. This regression is used to estimate the paths to variable Y1 in Figure 9.13; note that the coefficients match (in this case, the standardized coefficients, β) within errors of rounding. Also note the bolded portion of the output, with the β from variable X3 equal to .187 ($b = .188$, $SE_b = .045$, $t = 4.20$, $p < .001$). The bottom of Figure 9.14 shows the coefficients with X2 mistakenly not included in the regression. Note that the coefficient associated with X3 changes from .187 in the top part of the figure to .321 in the lower portion. Simply put, variable X2 is a common cause of X3 and Y1. When it was excluded from the model we got inaccurate estimates of the effects of X3 on Y1.

Now focus on Figure 9.15. For this model X2 is no longer a common cause of X3 and Y1 because when the other variables are taken into account X2 has an effect of zero on Y1. Once again, the top half of Figure 9.16 shows the regression of Y1 on the three variables pointing to it in Figure 9.15, and once again the coefficients match. The lower portion of Figure 9.16 shows the results of a regression that does not include variable X2. In this example, however, the coefficient for X3 to Y1 does not change; $\beta = .319$ in the top half of the figure and $\beta = .319$ in the bottom half of the figure. The coefficient associated with the path from X3 to Y1 does not change because X2 is not a common cause of X3 and Y1 because X2 does not affect X3.

Before moving to the next example note that Figure 9.15 shows an example of variable X3 completely mediating the effect of variable X2 on Y1. Thus, while exclusion or inclusion of X2 in the model does not change the apparent effect of variable X3, with variable X2 in the

Coefficients: Common Cause Example 1^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.020	.037		-.553	.580	-.093	.052
X1	.132	.039	.156	3.373	.001	.055	.208
X2	.260	.036	.362	7.175	.000	.189	.332
X3	.188	.045	.187	4.200	.000	.100	.276

a. Dependent Variable: Y1

Coefficients: Variable X2 not in the Model^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.035	.039		-.902	.367	-.111	.041
X1	.269	.036	.320	7.567	.000	.199	.339
X3	.323	.043	.321	7.583	.000	.239	.406

a. Dependent Variable: Y1

Figure 9.14 Regression results for common cause example 1. Variable X2 is a common cause of X3 and Y1. Neglecting it in the regression changes the estimate of the effect of X3.

Common cause 2

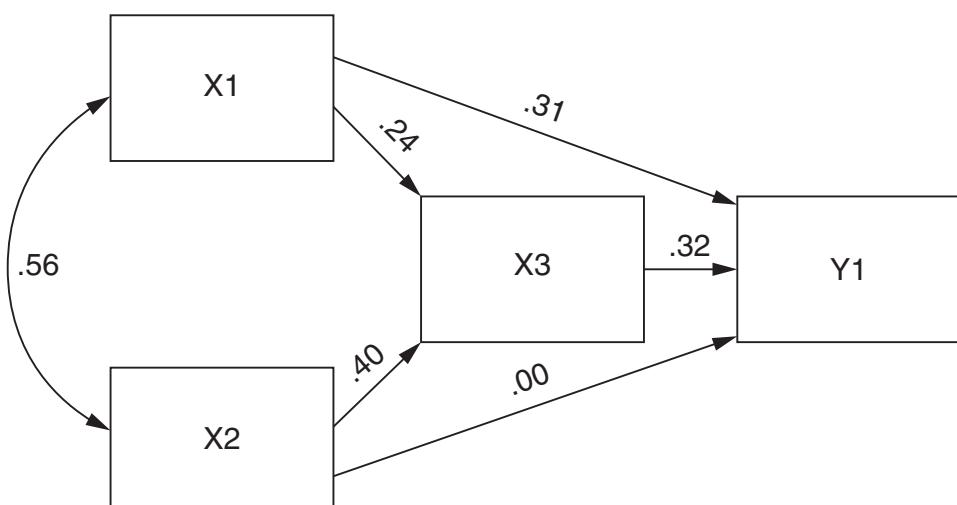


Figure 9.15 Common cause example 2. Variable X2 affects X3 but not Y1. It is not a common cause.

Coefficients: Common Cause Example 2^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.022	.039		-.553	.580	-.099	.055
X1	.265	.039	.315	6.713	.000	.188	.343
X2	-.001	.041	-.001	-.024	.981	-.081	.079
X3	.322	.046	.319	6.939	.000	.230	.413

a Dependent Variable: Y1

Coefficients: Variable X2 not in the Model^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.022	.039		-.553	.580	-.098	.055
X1	.265	.036	.314	7.389	.000	.194	.335
X3	.321	.043	.319	7.488	.000	.237	.405

a Dependent Variable: Y1

Figure 9.16 Regression results for common cause example 2. Here, X2 is not a common cause of X3 and Y1. The estimate of the effect of X3 does not change when X2 is not included in the regression.

Common cause 3

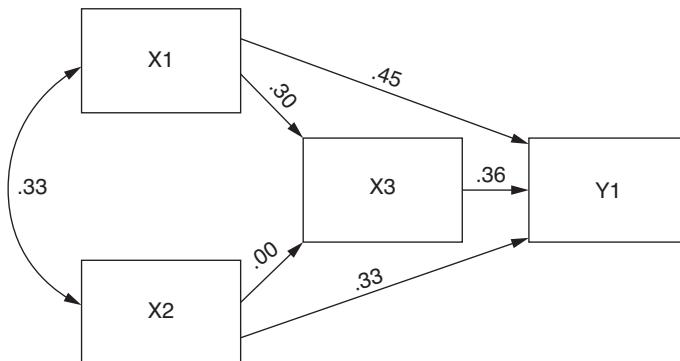


Figure 9.17 Common cause example 3. Variable X2 affects Y1 but not X3. Again, it is not a common cause.

model we have a more complete understanding of how these variables work. That is, variable X2 affects variable X3, which, in turn, affects variable Y1.

Finally, focus on Figure 9.17. Here, also, X2 is not a common cause of X3 and Y1. It affects Y1 but not X3. These effects are also shown in the first two portions of Figure 9.18: X2 has a

Coefficients: Common Cause Example 3^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.015	.027		-.553	.580	-.069	.038
X1	.465	.029	.454	15.811	.000	.407	.522
X2	.346	.029	.328	11.918	.000	.289	.403
X3	.378	.028	.364	13.378	.000	.322	.434

a. Dependent Variable: Y1

Coefficients: Paths to Variable X3^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.017	.043		-.383	.702	-.101	.068
X1	.292	.045	.296	6.514	.000	.204	.380
X2	-.001	.046	-.001	-.026	.980	-.092	.089

a. Dependent Variable: X3

Coefficients: Variable X2 not in the Model^a

Model	Unstandardized Coefficients			t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.037	.031		-1.209	.227	-.098	.023
X1	.576	.032	.562	18.246	.000	.514	.638
X3	.378	.032	.364	11.795	.000	.315	.441

a. Dependent Variable: Y1

Figure 9.18 Regression results for common cause example 3. Again, in this example X2 is not a common cause of X3 and Y1. The estimate of the effect of X3 does not change when X2 is not included in the regression.

standardized effect of .36 on Y1 ($b = .378$, $SE_b = .028$, $p < .001$), and X2 has an effect of zero on X3 ($b = -.001$, $SE_b = .046$, $p = .980$, $\beta = -.001$). And as shown in the bottom portion of the figure, if variable X2 were not included in the regression, the estimate of the effect of X3 on Y1 would be the same as it was in the first regression (and in Figure 9.17). A common belief is that we must include all background variables in a regression that affect an outcome if we are to get accurate estimates of the effect of our presumed cause on our presumed effect. This example shows that assumption to be incorrect. We need to include *common* causes of our presumed cause and presumed effect, not *all* causes of the presumed effect.

In my experience, it is not unusual for people to confuse common causes with moderation. When faced with a problem like that illustrated in Figure 9.19, sometimes a budding researcher will (often vaguely) posit that the Group variable “interacts” with the other

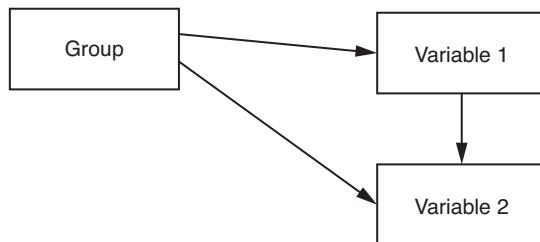


Figure 9.19 This example is an illustration of a common cause, not moderation.

variables in their effects. But this model does not illustrate interaction (moderation). If Group indeed affects both Variable 1 and Variable 2, it is a common cause. To use ANOVA-type lingo, Figure 9.19 illustrates Group having main effects on both Variable 1 and Variable 2, whereas moderation is analogous to an interaction in ANOVA.

It is also not unusual to refer to what I call “common causes” as confounding, as in “Variable 1 is confounded with Group membership” in trying to explain Variable 2. This is an accurate way to describing the model shown. But this term is also often used in a vague manner, in my experience, to mean situations like Figure 9.17, or even moderation. We will return to this discussion of common causes in Part 2; it is important in understanding the limits of nonexperimental research.

A FURTHER COMMENT ON LANGUAGE

It is not uncommon, at least in my reading, to see researchers say they are testing whether a variable mediates the relation (or relationship) between one variable and another. This use of the word relation or relationship, rather than effect or influence, no doubt grows out of a desire to avoid causal language in nonexperimental research. But if you examine the language more closely, it really is messy and adds confusion rather than clarity. Mediation is based on and requires causal thinking (Kenny, 2008, p. 355). Mediation is all about one variable affecting another via an intermediary variable. It is difficult, if not impossible, to conceive of mediation without an effect. In contrast, the use of the word relationship muddies the underlying thinking. Which is the presumed influence, and which is the effect? Relationship implies a correlation, an agnosticism about which variable we think of as the influence and which we think of as the effect. It implies that either variable could be the assumed influence and either the effect, when, by conducting a regression of one variable on others, we have implicitly assumed that one variable is, in some sense of the word, affected by others. *Mediation thinking requires causal thinking!* Non-causal language confuses mediation thinking. I encourage you to say that you are testing whether a variable mediates the effect (or presumed effect) of one variable on another, for example, whether homework mediates the effect of motivation on learning.

I believe that moderation thinking also requires causal thinking and is made clearer through the use of causal language. Suppose I say that I am testing whether sex moderates the relation between motivation and achievement. If the word “relation” implies correlation, what does this mean? Is it even possible? It probably is possible (e.g., different correlations for different groups) but usually means something different from what is intended by this imprecise wording. Again, we are really interested in whether sex moderates the presumed

effect of motivation on achievement; whether sex and motivation interact in their effect on achievement. The use of non-causal language confuses rather than illuminates and should, in my opinion, be avoided. Moderation thinking requires causal thinking, and avoiding causal language leads to confusion rather than clarity. We should not pretend that we are uninterested in the potential effect of one variable on another if, in fact, that is exactly what we are interested in (cf. Hayes, 2018, chap. 1).

Perhaps my cavalier use of causal language makes you feel uncomfortable. Or perhaps you agree with me but respond “I see your point, but my advisor (or my reviewers) won’t let me use that kind of language; they insist on ‘relation’ rather than ‘influence’ or ‘effect.’” One possible compromise is to add a paragraph explaining what we actually mean by such language when it is first introduced. Something like:

It is important to note that the data used in this research are nonexperimental in nature; there will be no (nor could there be) experimental manipulation of academic motivation to determine its subsequent effect on achievement. As a result, it should be understood that all statements that discuss the “effect” of one variable on another, or that focus on variables that “explain” an outcome are dependent on the validity of the regression model. In other words, if the model is a reasonable representation of reality, the estimates resulting from the model indeed show the extent of the influence of one variable on another. If the model is not a reasonable representation of reality, the estimates are not accurate estimates of those effects.

Once again, we will explore these topics, and especially what makes a model valid, more completely in Part 2 of this text.

One final note: when we really are talking about a non-causal relation, which is correct, “relation” or “relationship?” One of my early research texts was the excellent *Foundations of Behavioral Research* by Fred Kerlinger (1986). Kerlinger argued that people have relationships, whereas variables have relations (p. 58n). That explanation has always made sense to me.

SUMMARY

In this chapter we sought to clarify the concepts of moderation, mediation, and common cause. In my experience, these three concepts are often confused, so it is worthwhile to spend a little time making sure you understand the differences. Moderation is another term for interaction, and is used to describe the situation where the magnitude of the effect of one variable on another depends on a third variable. The use of the language “it depends” is a hint that we are likely talking about moderation. Chapters 7 and 8 focused on moderation with categorical (Chapter 7) and continuous (Chapter 8) variables.

Mediation is another term for indirect effect, and is useful for describing *how* an effect comes about. How does motivation affect achievement? Perhaps more motivated students complete more homework, and it is that homework, in turn, that raises achievement. Here, we have posited that homework may mediate the effect of motivation on achievement. Unlike common causes, mediators need not be included in a model to get accurate estimates of effects, but they are very interesting because they help explain how an effect may work. The chapter also explained how to test the statistical significance of a mediated effect. We first introduced the topic of indirect effects/mediation in Chapter 4, and will return to it in Part 2. There we will see how to test for mediation using structural equation modeling programs.

Common causes are variables that affect both your presumed cause and your presumed effect. They must be included in a regression or path model for the regression coefficients to provide accurate estimates of effects. The topic of common causes has been touched on

throughout the text so far, and will continue to be a topic of interest as we continue. They were discussed in this chapter to make sure that you understand how they differ from moderation and mediation. We also analyzed an example from the Chapter 4 exercises to show the effect of not including a common cause versus a non-common cause in a regression analysis. This example showed that a regression coefficient of interest is misleading when a common cause is excluded from the model, but stable when a non-common cause is excluded.

EXERCISES

1. Use a literature research database to find an article in an area of interest to you with the word *mediation* in either the title or the abstract. Read the abstract to make sure the term mediation refers to statistical mediation (rather than, say, legal mediation). Read the article. Do the authors also refer to mediation as an indirect effect? How do they test for mediation? Do they use any of the methods described in this chapter, or do they use structural equation modeling? Do you understand the test that was used?
2. Search for an article in your area of interest with the words *moderation* or *moderated regression* in the title or abstract. Read the abstract to make sure regression was used. Read the article. Is moderation also referred to as an interaction? Which variables interact? Were they continuous or categorical? Did the authors use techniques like those described in Chapters 7 and 8? Were the variables of interest centered prior to creating a cross product? Was the article understandable in light of the last three chapters?

10

Multiple Regression Summary, Assumptions, Diagnostics, Power, and Problems

Summary	195
“Standard” Multiple Regression	195
Explanation and Prediction	197
Three Types of Multiple Regression	198
Categorical Variables in MR	199
Categorical and Continuous Variables, Interactions, and Curves	199
Moderation, Mediation, and Common Cause	200
Assumptions and Regression Diagnostics	201
Assumptions Underlying Regression	201
Regression Diagnostics	202
Sample Size and Power	216
Problems with MR?	221
Exercises	225
Note	225

You should now have a reasonably complete, conceptual understanding of the basics of multiple regression analysis. This chapter will begin by summarizing the topics covered in Part 1. I will touch on some issues that you should investigate and understand more completely to become a sophisticated user of MR and will close the chapter with some nagging problems and inconsistencies that we have discussed off and on throughout Part 1 (and will try to resolve in Part 2).

SUMMARY

“Standard” Multiple Regression

For social scientists raised on statistical analyses appropriate for the analysis of experiments (ANOVA and its variations), multiple regression often seems like a different animal altogether. It is not. MR provides a close implementation of the general linear model, of which ANOVA is a part. In fact, MR subsumes ANOVA, and as shown in several places in this portion of the book, we can easily analyze experiments (ANOVA-type problems) using MR. The reverse is not the case, however, because MR can handle both categorical and continuous independent variables, whereas ANOVA requires categorical independent variables.

Those with such an experimental background may need to change their thinking about the nature of their analyses, but the underlying statistics are not fundamentally different. In my experience, this transition to MR tends to be more difficult for those with a background in psychology or education; in other social sciences, such as sociology and political science, experimentation (i.e., random assignment to treatment groups) is less common. Even in psychology and education the trend increasingly appears to be to focus on the general linear model, and multiple regression, early in students' research training, so the sometimes-difficult transition I mention here may not apply to you.

In early chapters we covered how to calculate the fundamental statistics associated with multiple regression. More practically, we discussed how to conduct, understand, and interpret MR using statistical analysis programs. R is the multiple correlation coefficient, and R^2 the squared multiple correlation. R^2 is an estimate of the variance explained in the dependent variable by all the multiple independent variables in combination; an R^2 of .2 means that the independent variables jointly explain 20% of the variance in the dependent variable. In applied social science research, R^2 's are often less than .5 (50% of the variance explained), unless some sort of pretest is included as a predictor of some posttest outcome, and R^2 's of .10 are not uncommon. A high R^2 does not necessarily mean a good model; it depends on the dependent variable to be explained. R^2 may be tested for statistical significance by comparing the variance explained (regression) to the variance unexplained (residual) using an F table, with degrees of freedom equal to the number of independent variables (k) and the sample size minus this number, minus 1 ($N_c - k - 1$).

R^2 provides information about the regression as a whole. The MR also produces information about each independent variable alone, controlling for the other variables in the model. The unstandardized regression coefficients, generally symbolized as b (or sometimes as B), are in the original metric of the variables used, and the b can provide an estimate of the likely change in the dependent variable for each 1-unit change in the independent variable (controlling for the other variables in the regression). For example, Salary, in thousands of dollars a year, may be regressed on Educational Attainment, in years, along with several other variables. If the b associated with Educational Attainment is 3.5, this means that for each additional year of schooling salary would increase, on average, by 3.5 thousand dollars per year. The b is equal to the slope of the regression line. The b 's may also be tested for statistical significance using a simple t test ($t = \frac{b}{SE_b}$), with the df equal to the df residual for the overall F test. This t simply tests whether the regression coefficient is statistically significantly different from zero. More interestingly, it is also possible to determine whether the b differs from values other than zero, either using a modification of the t test or by calculating the 95% (or 90%, or some other level) confidence interval around the b 's. Suppose, for example, that previous research suggests that the effect of Educational Attainment on Salary is 5.8. If the 95% CI around our present estimate is 2.6 to 4.4, this means that our present estimate is statistically significantly lower than are estimates from previous research. The use of confidence intervals is increasingly required by journals (see, for example, American Psychological Association, 2010).

We can also examine the standardized regression coefficients associated with each independent variable, generally symbolized as β . β 's are in standard deviation units, thus allowing the comparison of coefficients that have different scales. A β of .30 for the effect of Educational Attainment on Salary would be interpreted as meaning that each standard deviation increase in Educational Attainment should result in a .30 SD average increase in Salary.

The standardized and unstandardized regression coefficients serve different purposes and have different advantages. Unstandardized coefficients are useful when the scales of the independent and dependent variables are meaningful, when comparing results across samples and studies, when we wish to develop policy implications or interventions from our

research, and when interpreting the results of interaction (moderation) analyses. Unstandardized coefficients are also the coefficients that are tested for statistical significance. Standardized coefficients are useful when the scales of the variables used in the regression are not meaningful or when we wish to compare the relative importance of variables in the same regression equation.

The regression analysis also produces an intercept or constant. The intercept represents the predicted score on the dependent variable when all the independent variables have a value of zero. The regression coefficients and the intercept can be combined into a regression equation (e.g., $Y_{predicted} = \text{intercept} + b_1X_1 + b_2X_2 + b_3X_3$), which can be used to predict someone's score on the outcome from the independent variables.

The regression equation, in essence, creates an optimally weighted composite of the independent variables to predict the outcome variable. This composite is weighted so as to maximize the prediction and minimize the errors of prediction. We can graph this prediction by plotting the outcome (Y -axis) against the predicted outcome (X -axis). The spread of data points around the regression line illustrates the accuracy of prediction and the errors of prediction. Errors of prediction are also known as residuals and may be calculated as outcome scores minus predicted outcome scores. The residuals may also be considered as the outcome variable with the effects of the independent variables statistically removed.

Explanation and Prediction

MR may serve a variety of purposes, but these generally fall under one of two broad categories: prediction or explanation. If our primary interest is in explanation, then we are interested in using MR to estimate the effects or influences of the independent variables on the dependent variable. Underlying this purpose, whether we admit it or not, is an interest in cause and effect. To estimate such effects validly, we need to choose carefully the variables included in the regression equation; it is particularly important that we include any common causes of our presumed cause and presumed effect. An understanding of relevant theory and previous research can help one choose variables wisely. Throughout this text, I have assumed that in most instances we are interested in using MR in the service of explanation, and most of the examples have had an explanatory focus.

In contrast, MR may also be used for the general purpose of prediction. If prediction is our goal, we are not necessarily interested in making statements about the effect of one variable on another; rather, we only want to be as accurate as possible in predicting some outcome. A predictive purpose is often related to selection; a college may be interested in predicting students' first-year GPAs as an aid in determining which students should be admitted. If prediction is the goal, the larger the R^2 the better. One does not need to worry about common causes, or even cause and effect, if one's interest is in prediction, and thus variable selection for prediction is less critical. It may even be perfectly acceptable to have an "effect" predicting a "cause" if prediction is the goal. Theory and previous research can certainly help you choose the variables that will predict your outcome successfully, but they are not critical to the interpretation of your findings as they are when MR is used for explanation. If your interest is in prediction, however, you must refrain from making statements or coming to conclusions about the effects of one variable on another (an explanatory purpose). It is unfortunately common to see research in which the purpose is supposedly prediction, but then when you read the discussion you find explanatory (causal) conclusions are being made. Any time you wish to use MR to make recommendations for intervention or change (if we increase X , Y will increase), your primary interest is in explanation, not prediction. Explanation subsumes prediction. If you can explain a phenomenon well, then you can generally predict it well. The reverse does not hold, however; being able to predict something does not mean you can explain it.

Three Types of Multiple Regression

There are several types, or varieties, of multiple regression. The type of MR used in the earlier chapters of this book is generally referred to as simultaneous, or forced entry, or standard multiple regression. In *simultaneous regression*, all independent variables are entered into the regression equation at the same time. The regression coefficients and their statistical significance are used to make inferences about the importance and relative importance of each variable. Simultaneous regression is useful for explanation or prediction. When used in an explanatory context, the regression coefficients from simultaneous regression provide estimates of the direct effects of each independent variable on the outcome (taking the other independent variables into account); this is one of this method's major advantages. Its chief disadvantage is that the regression coefficients may change depending on which variables are included in the regression equation; this disadvantage is related to the exclusion of relevant common causes or the presence of intervening or mediating variables.

In sequential, or hierarchical, regression, each variable [or group or block of variables] is entered separately into the regression equation, sequentially, in an order determined by the researcher. With *sequential regression*, we generally focus on ΔR^2 from each step to judge the statistical significance of each independent variable. ΔR^2 is a stingy and misleading estimate of the *importance* of variables, however; the square root of ΔR^2 provides a better estimate of the importance of each variable (*given* the order of entry). Order of entry is critical with sequential regression because variables entered early in the sequential regression will appear, other things being equal, more important than variables entered later. Time precedence and presumed causal ordering are common methods for deciding the order of entry. The regression coefficients for each variable from the block in which it enters a sequential regression may be interpreted as the *total* effect of the variable on the outcome, including any indirect or mediating effects through variables entered later in the regression. To interpret sequential regression results in this fashion, variables must be entered in their correct causal order. Causal, or path, models are useful for both sequential and simultaneous regression and have been used to illustrate regression models and results throughout Part 1 of this text; they will be explored in more depth in Part 2. Sequential regression may be used for explanation or prediction. An advantage is that it can provide estimates of the total effects of one variable on another, given the correct order of entry. A chief disadvantage is that the apparent importance of variables changes depending on the order in which they are entered in the sequential regression equation.

Simultaneous and sequential regression may be combined in various ways. One combination is a method sometimes referred to as *sequential unique regression*. It is commonly used to determine the "unique" variance accounted for by a variable or a group of variables, after other relevant variables are accounted for. In this method, the other variables are entered in a simultaneous block, and a variable or variables of interest are entered sequentially in a second block. If a single variable is of interest, simultaneous regression may be used for the same purpose; if the interest is in the variance accounted for by a block of variables, this combination of simultaneous and sequential regression should be used. We made extensive use of this sort of combination of methods when we tested for interactions and curves in the regression line.

A final general method of multiple regression is stepwise regression and its variations. *Stepwise regression* operates in a similar fashion to sequential regression, except that the computer program, rather than the researcher, chooses the order of entry of the variables; it does so based on which variable will lead to the greatest single increment in ΔR^2 at each step. Although this solution seems a blessing—it avoids lots of hard thinking and potentially embarrassing statements about causal ordering—it is not. Using ΔR^2 or $\sqrt{\Delta R^2}$ as a measure

of the importance of variables is predicated on the assumption that the variables have been entered in the regression equation in the proper order. To also use ΔR^2 to determine the order of entry thus requires circular reasoning. For this reason, stepwise methods should be used only for prediction, not explanation. In the words of my friend Lee Wolfe, stepwise regression is “theoretical garbage” (1980, p. 206), meaning that its results will mislead rather than inform if you try to use it in explanatory research. And, in fact, stepwise regression may not be a particularly good choice even for prediction. If your interest is simply selecting a subset of variables for efficient prediction, stepwise regression may work (although I still wouldn’t recommend it); large samples and cross-validation are recommended. Whatever method of MR you use, be sure you are clear on the primary purpose of your research and choose your regression method to fulfill that purpose.

Categorical Variables in MR

It is relatively easy to analyze categorical, or nominal, variables in multiple regression. One of the easiest ways is to convert the categorical variable into one or more *dummy variables*. With dummy variables, a person is assigned a score of 1 or 0, depending on whether the person is a member of a group or not a member. For example, the categorical variable sex can be coded so that males are scored 0 and females 1, essentially turning it into a “female” variable on which those who are members of the group (females) receive a score of 1 and those who are not members (males) receive a score of 0. For more complex categorical variables, multiple dummy codes are required. We need to create as many dummy variables as there are categories, minus 1 ($g - 1$). When a categorical variable has more than two categories, thus requiring more than one dummy variable, one group will be scored 0 on all the dummy variables; this is essentially the reference group, or often the control group. When dummy variables are analyzed in MR, the intercept is equal to the mean score on the dependent variable for the reference group, and the b 's are equal to the mean deviations from that group for each of the other groups.

We demonstrated that MR results match those of ANOVA when the independent variables are all categorical: the F from the two procedures is the same, and the effect size η^2 from ANOVA is equal to the R^2 from MR. The coefficients from MR may be used to perform various post hoc procedures. There are other methods besides dummy coding for coding categorical variables for analysis in MR; we illustrated effect coding and criterion scaling. The different methods will provide the same overall results, but different contrasts from the regression coefficients.

Categorical and Continuous Variables, Interactions, and Curves

Our primary interest in discussing the analysis of categorical variables in MR was as preparation for combining categorical and continuous variables together in MR analyses. Analyses including both categorical and continuous variables are conceptually and analytically little different from those including only continuous variables. It is also possible to test for interactions between categorical and continuous variables. To do so, we centered the continuous variable and created a new variable that was the cross product of the dummy variable and the centered continuous variable. If there are multiple dummy variables, then there will also be multiple cross products. These cross products are then entered as the second, sequential step in a regression following the simultaneous regression with all other independent variables (including the categorical and continuous variables used to create the cross products). The statistical significance of the ΔR^2 associated with the cross products is the test of the statistical significance of the interaction. With multiple dummy variables, and thus multiple

cross products, the ΔR^2 associated with the *block* of cross products is used to determine the statistical significance of the interaction.

Given the presence of a statistically significant interaction, the next step is to graph the interaction to provide an understanding of its nature, perhaps followed by additional regressions across the values of the categorical variable or other post hoc probing. Tests of predictive bias and attribute-treatment interactions are specific examples of analyses that should use this MR approach. ANCOVA can also be considered as MR with categorical and continuous variables, but researchers using MR can also test for possible interactions between the covariate and the treatment.

It is equally possible to test for interactions between two continuous variables in MR. The same basic procedure is used: the continuous variables are centered and multiplied, and this cross product is entered sequentially in a regression equation. Follow-up of this type of interaction may be a little more difficult, but the first step again is generally to graph the interaction. Several methods were discussed for graphing and exploring interactions between continuous variables. All types of interactions are often well described using the phrase “it depends.”

A special type of interaction between continuous variables is when a variable interacts with itself, meaning that its effects depend on the *level* of the variable. For example, we found that the effect of homework depends on the amount of homework being discussed; homework has a stronger effect on achievement for fewer hours of homework than for higher levels of homework. This type of interaction shows up as curves in the regression line. We test for curves in the regression line by multiplying a variable times itself and then entering this squared variable last in a combined simultaneous-sequential regression. We can test for more than one curve by entering additional product terms (variable-cubed, to the fourth power, etc.). Again, graphs were recommended as a method for understanding the nature of these curvilinear effects.

Moderation, Mediation, and Common Cause

Interactions in multiple regression also go by the name of “moderation.” To say that sex moderates the effect of self-concept on achievement means the same thing as saying that sex and self-concept interact in their effect on achievement, or that self-concept has differential effects on achievement by sex. Why do we use different terms to mean what is essentially the same thing? Thompson’s contention that we do so to “confuse the graduate students” seems as plausible as any other (Thompson, 2006, p. 4). The term moderation is sometimes confused with that of mediation. Mediation describes the process by which one variable has an indirect effect on another variable through another mediating variable. If homework mediates the effect of motivation on achievement, this means that motivation affects homework, which in turn affects achievement. In Chapter 9 we discussed several methods for testing for mediation in multiple regression, but also noted that it is often easier to understand and test for mediation in the context of path analysis and SEM (as in Part 2). Indeed, we used path diagrams extensively to illustrate mediation. Although I tend to use the terms “mediation” and “indirect effect” fairly interchangeably, others suggest that the term mediation should be reserved for analyses involving longitudinal data (e.g., Kline, 2016, chap. 6). Fewer writers discuss the issue of common cause (and there are also several terms used to discuss this concept). A common cause is a variable that affects both our presumed influence and our presumed outcome; such variables *must* be included in multiple regression for the results to provide valid estimates of “effects.” It is not unusual to see and hear this concept confused with that of moderation. When you hear researchers vaguely state that two variables likely interact in some way, pay attention. Do they really mean interaction/moderation? Or are they really talking about a potential common cause? Again, this is a topic that becomes clearer with the presentation of path diagrams (as used in Chapter 9) and is an important topic in Part 2 of this book.

ASSUMPTIONS AND REGRESSION DIAGNOSTICS

We have postponed discussion of several important topics until you had a more complete understanding of multiple regression and how to conduct and interpret results of multiple regression analyses. Now it is time to discuss assumptions underlying our multiple regressions, as well as how to diagnose various problems that can affect regression analyses and what to do about these problems. References are given to sources that provide more detail about these topics.

Assumptions Underlying Regression

What assumptions underlie our use of multiple regression? If we are to be able to trust our MR results and interpret the regression coefficients, we should be able to assume the following:

1. The dependent variable is a linear function of the independent variables.
2. Each person (or other observation) should be drawn independently from the population. Recall one general form of the regression equation: $Y = a + bX_1 + bX_2 + e$. This assumption means that the errors (e 's) for each person are independent from those of others.
3. The variance of the errors is not a function of any of the independent variables. The dispersion of values around the regression line should remain fairly constant for all values of X . This assumption is referred to as homoscedasticity.
4. The errors are normally distributed.

The first assumption (linearity) is the most important. If it is violated, then all of the estimates we get from regression— R^2 , the regression coefficients, standard errors, tests of statistical significance—may be biased. To say the estimates are biased means that they will likely not reproduce the true population values. When assumptions 2, 3, and 4 are violated, regression coefficients are unbiased, but standard errors, and thus significance tests, will not be accurate. In other words, violation of assumption 1 threatens the meaning of the parameters we estimate, whereas violation of the other assumptions threatens interpretations from these parameters (Darlington, 1990, p. 110). Assumptions 3 and 4 are less critical, because regression is fairly robust to their violation (Kline, 1998). The violation of assumption 4 is only serious with small samples. We have already discussed methods of dealing with one form of nonlinearity (curvilinearity, in Chapter 8) and will discuss here and later methods for detecting and dealing with violations of the other assumptions.

In addition to these basic assumptions, to interpret regression coefficients as the *effects* of the independent variables on the dependent variable, we need to be able to assume that the errors are uncorrelated with the independent variables. This assumption further implies the following:

5. The dependent variable does not influence any of the independent variables. In other words, the variables we think of as causes must in fact be the causes, and those that we think of as the effects must be the effects.
6. The independent variables are measured without error, with perfect reliability and validity.
7. The regression must include all common causes of the presumed cause and the presumed effect (Kenny, 1979, p. 51).

We have already discussed assumptions 5 and 7 and will continue to develop them further in Part 2. Assumption 6 is a concern, because in the social sciences we rarely have perfect measurement. Again, we will discuss the implications of violation of this assumption in Part 2. There are a number of very readable, more detailed explanations of these seven assumptions. Allison (1999), Berry (1993), and Cohen and colleagues (2003) are particularly useful.

Regression Diagnostics

Here and in earlier chapters I noted that a good habit in any data analysis is to examine the data to make sure the values are plausible and reasonable. Always, always, always check your data. Regression diagnostics take this examination to another level and can be used to probe violations of assumptions and spot impossible or improbable values and other problems with data. In this section I will briefly describe regression diagnostics, illustrate their use for the data from previous chapters, and discuss what to do with regression diagnostic results. I will emphasize a graphic approach.

Diagnosing Violations of Assumptions

Nonlinearity

In Chapter 8, we examined how to deal with nonlinear data by adding powers of the independent variable to the regression equation. In essence, by adding both Homework and Homework² to the regression equation, we turned the nonlinear portion of the regression line into a linear one and were thus able to model the curve effectively using MR.

This approach thus hints at one method for determining whether we have violated the assumption of linearity: If you have a substantive reason to suspect that an independent variable may be related to the outcome in a curvilinear fashion, add a curve component (variable²) to the regression equation to see whether this increases the explained variance.

The potential drawback to this approach is that the curve modeled by variable² may not adequately account for the departure from linearity. Therefore, it is useful to supplement this approach with a more in depth examination of the data using scatterplots. Rather than plotting the dependent variable of interest against the independent variable, however, we will plot the *residuals* against the independent variables; the residuals should magnify departures from linearity. Recall that the residuals represent the predicted values of the dependent variable minus the actual values of the dependent variable ($\hat{Y} - Y$). They are the errors in prediction.

To illustrate, we will use the example from Chapter 8 that was used to illustrate testing for curves in MR: the regression of Grades on SES, previous Achievement, and time spent on Homework out of school. The addition of a Homework² variable was statistically significant, indicating (and correcting) a departure from linearity in the regression. Let's see if we can pick up this nonlinearity using scatterplots.

I reran the initial regression (without the Homework² variable and using the original uncentered metric) and saved the residuals (regression programs generally allow you to save unstandardized residuals as an option). Figure 10.1 shows the plot of the residuals against the original variable Homework. Note the two lines in the graph. The straight, horizontal line is the mean of the residuals. The line should also represent the regression line of the residuals on Homework. That line would be horizontal because the residuals represent Grades with the effects of Homework (and the other independent variables) removed. Because Homework has been removed, it is no longer related to the residuals. Recall that when two variables are unrelated our best prediction for Y is its mean for all values of X . The regression line is thus equal to the line drawn through the mean of the residuals. The other, almost straight line is what is called a *lowess* (or *loess*) fit line, which represents the *nonparametric* best fitting

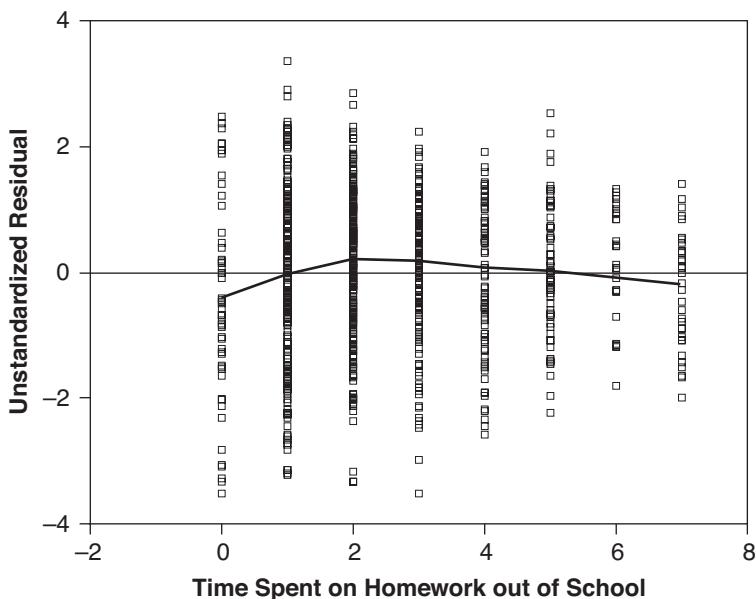


Figure 10.1 Plot of the unstandardized residuals against one independent variable (Homework). The lowess line is fairly straight.

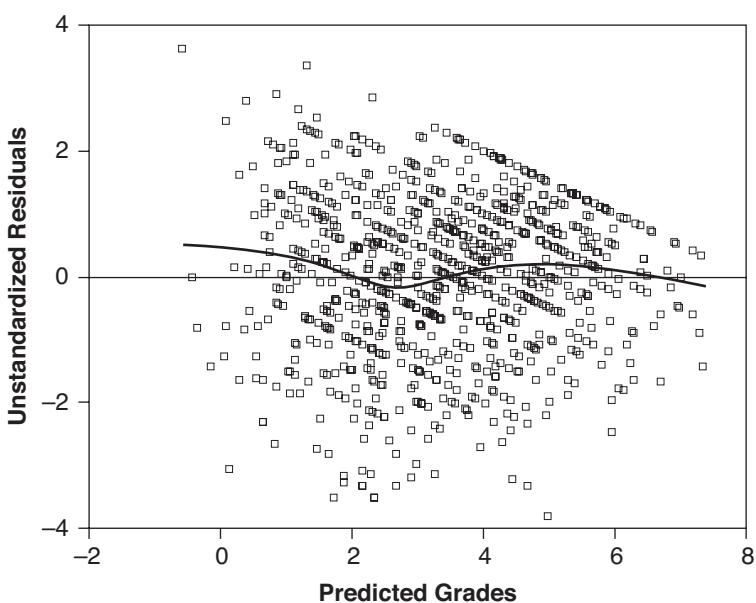


Figure 10.2 Plot of unstandardized residuals against the predicted Grades (a composite of the independent variables).

line, one that does not impose the requirement of linearity. Most computer programs can easily add this line to a regression scatterplot.

If there is no departure from linearity in the data, we would expect the lowess line to come close to the regression line; Cohen and colleagues note that the lowess line should look like “a young child’s freehand drawing of a straight line” (2003, p. 111). With a significant departure from linearity, you would expect the lowess line to be curved, something more similar to the curvilinear regression lines shown in Chapter 8

(e.g., Figure 8.10) but without the upward slope. The lowess line in this plot indeed approaches the straight regression line. Figure 10.2 shows another useful plot: the residuals and the predicted values for the Grades dependent variable. Recall in Chapter 3 that we demonstrated that the predicted Y is an optimally weighted composite of the independent variables. It is, then, a variable that represents all independent variables in combination. Again, the lowess line comes close to the regression line and does not suggest a departure from linearity.

In this example, the test of the addition of a curve component (Chapter 8) was more successful in spotting a departure from linearity than was the inspection of data through scatterplots. This will not always be the case, and thus I recommend that you use both methods if you suspect a violation of this assumption. If theory or inspection suggests a departure from linearity, a primary method of correction is to build nonlinear terms into the regression (e.g., powers, logarithms). The method is discussed in Chapter 8; see also Cohen and colleagues (2003) and Darlington and Hayes (2017) for more depth.

Nonindependence of Errors

When data are not drawn independently from the population, we risk violating the assumption that errors (residuals) will be independent. As noted in the section on multilevel modeling in the next chapter, the NELS data, with students clustered within schools, risks violation of this assumption. Violation of this assumption does not affect regression coefficients but does affect standard errors. When clustered as described, we risk underestimating standard errors and thus labeling variables as statistically significant when they are not. This danger is obviated, to some degree, with large samples like the NELS data used here, especially when we are more concerned with the magnitude of effects than with statistical significance.

Are the residuals from the regression of Grades on SES, Previous Achievement, and Homework nonindependent? Is there substantial variation within schools? Unfortunately, this assumption is difficult to test with the NELS data included on the Web site because, with the subsample of 1000 cases, few of the schools had more than one or two students. Therefore, I used the original NELS data and selected out 414 cases from 13 schools. I conducted a similar regression analysis (Grades on SES, Previous Achievement, and Homework) and saved the residuals.

One way to probe for the violation of this assumption is through a graphing technique called *boxplots*. The boxplots of residuals, clustered by schools, are shown in Figure 10.3. The center through each boxplot shows the median, with the box representing the middle 50% of cases (from the 25th to the 75th percentile). The extended lines show the high and low values, excluding outliers and extreme values. For the purpose of exploring the assumption of independence of errors, our interest is in the variability of the boxplots. There is some variability up and down by school, and thus this clustering may indeed be worth taking into account. Another, quantitative test of the independence of observations uses the intraclass correlation coefficient, which compares the between-group (in this case, between-schools) variance to the total variance (for an example, see Stapleton, 2006). The intraclass correlation could be computed on the residuals or on a variable (e.g., Homework) that you suspect might vary across schools.

One option for dealing with a lack of independence of errors is to include categorical variables (e.g., using criterion scaling; see Chapter 6) that take the clustering variable into account. Another option is the use of multilevel or hierarchical linear modeling, discussed briefly in the next chapter. This assumption can also be violated in longitudinal designs in which the same tests or scales are administered repeatedly. We will deal with this issue briefly in Part 2.

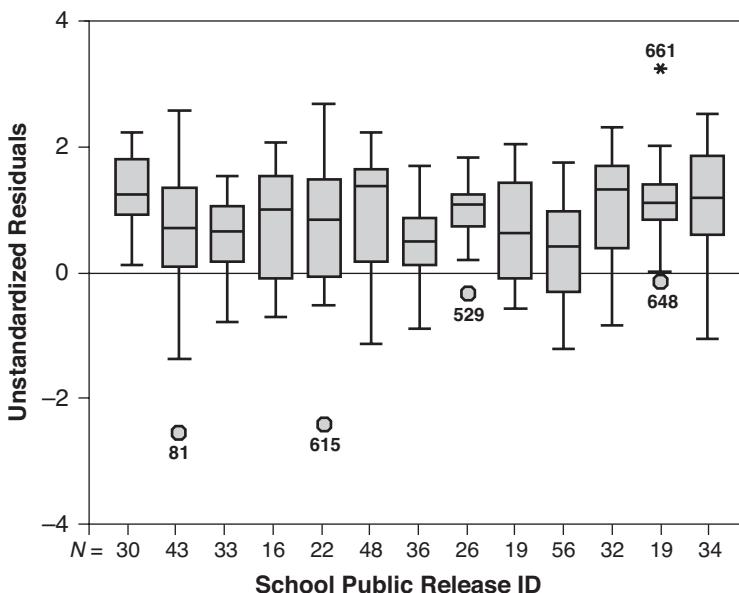


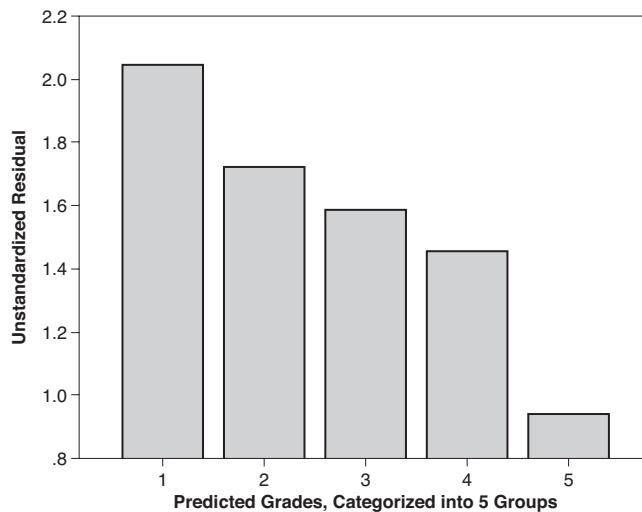
Figure 10.3 Boxplots of residuals, grouped by the school from which NELS students were sampled. The data are 414 cases from the full NELS data.

Homoscedasticity

We assume that the variance of errors around the regression line is fairly consistent across levels of the independent variable. In other words, the residuals should spread out consistently across levels of X . Violation of this assumption affects standard errors and thus statistical significance (not the regression coefficients), and regression is fairly robust to its violation. Scatterplots of residuals with independent variables or predicted values are also helpful for examining this assumption.

Return to Figure 10.1, the scatterplot of Homework with the Residuals from the regression of Grades on SES, Previous Achievement, and Homework. Although the residuals are spread out more at lower levels of homework than at upper levels, the difference is slight; visual inspection suggests that heteroscedasticity (the opposite of homoscedasticity) is not a problem. A common pattern of heteroscedasticity is a fan shape with, for example, little variability at lower levels of Homework and large variability at higher levels of Homework. Butterfly shapes are also possible (residuals constricted around the middle level of Homework), as is the opposite shape (a bulge in the middle).

Focus again on Figure 10.2. Notice how the residuals bunch up at higher levels of the Predicted Y ; the plot has something of a fan shape, narrowing at upper levels of the predicted values. Do these data violate the assumption of homoscedasticity? To test this possibility, I collapsed the predicted Grades into five equal categories so that we can compare the variance of the residuals at each of these five levels. The data are displayed in Figure 10.4 as both a bar chart and table. As shown in the table, for the lowest category of predicted values, the variance of the residuals was 2.047, versus .940 for the highest category. There is a difference, but it is not excessive. One rule of thumb is that a ratio of high to low variance of less than 10 is not problematic. Statistical tests are also possible (Cohen et al., 2003).



Report

RES_2 Unstandardized Residual

NPRE_2 predicted	Mean	N	Std. Deviation	Variance
1	.1813529	173	1.43089123	2.047
2	-.2252563	178	1.31232697	1.722
3	-.0820244	182	1.25877627	1.585
4	.0519313	182	1.20728288	1.458
5	.0784449	181	.96944000	.940
Total	.0000000	896	1.24815811	1.558

Figure 10.4 Comparison of the variance of residuals for different levels of predicted Grades.

Normality of Residuals

The final assumption we will deal with is that the errors, or residuals, are normally distributed. What we are saying with this assumption is that if we plot the values of the residuals they will approximate a normal curve. This assumption is fairly easily explored because most MR software has tools built in to allow such testing.

Figure 10.5 shows such a plot: a bar graph of the residuals from the NELS regression of Grades on SES, Previous Achievement, and Homework (this graph was produced as one of the plot options in regression in SPSS). The superimposed normal curve suggests that the residuals from this regression are indeed normal. Another, more exacting, method is what is known as a q-q plot (or, alternatively, a p-p plot) of the residuals. A q-q plot of the residuals shows the value of the residuals on one axis and the expected value (if they are normally distributed) of the residuals on the other. Figure 10.6 shows the q-q plot of the residuals from the Grades on SES, Previous Achievement, Homework regression. If the residuals are normally distributed, the thick line (expected versus actual residuals) should come close to the diagonal straight line. As can be seen from the graph, the residuals conform fairly well to the superimposed straight line. The reason this method is more exact is that it is easier to spot a deviation from a straight line than a normal curve (Cohen et al., 2003). Some programs (e.g., SPSS) produce a p-p plot of the residuals as an option in multiple regression. A p-p plot

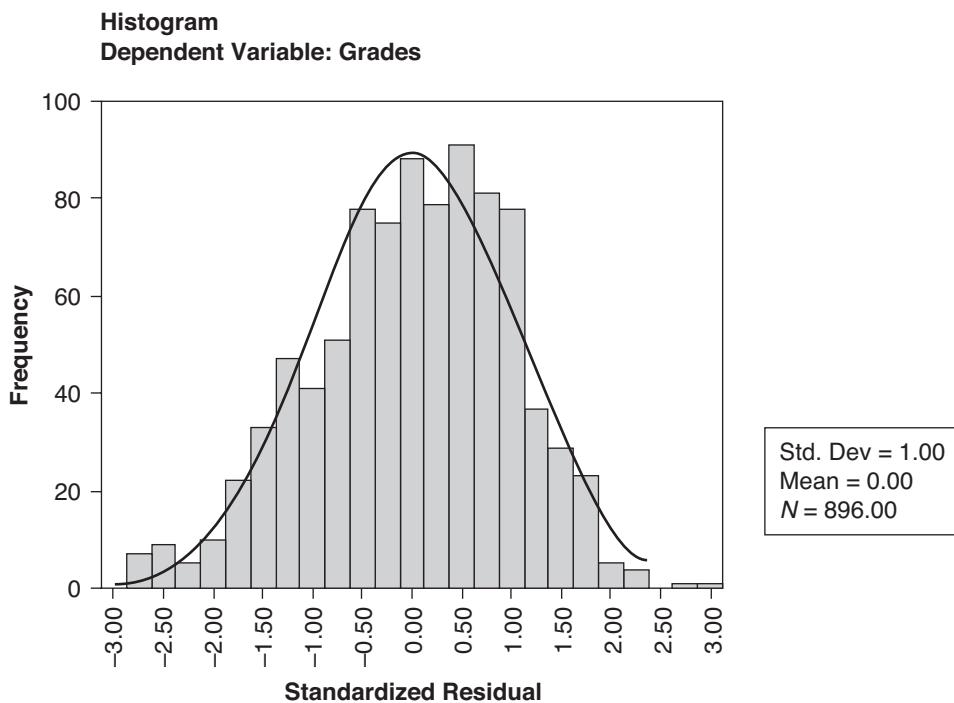


Figure 10.5 Testing for the normality of residuals. The residuals form a nearly normal curve.

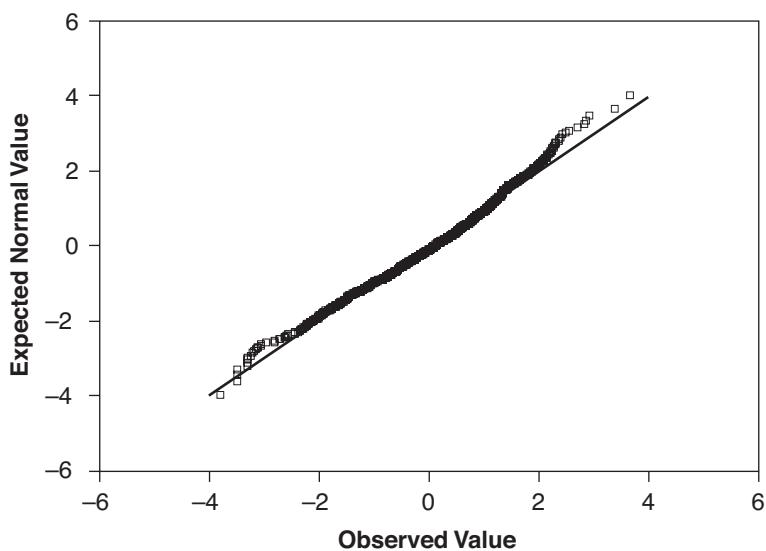


Figure 10.6 A q-q plot of the residuals. The residuals' adherence to a nearly straight line supports their normality

uses the cumulative frequency and is interpreted in the same fashion (looking for departures from a straight line).

Excessive heteroscedasticity and nonnormal residuals can sometimes be corrected through transformation of the dependent variable. Eliminating subgroups from the regression may

also be useful. Finally, there are alternative regression methods (e.g., weighted least squares regression) that may be useful when these assumptions are seriously violated (see Cohen et al., 2003, and Darlington, 1990, for more information).

Diagnosing Data Problems

Regression diagnostics for spotting problematic data points focus on three general characteristics: distance, leverage, and influence. Conceptually, how would you spot unusual or problematic cases, commonly referred to as outliers or as extreme cases? Focus on Figure 10.7, a reprint of the earlier Figure 3.7. The figure is a byproduct of the regression of students' Grades on Parent Education and Homework. Recall that we saved the variable Predicted Grades, which I demonstrated was an optimally weighted composite of the two independent variables, weighted so as to best predict the outcome. The figure shows students' GPA plotted against their Predicted GPAs. Note the case circled in the lower right of the figure. This case is among the farthest from the regression line; this is one method of isolating an extreme case, called *distance*. *Leverage* refers to an unusual pattern on the independent variables and does not consider the dependent variable. If you were using homework in different academic areas to predict overall GPA, it would not be unusual to find a student who spent 1 hour per week on math homework nor would it be unusual to find a student who spent 8 hours per week on English homework. It would likely be unusual to find a student who combined these, who spent only 1 hour per week on math while spending 8 hours per week on English. This case would likely have high leverage. Because leverage is not calculated with respect to the dependent variable, the graph shown here may not be informative as to leverage; a graph of the two independent variables may be more useful (as we will soon see). The final characteristic of interest is

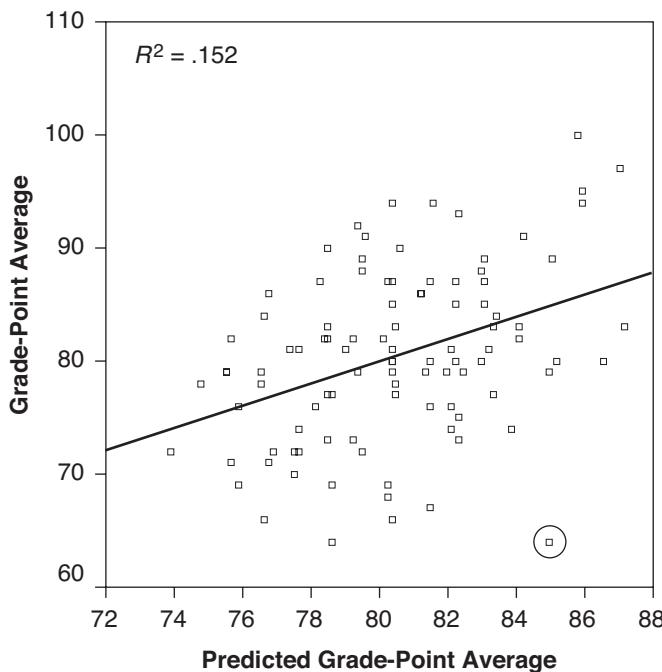


Figure 10.7 Predicted versus actual Grades plot from Chapter 3. The circled case is a potential extreme case, a long distance from the regression line.

influence. As the name implies, a case that has high influence is one that, if removed from the regression, results in a large change in the regression results. Cases with high influence are those that are high on both distance and leverage. The circled case would likely fit this description as well. If it were deleted from the regression, the regression line would likely be somewhat steeper than it is in the figure.

Distance

Common measures of distance are derived from the residuals. In Figure 10.7, the residual for the circled case is the point on the regression line above the case (approximately 85) minus the actual value of the case (64). This definition matches well the conceptual definition of distance given previously.

In practice, the unstandardized residuals are less useful than are standardized versions of residuals. Table 10.1 shows some of the cases from this data set. The first column shows the case number, followed by the dependent variable Grades and the two independent variables Parent Education and Homework. Column five shows the Predicted Grades used to create the graph in Figure 10.7. The remaining columns show various regression diagnostics. The first row of the table shows the names assigned these variables in SPSS, under which I have included a brief explanation. Column six, labeled ZRE_1, shows the standardized residuals, which are the residuals standardized to approximately a normal distribution. Think of them like *z* scores, with values ranging from 0 (very close to the regression line) to ± 3 or more. The next column (SRE_1) represents the standardized residuals converted to a *t* distribution (the *t* distribution is also referred to as Student's *t*, hence the S), which are generally called the studentized or *t* residuals. The advantage of this conversion is that the *t* residuals may be tested for statistical significance (see Darlington, 1990, p. 358). In practice, however, researchers often simply examine large positive or negative standardized or studentized residuals or, with reasonable sample size, those greater than an absolute value of 2 (with very large samples, there may be many of these).

The cases shown in Table 10.1 were chosen for display because they have high values for distance, leverage, or influence. As shown in the table, cases 34 (-3.01) and 83 (2.06) show high values for studentized residuals.

Figure 10.8 shows the same plot of Predicted and actual Grades, with a few of the cases identified. Note the case that was originally circled is case number 34, the highest negative studentized (and standardized) residual. As can be seen, case 83, with a high positive studentized residual, is also far away from the regression line. It might be worth investigating these cases with high residuals further to make sure that they have been coded and entered correctly.

Leverage

Leverage gets at the unusualness of a pattern of independent variables, without respect to the dependent variable. The column in Table 10.1 labeled LEV_1 provides an estimate of leverage (this measure is also often referred to as *h*). Leverage ranges from 0 to 1, with an average value of $(k + 1)/n$ (*k* = number of independent variables); twice this number has been suggested as a rule of thumb for high values of leverage (Pedhazur, 1997, p. 48). Case 16 in the table had the highest value for leverage (.098), followed by cases 36 (.088) and 32 (.084). Both these values are higher than the rule of thumb would suggest:

$$2\left(\frac{k+1}{N}\right) = 2\left(\frac{3}{100}\right) = .06.$$

Table 10.1 Regression Diagnostics for the Regression of Grades on Parent Education and Homework (Data from Chapter 3).

Casenum	Grades	Pared	Hwork	Predgrad	ZRE_1	SRE_1	SDR_1	tResid, deleted	Standardized DF Beta			Standardized DF Beta			
									Studentized, t residual		Cook	Leverage	intercept	pared	hwork
									standardized residual	t residual					
12.00	72.00	13.00	5.00	79.48435	-1.05539	-1.06231	-1.06302	0.00495	0.00299	-0.07044	0.05836	-0.01163			
13.00	66.00	12.00	3.00	76.63804	-1.50010	-1.52071	-1.53122	0.02134	0.01693	-0.19095	0.12503	0.11783			
14.00	79.00	14.00	4.00	79.36713	-0.05177	-0.05211	-0.05184	0.00001	0.00303	-0.00098	-0.00072	0.00287			
15.00	76.00	10.00	4.00	75.88464	0.01627	0.01673	0.01664	0.00001	0.04405	0.00377	-0.00347	0.00009			
16.00	80.00	20.00	6.00	86.56656	-0.92597	-0.98069	-0.98049	0.03901	0.09848	0.30209	-0.32258	0.04489			
17.00	91.00	15.00	8.00	84.18914	0.96042	0.97535	0.97510	0.00994	0.02038	-0.04474	0.01145	0.13224			
32.00	83.00	15.00	11.00	87.15267	-0.58558	-0.61536	-0.61338	0.01317	0.08446	0.03559	0.01979	-0.18448			
33.00	78.00	13.00	6.00	80.47220	-0.34861	-0.35156	-0.34996	0.00070	0.00669	-0.02205	0.02422	-0.02180			
34.00	64.00	17.00	7.00	84.94254	-2.95316	-3.00886	-3.14360	0.11492	0.02668	0.45737	-0.42923	-0.16864			
35.00	82.00	13.00	4.00	78.49651	0.49404	0.49765	0.49571	0.00121	0.00448	0.03455	-0.02020	-0.01998			
36.00	81.00	17.00	1.00	79.01546	0.27984	0.29462	0.29322	0.00313	0.08776	-0.03788	0.06746	-0.07800			
37.00	73.00	13.00	4.00	78.49651	-0.77508	-0.78075	-0.77917	0.00298	0.00448	-0.05430	0.03175	0.03141			
80.00	72.00	10.00	5.00	76.87248	-0.68708	-0.70760	-0.70576	0.01012	0.04714	-0.15793	0.15778	-0.04060			
81.00	79.00	17.00	4.00	81.97900	-0.42008	-0.42961	-0.42780	0.00283	0.03391	0.05808	-0.07712	0.04376			
82.00	93.00	14.00	7.00	82.33067	1.50451	1.51942	1.52989	0.01533	0.00954	0.01337	-0.04408	0.15087			
83.00	100.00	18.00	7.00	85.81316	2.00052	2.05698	2.09249	0.08073	0.04414	-0.41586	0.40491	0.08101			
84.00	90.00	13.00	4.00	78.49651	1.62214	1.63401	1.64841	0.01307	0.00448	0.11487	-0.06717	-0.06645			
85.00	69.00	10.00	4.00	75.88464	-0.97082	-0.99817	-0.99815	0.01898	0.04405	-0.22643	0.20833	-0.00511			

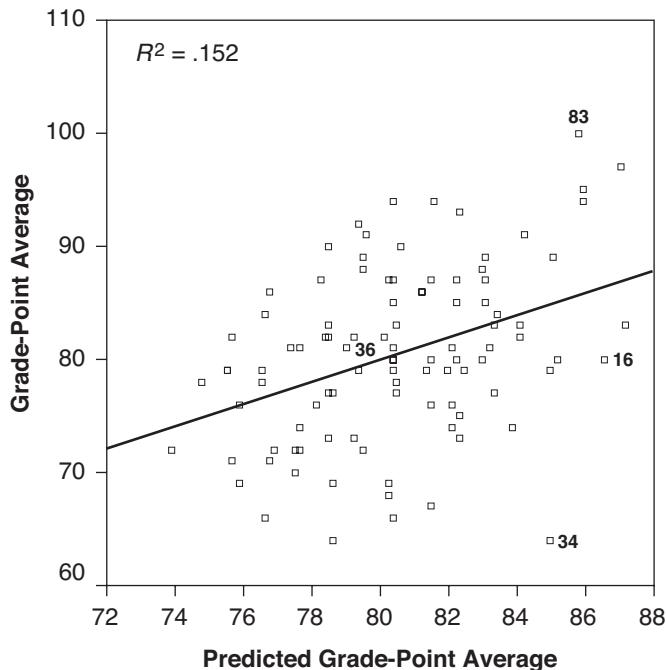


Figure 10.8 Plot from Figure 10.7 with several noteworthy cases highlighted.

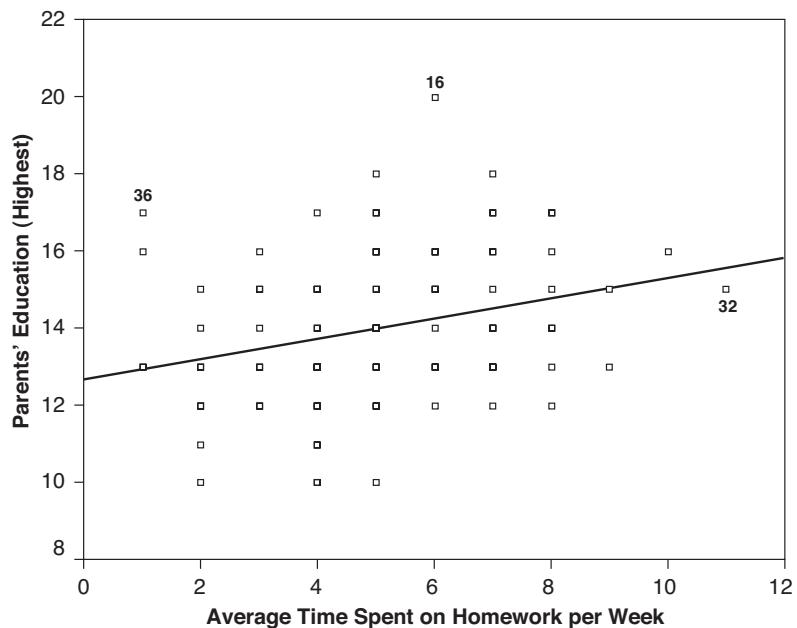


Figure 10.9 Leverage illustrated.

As can be seen in Figure 10.8, you might suspect that case 16 was unusual from a visual display (because it is on one edge of the graph), but case 36 is right in the middle of the graph. Recall, however, that leverage does not depend on the dependent variable. Figure 10.9 shows a plot of the two independent variables. Cases 16, 36, and 32 are outside the “swarm”

of most of the cases; they indeed represent an unusual combination of independent variables. These cases may also be worth checking.

Influence

Influence means what the name suggests: a case that is highly influential on the intercept or the regression line. The column labeled Coo_1 (for Cook's Distance) in Table 10.1 provides values of an estimate of influence; cases with large values are worth inspecting. The cases with the largest Cook's D values were cases 34 (.115) and 83 (.081). The regression plane would move the most if these cases were omitted.

Most computer programs also compute estimates of *partial* influence (as in influence, with the effects of the other independent variables accounted for). The DF Betas, standardized, listed in the last three columns are estimates of partial influence. The first of these columns (SDB0_1) pertains to the regression intercept, the second (SDB1_1) to the first independent variable (Parent Education), and the third (SDB2_1) to the second independent variable (Homework). The values shown are the change in each parameter, if a particular case were removed. A negative value means that the particular case lowered the value of the parameter, whereas a positive value means that the case raised the parameter. So, for example, case 34 had standardized DF Beta values of .457, -.429, and -.169. Case 34 served to raise the intercept and lower the regression coefficient for Parent Education and Homework. Although the unstandardized DF Betas are not shown in Table 10.1, they were 2.29, -.158, and -.058. If you run the regression without case 34, you will find that the intercept reduces by 2.29, the Parent Education *b* increases by .158, and the Homework *b* increases by .058.

An inspection of the standardized DF Betas showed large negative values by case 83 for the intercept (-.416) and large positive value for case 34 (.457). These two cases were also very influential for the Parent Education regression coefficient, although reversed: case 34 (-.429), case 83 (.405). The partial influence values for the Homework variable were considerably smaller. Cases 21 and 29 had the highest values (.334 and .335).

Uses

What do these various regression diagnostics tell us? In the present example, cases 34 and 83 showed up across measures; it would certainly be worth inspecting them. But inspecting them for what? Sometimes these diagnostics can point out errors or misentered data. A simple slip of the finger may cause you to code 5 hours of homework as 50. This case will undoubtedly show up in the regression diagnostics, thus alerting you to the mistake. Of course, a simple careful inspection of the data will likely spot this case as well! Think about the example I used initially to illustrate leverage, however, someone who reports 1 hour of Math Homework and 8 hours of English Homework. This case will not show up in a simple inspection of the data, because these two values are reasonable and, taken by themselves, only become curious when taken together. The case will likely be spotted in an analysis of both leverage and influence; we might well discover that errors were made in entering this datum as well.

If there are not obvious errors for the variables spotted via regression diagnostics, then what? In our present example, cases 34 and 83, although outliers, are reasonable. A check of the raw data shows that case 34 had well-educated parents, higher than average homework, but poor grades. Case 83 simply had an excellent GPA and higher than average homework. On further investigation, I might discover that case 34 had a learning disability, and I might decide to delete this case and several other similar cases. Or I might decide that the variation is part of the phenomenon I am studying and leave case 34 in the analysis. Another option is additional analysis. If a number of outliers share characteristics in common and are

systematically different from other cases, it may suggest that a different regression is needed for these participants or the advisability of including an interaction term in the analysis (e.g., Disability Status by Parent Education). It might also suggest the inclusion of an important common cause (e.g., disability status affecting both time spent on homework and subsequent grades).

Obviously, unless clear-cut errors are involved, considerable judgment is involved in the inspection of regression diagnostics. Note that deletion of case 34 will increase the regression weight for Homework; if I did delete this case, I will need to be sure that my deletion is based on a concern about its extremity rather than a desire to inflate the apparent importance of my findings. If you do delete cases based on regression diagnostics, you should note this in the research write-up and the reasons for doing so. With the present example and after examining cases with high values on all the regression diagnostics, I would first double-check each of these values against the raw data but would likely conclude in the end that all the cases simply represented normal variation. I would then leave the data in their present form.

Again, I have barely scratched the surface of an important topic; it is worth additional study. Darlington (1990, chap. 14), Darlington and Hayes (2017), Fox (2008), and Pedhazur (1997) each devote chapters to regression diagnostics and are worth reading.

Multicollinearity

I mentioned briefly when discussing interactions the potential problem of multicollinearity (also called collinearity). Briefly, multicollinearity occurs when several independent variables correlate at an excessively high level with one another or when one independent variable is a near linear combination of other independent variables. Multicollinearity can result in misleading and sometimes bizarre regression results.

Figure 10.10 shows some results of the regression of a variable named Outcome on two independent variables, Var1 and Var2. The correlations among the three variables are also shown. The results are not unusual and suggest that both variables have positive and statistically significant effects on Outcome.

Now focus on Figure 10.11. For this analysis, the two independent variables correlated at the same level with the dependent variable as in the previous example (.3 and .2). However, in this example, Var1 and Var2 correlate .9 with each other (versus .4 in the previous example). Notice the regression coefficients. Even though all variables correlate positively with

		Correlations		
		OUTCOME	VAR1	VAR2
Pearson Correlation	OUTCOME	1.000	.300	.200
	VAR1	.300	1.000	.400
	VAR2	.200	.400	1.000
Sig. (1-tailed)	OUTCOME	.	.000	.000
	VAR1	.000	.	.000
	VAR2	.000	.000	.
N	OUTCOME	500	500	500
	VAR1	500	500	500
	VAR2	500	500	500

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error				Lower Bound	Upper Bound	Tolerance	VIF
	64.286	5.133		12.524	.000	54.201	74.370		
1	(Constant)	.262	.046	.262	.5633	.000	.171	.353	.840
	VAR1	.952E-02	.046	.095	2.048	.041	.004	.187	.840
	VAR2								1.190

a. Dependent Variable: OUTCOME

Figure 10.10 Regression of Outcome on Var1 and Var2. The results are reasonable.

		Correlations		
		OUTCOME	VAR1	VAR2
Pearson Correlation	OUTCOME	1.000	.300	.200
	VAR1	.300	1.000	.900
	VAR2	.200	.900	1.000
Sig. (1-tailed)	OUTCOME	.	.000	.000
	VAR1	.000	.	.000
	VAR2	.000	.000	.
N	OUTCOME	500	500	500
	VAR1	500	500	500
	VAR2	500	500	500

		Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	73.684	4.373		16.848	.000	65.092	82.277		
	VAR1	.632	.097	.632	6.527	.000	.441	.822	.190	5.263
	VAR2	-.368	.097	-.368	-3.807	.000	-.559	-.178	.190	5.263

a. Dependent Variable: OUTCOME

Figure 10.11 Regression of Outcome on Var1 and Var2 when Var1 and Var2 are very highly correlated (collinear). The results are puzzling, and the interpretation will likely be misleading.

one another, Var1 seems to have a positive effect on Outcome, whereas Var2 has a negative effect. As noted previously, multicollinearity can produce strange results such as these; standardized regression coefficients greater than 1 are also common. Notice also that the standard errors of the b 's are also considerably larger for the second example than for the first. Multicollinearity also inflates standard errors; sometimes two variables will correlate at similar levels with an outcome, but one will be a statistically significant predictor of the outcome, while the other will not, as a result of multicollinearity.

Conceptually, multicollinearity suggests that you are trying to use two variables in a prediction that overlap completely or almost completely with one another. Given this definition, it makes intuitive sense that multicollinearity should affect standard errors: the more that variables overlap, the less we can separate accurately the effects of one versus the other. Multicollinearity is often a result of a researcher including multiple measures of the same construct in a regression. If this is the case, one way to avoid the problem is to combine the overlapping variables in some way, either as a composite or, as is done in Part 2, using the variables as indicators of a latent variable. Multicollinearity is also often a problem when researchers use a kitchen-sink approach: throwing a bunch of predictors into regression and using stepwise regression, thinking it will sort out which are important and which are not.

Given the example, you may think you can spot multicollinearity easily by examining the zero-order correlations among the variables, with high correlations alerting you to potential problems. Yet multicollinearity can occur even when the correlations among variables are not excessive. A common example of such an occurrence is when a researcher, often inadvertently, uses both a composite and the components of this composite in the same regression. For example, in Figure 10.12 I regressed BYTests on grades in each academic area, in addition to a composite Grades variable (BYGrads). Notice the results: the overall R^2 is statistically significant, but none of the predictors is statistically significant. In this example, the largest individual correlation was .801, however, not overly large. The zero-order correlations are not always useful in spotting collinearity.

How can you avoid the effects of multicollinearity? Computer programs provide, on request, collinearity diagnostics. Such statistics are shown in Figures 10.10 through 10.12. Tolerance is a measure of the degree to which each variable is independent of (does not

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.558 ^a	.311	.307	7.10940

a. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYS81B math88-grades, BYS81A English88-grade, BYS81D sstudies88-grades, BYS81C science88-grades

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20119.51	5	4023.903	79.612	.000 ^a
	Residual	44579.48	882	50.544		
	Total	64699.00	887			

a. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYS81B math88-grades, BYS81A English88-grade, BYS81D sstudies88-grades, BYS81C science88-grades

b. Dependent Variable: BYTESTS 8th-grade achievement tests (mean)

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error				Tolerance	VIF
1	(Constant)	33.123	5.370				
	BY81A English88-grade	8.19E-02	1.501	.009	.055	.956	.027
	BY81B math88-grades	-.698	1.490	-.077	-.469	.639	.029
	BY81C science88-grades	.767	1.499	.090	.511	.609	.025
	BY81D sstudies88-grades	-.125	1.487	-.015	-.084	.933	.026
	BYGRADS GRADES COMPOSITE	6.241	6.008	.538	1.039	.299	.003

a. Dependent Variable: BYTESTS 8th-grade achievement tests (mean)

Figure 10.12 Another cause of multicollinearity. A composite and its components are both used in the regression.

overlap with) the other independent variables (Darlington & Hayes, 2017). Tolerance can range from 0 (no independence from other variables) to 1 (complete independence); larger values are desired. The variance inflation factor (VIF) is the reciprocal of tolerance and is “an index of the amount that the variance of each regression coefficient is increased” over that with uncorrelated independent variables (Cohen et al., 2003, p. 423). Small values for tolerance and large values for VIF signal the presence of multicollinearity. Cohen and colleagues (2003, p. 423) note that a common rule of thumb for a large value of VIF is 10, which means that the standard errors of b are more than three times as large as with uncorrelated variables ($\sqrt{10} = 3.16$), but that this value is probably too high. Note that use of this value will lead to an inspection and questioning of the results in Figure 10.12, but not those in Figure 10.11. Values for the VIF of 6 or 7 may be more reasonable as flags for excessive multicollinearity (cf. Cohen et al., 2003). These values of the VIF correspond to tolerances of .10 (for a VIF of 10), .14 (VIF of 7), and .17 (VIF of 6), respectively.

Factor analysis of independent variables and “all subsets” regression can also be useful for diagnosing problems. When you get strange regression results, you should consider and investigate multicollinearity as a possible problem. Indeed, it is a good idea to routinely examine these statistics. A method known as ridge regression can be used when data are excessively collinear.

Obviously, I have just touched the surface of this important topic; it is worth additional study. Pedhazur (1997) presents a readable, more detailed discussion of the topic, as does Darlington (1990, chaps. 5, 8). Darlington and Hayes (2017, chap. 4) offer useful suggestions for dealing with collinearity.

SAMPLE SIZE AND POWER

“How large a sample do I need?” Anyone who has advised others on the use of multiple regression (or any other statistical method) has heard this question more times than he or she can count. This question may mean several things. Some who ask it are really asking, “Is there some minimum sample size that I can’t go below in MR?” Others are looking for a rule of thumb, and there is a common one: 10 to 20 participants for each independent variable. Using this rule, if your MR includes 5 independent variables, you need at least 50 (or 100) participants. I’ve heard this rule of thumb many times but have no idea where it comes from. We will examine it shortly to see if it has any validity for the types of MR problems we have been studying. Finally, more sophisticated researchers will ask questions about what sample size they need to have a reasonable chance of finding statistical significance.

I hope you recognize this final version of the question as one of the *power* of MR. I have alluded to power at several points in this text (e.g., in the discussion of interactions in MR, testing for mediation), but, as you will see, we have really sidestepped the issue until this point by our use of the NELS data. With a sample size of 1000, we had adequate power for all the analyses conducted. You can’t always count on sample sizes in the thousands, however, so let us briefly turn to the issue of power and sample size.

Briefly, power generally refers to the ability correctly to reject a false null hypothesis. It is a function of the magnitude of the effect (e.g., whether Homework has a small or a large effect on Grades); the alpha, or probability level chosen for statistical significance (e.g., .05, .01, or some other level); and the sample size used in the research. Likewise, the necessary sample size depends on effect size, chosen alpha, and desired power. The needed sample size increases as desired power increases, effect size decreases, and alpha gets more stringent (i.e., as the probability chosen gets smaller). Common values for power are .8 or .9, meaning that given a particular effect size one would like to have an 80% or 90% chance of rejecting a false null hypothesis of no effect. Like alpha, and despite conventions, power levels should be chosen based on the needs of a particular study.

This short section is, of course, no treatise on power analysis. What I do plan to do here is to examine power and sample size for the rule of thumb given previously, as well as some of the examples we have used in this book, to give you some sense of what sorts of sample sizes are needed with the kinds of problems used in this book. Fortunately, there are some excellent books on power analysis, including Cohen’s classic book on the topic (1988). The Darlington and Hayes (2017) and Cohen and colleagues (2003) text is useful on this topic as well and many others; for experimental research, I found Howell’s (2013) introduction to power especially clear. If you intend to conduct research using MR (or other methods), I recommend that you read further on this important issue. You should and can also have access to a program for conducting power analysis. The examples that follow use G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007), a free power analysis program available for download (www.gpower.hhu.de/, or just search for “GPower”). I have also used SamplePower from SPSS, and the PASS (Power Analysis and Sample Size) program from NCSS (www.ncss.com); they also are easy to use and work well.

First, let’s examine several of the examples in this text. In Chapter 4, we regressed GPA in 10th grade on Parent Education, In School Homework, and Out of School Homework in a simultaneous regression. The R^2 for the overall regression was .155, with a sample size of 909.

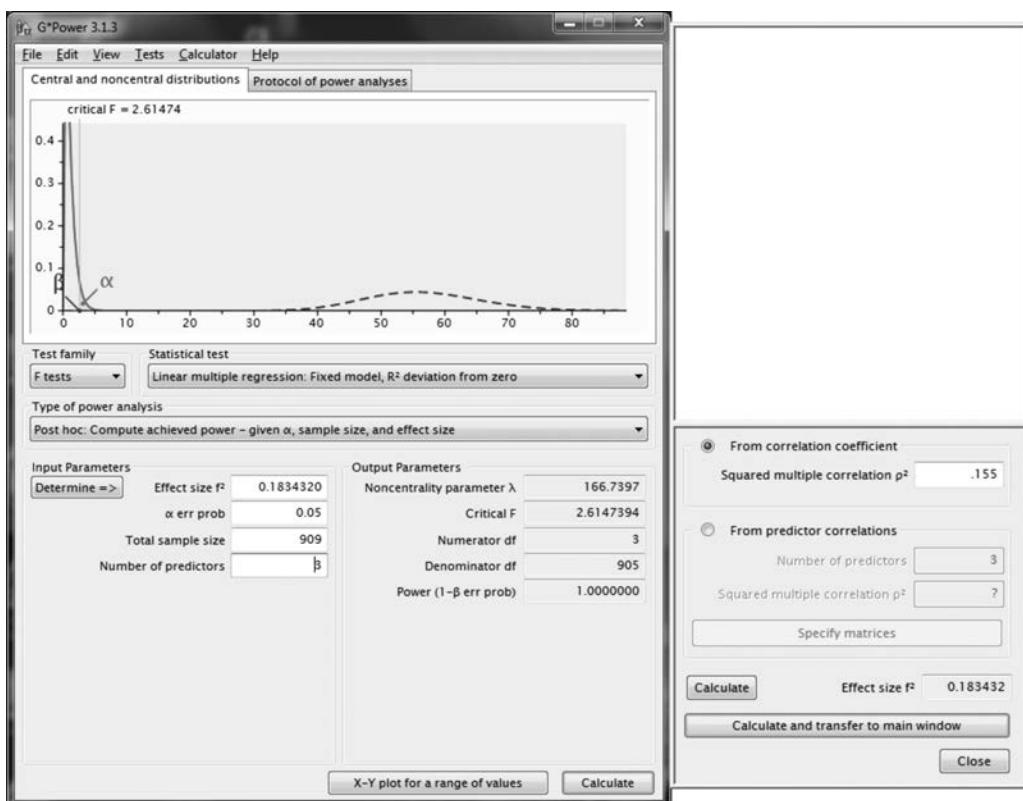


Figure 10.13 Power analysis for the overall regression of GPA on Parent Education, In-School Homework, and Out-of-School Homework from Chapter 4.

What sort of power did we have with this simultaneous regression? According to G*Power, this example had a power of 1.0 (for this and the other examples, I will assume an alpha of .05) for the overall regression. In other words, given the information previously, we had a 100% chance of correctly rejecting a false null hypothesis. Figure 10.13 shows the relevant screen shot. We are interested in an F test (Test Family), and are interested in the overall regression (e.g., the statistical significance of the overall R^2), so choose “Fixed model . . . R^2 deviation from zero.” G*Power uses f^2 as its measure of effect size, but it is easy to convert R^2 and ΔR^2 into f^2 (see chapters 4 and 5); indeed, G*Power will do these calculations for you, as shown on the smaller right-hand screen (to get this screen, click on the “Determine” button under “Input Parameters.”). The figure also shows the results.

These findings are for a post-hoc power analysis; that is, we conducted the regression and then wondered what the power was. Much more useful for most researchers is an a priori power analysis, in which we plan the research and then calculate the needed sample size to have a good chance of rejecting a false null hypothesis. With these three variables and an R^2 of .155, we will have a power of .8 with 64 participants and a power of .9 with 82 participants. Figure 10.14 shows a graph of power (Y-axis) as a function of sample size, given an alpha of .05 and an R^2 of .155.

We often are interested in the power of the addition of one variable or a block of variables to the regression equation, with other variables (background variables or covariates) controlled. For example, in Chapter 5 we considered the sequential regression in which we added Locus of Control and Self-Esteem to the regression, with SES and Previous Grades already

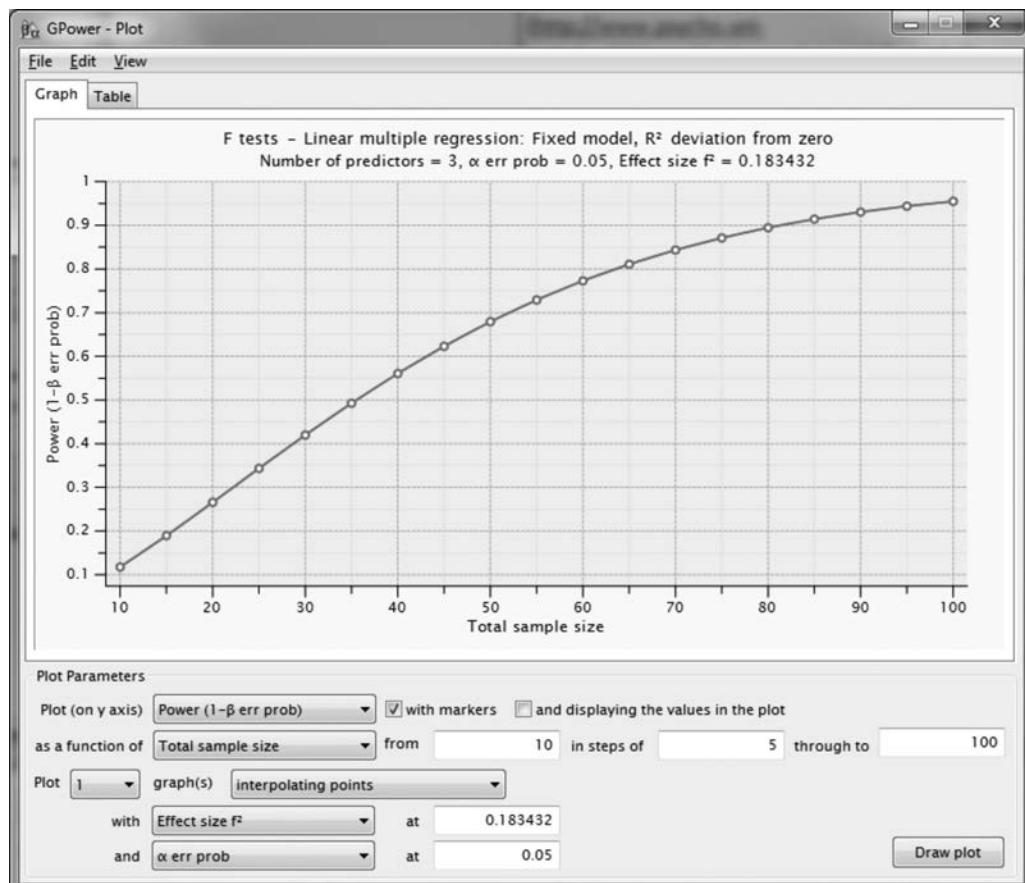


Figure 10.14 Power to detect a statistically significant R^2 as a function of sample size. This figure refers to the same regression as Figure 10.13, both from Chapter 4.

in the equation. The R^2 with two variables in the equation was .328, and the psychological variables added another .010 to the R^2 . What sort of power was associated with this block? Given the sample size of 887, this final block in the regression had a power of .92; given this information, we had a 92% chance to reject correctly a false null hypothesis of no effect for the psychological variables. Given these same numbers, a sample size of 641 (see Figure 10.15) would be needed for a power .80 and sample size of 841 for a power of .90 for this block. The top of Figure 10.15 shows the input values for G*Power; the lower portion shows the sample size graph.

Consider the regressions in which we added interaction terms to the regression. In Chapter 7 we tested the interaction of Previous Achievement and Ethnic origin in their possible effect on Self-Esteem. The categorical and continuous variable accounted for 2% of the variance in Self-Esteem, and the cross product added another .8% (which I will round off to 1%) to the variance explained, with a sample size of approximately 900. In this example, the test of the interaction term had a power of .86 (post hoc) and .80 power would be achieved with a sample size of 764 (a priori). Although the test of the interaction has lower power than the initial variables, with this sample size we still had adequate power to examine the statistical significance of the interaction.

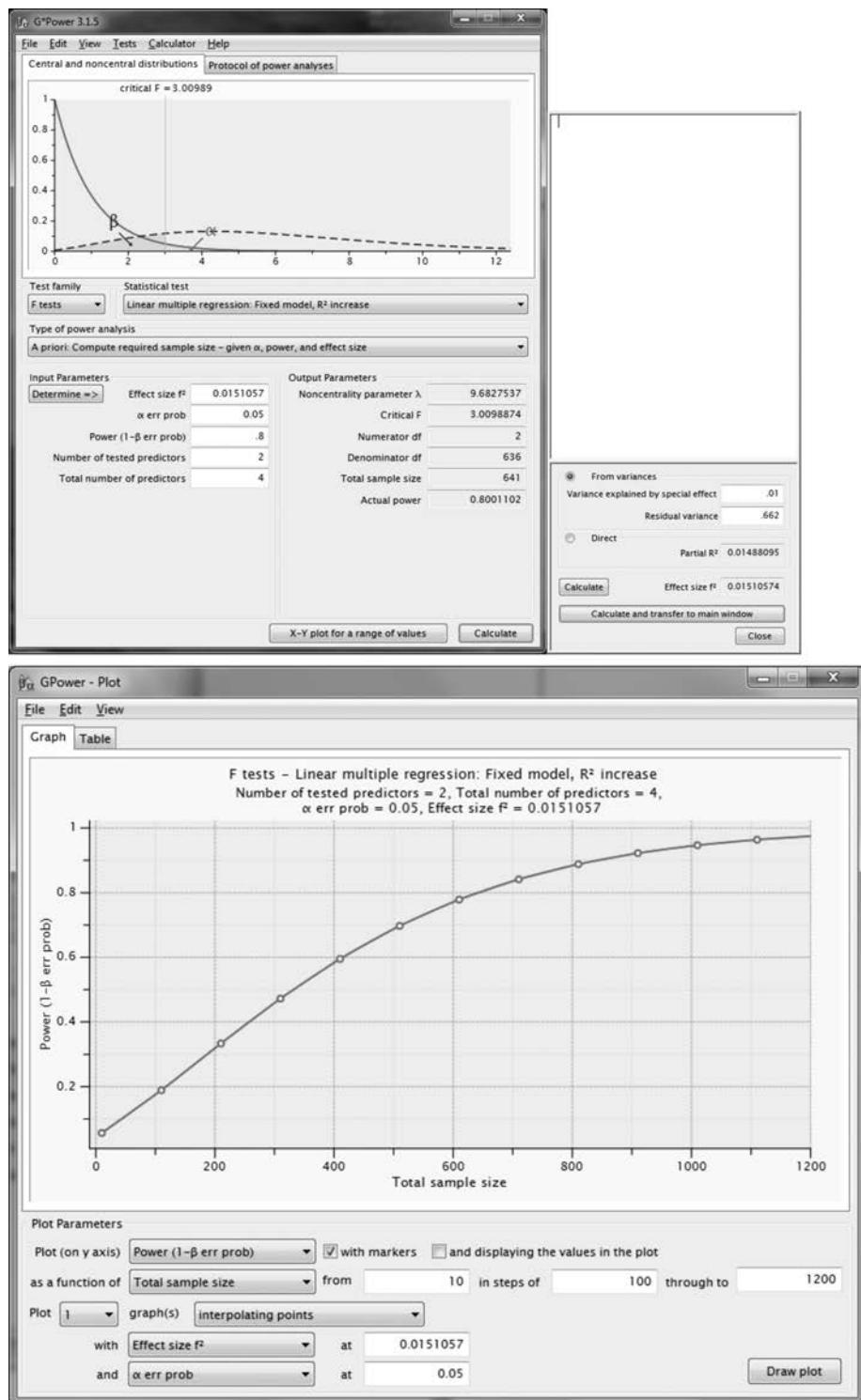


Figure 10.15 Power analysis for ΔR^2 for one of the sequential regressions from Chapter 5.

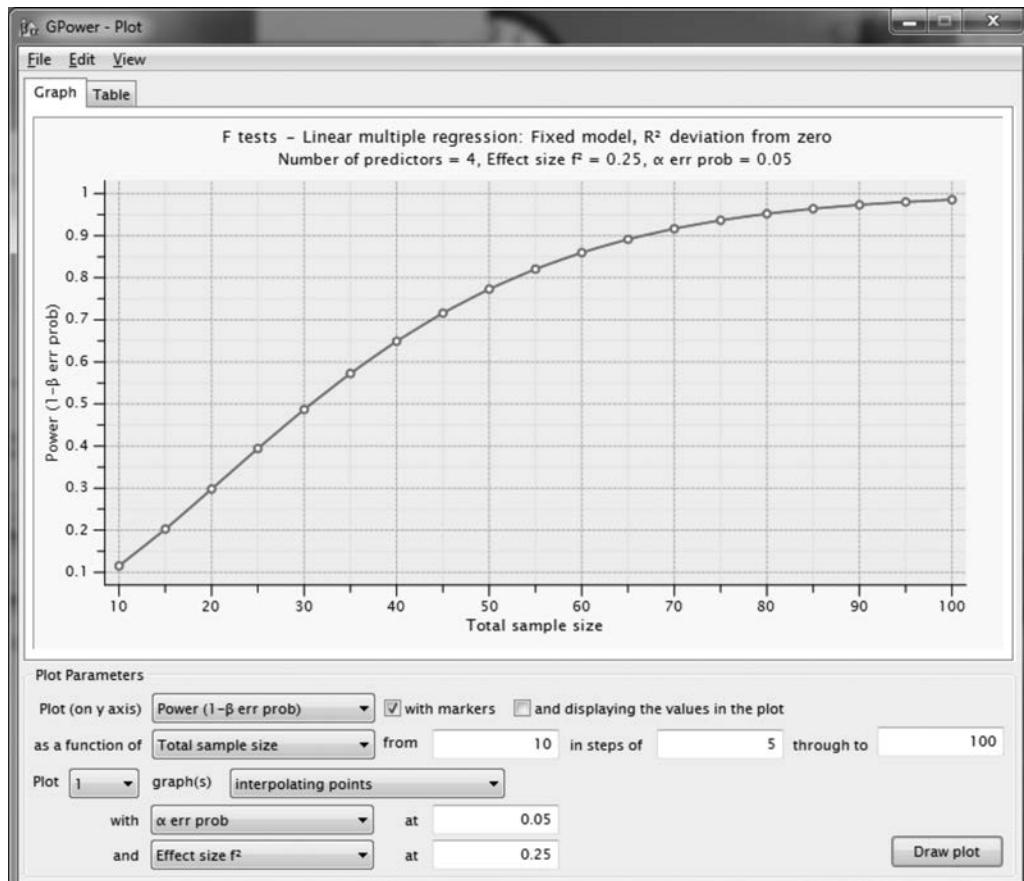


Figure 10.16 Power as a function of sample size for $R^2 = .20$ ($f^2 = .25$). The example illustrates potential problems with a common rule of thumb for sample size in multiple regression.

Finally, consider the 10 to 20 participants per independent variable rule of thumb. Let's model this on some of the other regressions discussed here. Suppose four independent variables account for 20% of the variance in the outcome ($f^2 = .25$), a value that seems reasonable given our examples. Will a sample size of 40 to 80 produce adequate power? Forty cases will produce a total power of only .65, but 80 cases will result in a power of .95. The relevant graph is shown in Figure 10.16. If the R^2 for these four variables was .30 ($f^2 = .43$) instead of .20, then the power associated with 40 cases is .89 (no graph shown). Suppose instead that you were interested in the power associated with one variable that increased the R^2 by .05 above an $R^2 = .20$ ($\Delta f^2 = .067$) from the first four variables in the regression. You will need a sample size of 120 to have a power of .80 for this final variable (see Figure 10.17). It appears that this rule of thumb, although sometimes accurate, will produce low power in many real-world research problems.¹

In real-world research, you should, of course, conduct these power calculations prior to the research to make sure you collect data on the needed number of participants. You will not know the exact effect size but can generally estimate effect sizes from previous research and your knowledge of relevant theory in the area. Most programs use R^2 or ΔR^2 as the measure of effect size, or the easily calculable f^2 or Δf^2 (as in the previous examples). You can, of course, get estimates of ΔR^2 if researchers have used sequential regression or by squaring

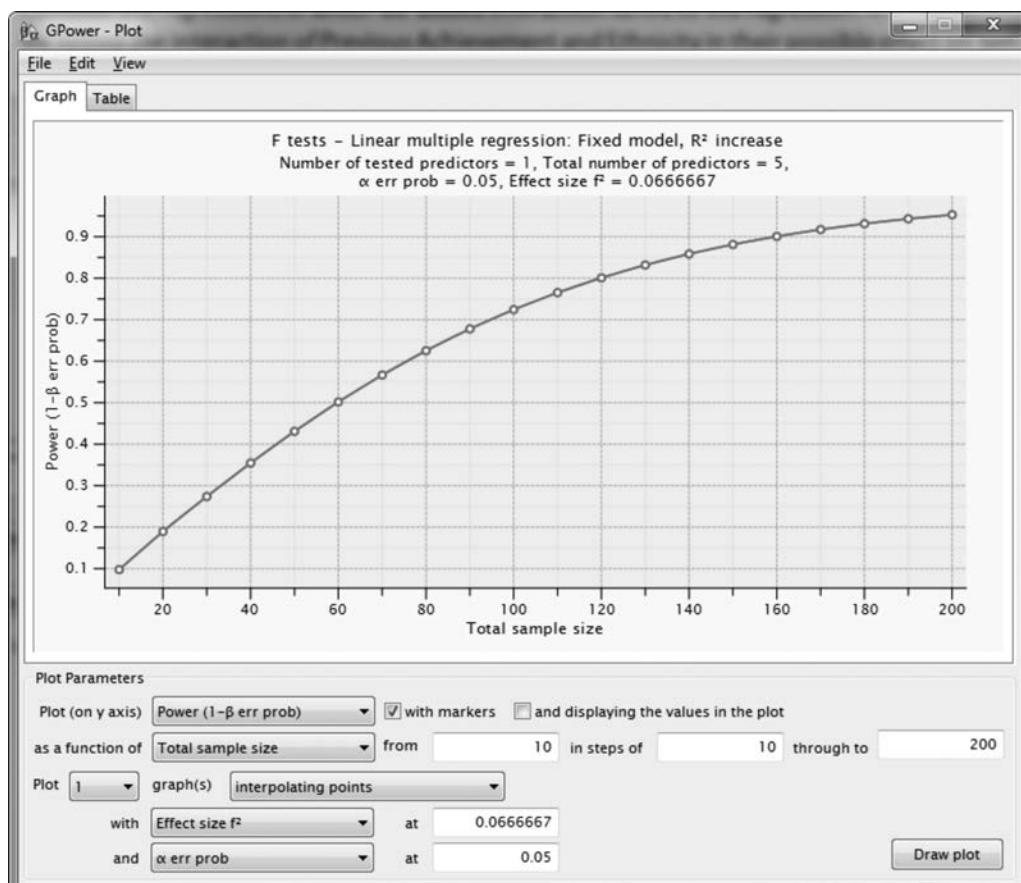


Figure 10.17 Power as a function of sample size for $\Delta R^2 = .05$ (with $1 - R^2 = .75$).

the semipartial correlations (which you can calculate using t values, if necessary). If you have no previous research to go on, you can use common rules of thumb (e.g., R^2 's of .01, .09, and .25; f^2 's of .02, .13, and .35 represent small, medium, and large effects in the social sciences; Cohen et al., 2003). A medium effect size is generally recognized as one noticeable to a knowledgeable observer (Howell, 2013).

As you plan your own research, I encourage you to investigate power more completely and spend some time estimating the sample size you will need in your research (assuming you are not using a large data set like NELS). You don't want to be filled with regrets after having conducted the research and finding nothing of statistical significance and then wishing that you had collected data from 10, or 100, additional participants!

PROBLEMS WITH MR?

Let's revisit some of the interpretive problems we've dealt with throughout this part of the book. I conducted three multiple regressions of high school Achievement on Family Background (SES), Intellectual Ability, Academic Motivation, and Academic Coursework in high school. Our interest is in the effects of these variables on students' high school achievement. We will briefly examine the results of a simultaneous, a sequential, and a stepwise multiple regression, with a focus on the different conclusions we can reach using the different

methods. Because our primary interest is in the differences across methods, I won't define the variables in any more detail. The data are taken from Keith and Cool (1992), however, if you are interested in learning more. For this example, rather than simulating the data, I have conducted the regressions using a portion of the correlation matrix as presented in the article. The file "problems w MR 3.sps" illustrates how to conduct a MR using a correlation matrix in SPSS. You may want to save or print this file; it's a useful method and one you can use to reanalyze any published correlation matrix.

Figure 10.18 shows the primary results from a simultaneous MR of Achievement on the four explanatory variables. The regression is statistically significant, and over 60% of the variance in Achievement is explained by these four variables ($R^2 = .629$). The table of coefficients in the figure provides information about the relative influence of the variables. All the variables appear important, with the exception of Academic Motivation. The effects of Motivation appear very small ($\beta = .013$) and are not statistically significant. Motivation, it seems, has no effect on high school Achievement. Turning to the other variables and based on the β 's, Ability appears the most important influence, followed by high school Coursework; both effects were large. Family Background, in contrast, had a small but statistically significant effect on Achievement.

Figure 10.19 shows the same data analyzed via sequential MR. For this problem, the explanatory variables were entered in the order of presumed time precedence. Parents' background characteristics generally come prior to their children's characteristics; Ability, a relatively stable characteristic from an early age, comes prior to the other student characteristics; Motivation determines in part the courses students take in high school; and these courses, in turn, determine in part a high school student's Achievement. Thus, achievement was regressed on Family Background, then Ability, then Motivation, and finally Coursework. Relevant results of this regression are shown in Figure 10.19.

There are several differences in these results and those from the simultaneous MR. What is more disturbing is that we will likely come to different conclusions depending on which printout we examine. First, with the sequential regression and focusing on the statistical significance of ΔR^2 for each step, it now appears that Academic Motivation *does* have a statistically significant effect on Achievement ($\Delta R^2 = .009$, $F[1, 996] = 19.708$, $p < .001$). Second,

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.793 ^a	.629	.629	421.682	4	995	.000

a. Predictors: (Constant), COURSES, FAM_BACK, MOTIVATE, ABILITY

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant) 6.434	1.692		3.803	.000	3.114	9.753
	FAM_BACK .695	.218	.069	3.194	.001	.268	1.122
	ABILITY .367	.016	.551	23.698	.000	.337	.398
	MOTIVATE 1.26E-02	.021	.013	.603	.547	-.028	.054
	COURSES 1.550	.120	.310	12.963	.000	1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 10.18 Simultaneous regression of Achievement on Family Background, Ability, Motivation, and Academic Coursework.

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.417 ^a	.174	.174	210.070	1	998	.000
2	.747 ^b	.558	.384	865.278	1	997	.000
3	.753 ^c	.566	.009	19.708	1	996	.000
4	.793 ^d	.629	.063	168.039	1	995	.000

a. Predictors: (Constant), FAM_BACK

b. Predictors: (Constant), FAM_BACK, ABILITY

c. Predictors: (Constant), FAM_BACK, ABILITY, MOTIVATE

d. Predictors: (Constant), FAM_BACK, ABILITY, MOTIVATE, COURSES

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B		
	B	Std. Error				Lower Bound	Upper Bound	
1	(Constant)	50.000	.288	173.873	.000	49.436	50.564	
	FAM_BACK	4.170	.288	.417	.000	3.605	4.735	
2	(Constant)	4.557	1.559		2.923	.004	1.498	7.617
	FAM_BACK	1.328	.232	.133	5.729	.000	.873	1.782
	ABILITY	.454	.015	.682	29.416	.000	.424	.485
3	(Constant)	.759	1.766		.430	.667	-2.706	4.224
	FAM_BACK	1.207	.231	.121	5.221	.000	.753	1.661
	ABILITY	.445	.015	.667	28.768	.000	.414	.475
	MOTIVATE	9.53E-02	.021	.095	4.439	.000	.053	.137
4	(Constant)	6.434	1.692		3.803	.000	3.114	9.753
	FAM_BACK	.695	.218	.069	3.194	.001	.268	1.122
	ABILITY	.367	.016	.551	23.698	.000	.337	.398
	MOTIVATE	1.26E-02	.021	.013	.603	.547	-.028	.054
	COURSES	1.550	.120	.310	12.963	.000	1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 10.19 Sequential regression results for the same data.**Model Summary**

Model	R	R Square	Adjusted R Square	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.737 ^a	.543	.543	.543	1186.615	1	998	.000
2	.791 ^b	.625	.624	.082	217.366	1	997	.000
3	.793 ^c	.629	.628	.004	10.453	1	996	.001

a. Predictors: (Constant), ABILITY

b. Predictors: (Constant), ABILITY, COURSES

c. Predictors: (Constant), ABILITY, COURSES, FAM_BACK

Figure 10.20 Stepwise regression of Achievement on the same four school learning variables.

although we still conclude that Ability was the most important variable, we now conclude that Family Background was second in importance ($\sqrt{\Delta R^2} = .620, .417, .251, .095$, for Ability, Family Background, Coursework, and Motivation, respectively; of course this rank order would stay the same if we were to focus on ΔR^2 instead).

Figure 10.20 shows the results from a stepwise regression of these same variables. Again, Academic Motivation appears unimportant, because it never entered the regression equation.

Table 10.2 Regression Coefficients from the Simultaneous versus Sequential Regression of Achievement on Family Background, Ability, Academic Motivation, and Academic Coursework.

Variable	Simultaneous Regression	Sequential Regression
Family Background	.695 (.218) .069	4.170 (.288) .417
Ability	.367 (.016) .551	.454 (.015) .682
Academic Motivation	.013 (.021) .013	.095 (.021) .095
Academic Coursework	1.550 (.120) .310	1.550 (.120) .310

Note. The first row for each variable shows the unstandardized coefficient followed by the standard error (in parentheses). The second row shows the standardized coefficient.

And again, the order of “importance” changed. In the stepwise regression, Ability entered the equation first, followed by Coursework, followed by Family Background. The stepwise regression thus seems to paint yet another picture of the importance of these variables for Achievement.

How do we resolve these differences? First, we can ignore the results of the stepwise regression, because this is an explanatory problem and stepwise regression is not appropriate for explanatory research. But we still have the differences between the simultaneous and the sequential regressions, both of which are appropriate for explanation.

We have touched on these differences in previous chapters. As noted primarily in Chapter 5, simultaneous regression focuses the *direct* effects of variables on an outcome, whereas sequential regression focuses on *total* effects. Thus, the two approaches may well produce different estimates, even when they are based on the same underlying model and even when one interprets the same statistics. Table 10.2 shows the relevant regression coefficients from Figures 10.18 (simultaneous regression) and 10.19 (sequential regression). For the sequential regressions, the coefficients are from the step at which each variable was entered (shown in italic boldface in the table of coefficients in Figure 10.19). Note the differences in the coefficients; many of the differences are large. Family Background, for example, has an effect of .069 (standardized) in the simultaneous regression versus .417 in the sequential regression.

Again, these differences are not so startling if we know that the simultaneous regression focuses on direct effects versus total effects for sequential regression. But many users of multiple regression seem unaware of this difference. Likewise, many users of MR seem unaware that their regression, when used for explanatory purposes, implies a model and that this model should guide the analysis. The model that underlies these regressions is shown in Figure 10.21, and it can be used to illustrate the differences in coefficients between simultaneous and sequential regression. The simultaneous regression estimates the direct effects, labeled a, b, c, and d in the figure. The sequential regression estimates aspects of the total effects. Thus for the variable motivation, the coefficient for Motivation is the direct effect of Motivation on Achievement (path b) plus the indirect effect of Motivation on Achievement through Academic Coursework (path e times path a).

In Part 2 of this book we will develop such models in considerably more detail and, along the way, gain a deeper understanding of MR and our current difficulties in interpretation. Even if you are using this book for a class in MR only and focusing on Part 1 only, I urge you

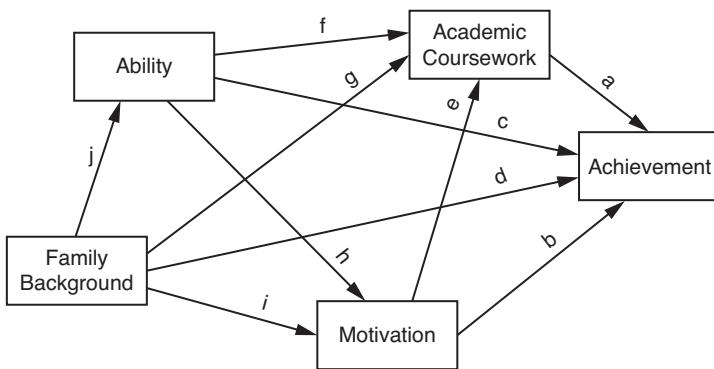


Figure 10.21 Model underlying the simultaneous and sequential regressions of Achievement on Family Background, Ability, Academic Motivation, and Academic Coursework.

to read Part 2 (at least the first two chapters). I think you will find they help you resolve many of the issues that have vexed us—and apparently others—in the use and interpretation of multiple regression. If nothing else, these chapters will give you a more complete heuristic aid in understanding MR results.

EXERCISES

1. Return to the first regression we did with the NELS data. Regress 10th-grade GPA (FFU-Grad) on Parent Education (BYParEd) and Time Spent on Homework Out of School (F1S36A2) (see the exercises in Chapter 2). Save the unstandardized residuals and predicted values. Use the residuals to test for linearity in the Homework variable and for the overall regression. Are the residuals normally distributed? Is the variance of the errors consistent across levels of the independent variables (to conduct this final analysis, I suggest you reduce the Predicted Grades variable into a smaller number of categories)?
2. Rerun the regression; save standardized and studentized residuals, leverage, Cook's Distance, and standardized DF Betas. Check any outliers and unusually influential cases. Do these cases look okay on these and other variables? What do you propose to do? Discuss your options and decisions in class. (To do this analysis, you may want to create a new variable equal to the case number [e.g., COMPUTE CASENUM=\$CASENUM in SPSS]. You can then sort the cases based on each regression diagnostic to find high values, but still return the data to their original order.)
3. Do the same regression, adding the variable BYSES to the independent variables (BYParEd is a component of BYSES). Compute collinearity diagnostics for this example. Do you note any problems?

Note

- 1 Two slightly more sophisticated rules of thumb are $N > 50 + 8k$ for calculating the N needed for adequate power in an overall regression and $> 104 + k$ for testing the statistical significance of a single variable (with k representing the number of independent variables). Green (1991) evaluated these and other rules of thumb and, although they work somewhat better than the simple $N > 10k$ rule mentioned in this chapter, they also fall short, because they do not take effect sizes into account. Indeed, the second rule would underestimate the sample size needed for the final example given here. Green also developed several additional rules of thumb that take effect size into account and are therefore more useful. I recommend you use a power analysis program rather than rules of thumb, but this article is still worth reading.

11

Related Methods

Logistic Regression and Multilevel Modeling

Logistic Regression	227
<i>Predicting Optimism Versus Pessimism</i>	227
<i>Multiple Regression Analysis</i>	228
<i>Problems With the MR Analysis</i>	228
<i>Logistic Regression: Transforming the Dependent Variable to Log Odds</i>	232
<i>Conducting the Logistic Regression and Understanding the Output</i>	235
<i>Categorizing a Continuous Variable</i>	239
<i>Appropriate Uses of Logistic Regression</i>	240
<i>Logistic Regression Versus Discriminant Analysis</i>	241
Multilevel Modeling	241
<i>Effects of SES on Achievement</i>	242
<i>Multiple Regression Analysis</i>	242
<i>Multilevel Analysis of the Effects of SES on Achievement</i>	245
<i>MLM: Next Steps</i>	252
Summary	252
Exercises	253

This chapter will focus on two methods that are similar to multiple regression—logistic regression and multilevel modeling (aka hierarchical linear modeling)—but that require specialized analysis with something other than the linear regression procedures in general statistics programs. These are methods you are likely to encounter in your reading, so it is useful to have a conceptual understanding of them. Logistic regression is useful when the dependent variable you are interested in is dichotomous (or categorical with more than two categories). Multilevel modeling takes into account the often-nested or clustered nature of our data, such as children within schools, or individuals within families.

The intent of this chapter is not to teach you how to conduct such analyses. Instead, here I hope to illustrate these methods from a now-familiar multiple regression orientation. The chapter will use an example of each method with a brief look at the output from the analysis and what it means. You won't be adept at the methods after reading this chapter, but I hope that you will have a conceptual understanding of them and an idea of where you can go to learn more. As a result I hope when you do encounter these methods either in your reading or in subsequent coursework that you will have a mental schema in place that will help you understand them more fully.

LOGISTIC REGRESSION

In all our regression examples in this book, the outcome variable has been continuous. Achievement test scores, Grades, Self-Esteem ratings, and so on, are all continuous variables.

Starting in Chapter 6 we spent considerable time discussing how to use dummy and effect coding to analyze categorical independent variables. But what do you do when you have a categorical *dependent* variable? Logistic regression is one method of dealing with this problem. It is most commonly used to predict a dichotomous dependent variable from multiple continuous or continuous and categorical independent variables. Let's illustrate the method with an example.

Predicting Optimism Versus Pessimism

I created an Optimism composite variable as the average of a series of questions about students' outlook toward the future (F1S64A through F1S64K): "Think about how you see the future. What are the chances that:

- You will graduate from high school?
- You will go to college?
- You will have a job that pays well?
- You will be able to own your own home?
- You will have a job that you enjoy doing?
- You will have a happy family life?
- You will stay in good health most of the time?
- You will be able to live wherever you want in the country?
- You will be respected in your community?
- You will have good friends you can count on?
- Life will turn out better for you than it has for your parents?"

Possible answers ranged from 1 (very low) to 3 (about 50–50) to 5 (very high). Students with high scores on the composite had a fairly optimistic view of the future, whereas those with low scores were more pessimistic about the future. This composite is used in its continuous form in Appendix C on Partial and Semipartial correlation. For illustration of logistic regression I divided this continuous variable into an Optimism/Pessimism dichotomous variable. Students who had an average rating of less than 4 were categorized as pessimistic, whereas those with an average rating of 4 or higher were categorized as optimistic (4 corresponded to an answer of "high" on each item). This dichotomization seems reasonable, as many people think of this a categorical variable anyways. But we will discuss this dichotomization more, later. As shown in Figure 11.1, using this categorization, about 36% of the NELS sample would be classified as pessimistic versus 64% optimistic.

Optim_Press Optimistic or Pressimistic

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00 Pessimistic	321	32.1	35.5	35.5
	1.00 Optimistic	583	58.3	64.5	100.0
	Total	904	90.4	100.0	
Missing	System	96	9.6		
	Total	1000	100.0		

Figure 11.1 Breakdown of the optimistic/pessimistic dichotomy.

Suppose you were interested in predicting whether high school students were more likely to be optimistic or pessimistic. What variables might you use? I picked two of our old standbys: Family Background, or SES, as referenced by the NELS variable BYSES, and 8th-grade achievement test scores (BYTests). My thinking is that students who come from more advantaged backgrounds (higher on the BYSES variable), and who achieve at a higher level are more likely to be optimistic about the future than those who are lower on these two variables.

Two other composites were created and used in the prediction. Substance was a measure of self-reported substance use in 10th grade, and was a composite of F1S77 (How many cigarettes smoked per day), F1S78 (In lifetime, number of times had alcohol to drink), and F1S80aa (In lifetime, number of times used marijuana). Religious was the average of F1S82 (How often attend religious services) and F1S83 (Thinks is a religious person). All these items were measured in 10th grade, and were converted to z-scores prior to averaging. It seems likely that students who report lower substance use and higher levels of religiosity should be more likely to be optimistic than those who report more substance use and lower levels of religiosity. We will see if these predictions are accurate. Histograms for all variables are shown in Figure 11.2.

Multiple Regression Analysis

Other than the dichotomous dependent variable, this sounds like a standard multiple regression-type analysis: predict, or explain, a dependent variable from four independent variables. We could indeed treat it like a multiple regression and regress the optimism/pessimism variable on those four predictors. Some of the output from this analysis is shown in Figure 11.3. As shown in the Figure, the four IVs explained about 12% of the variance in the dichotomous optimism/pessimism outcome variable, and this was statistically significant ($F(4, 799) = 28.288, p < .001$). In addition, each of the four independent variables was statistically significant, with (judging by the β s) religiosity, SES, and 8th-grade achievement all having moderate effects, and substance use having a small but statistically significant negative effect. Higher achieving, more advantaged, and more religious students were more likely to be optimistic than were their peers who were lower on these scales, and 10th-graders who used substances like tobacco, alcohol, and marijuana were more likely to be pessimistic than were their peers who did not use (or used less of) these substances.

Problems With the MR Analysis

What is wrong with this analysis? It works and is interpretable, correct? Yet there are reasons why a MR approach is not the best one. First, the magnitude of the correlations between the dichotomous dependent variable and the independent variables depends, in part, on the split used in the dichotomization, with the maximum correlation decreasing as we depart from a 50/50 split (this is actually more of a criticism of turning a continuous variable into a categorical one, and we will return to it at the end of this section). And, of course, the multiple regression results, including R^2 and all of the regression coefficients, depend on the correlations of the dependent variable with the independent variables (see Chapter 3). So, for example, with the variables used in the regression shown in Figure 11.3, 35.5% of students were classified as pessimistic versus 64.5% optimistic. With this dichotomization, the correlation between BYTests and the optimism/pessimism dichotomous variable was .210. If, however, we had used a value of 3 on the original (continuous) optimism scale as the cut-point for creating a dichotomy, the pessimistic/optimistic split would have been 2.9/97.1%. The correlation of this version of the DV with BYTests would be .172.

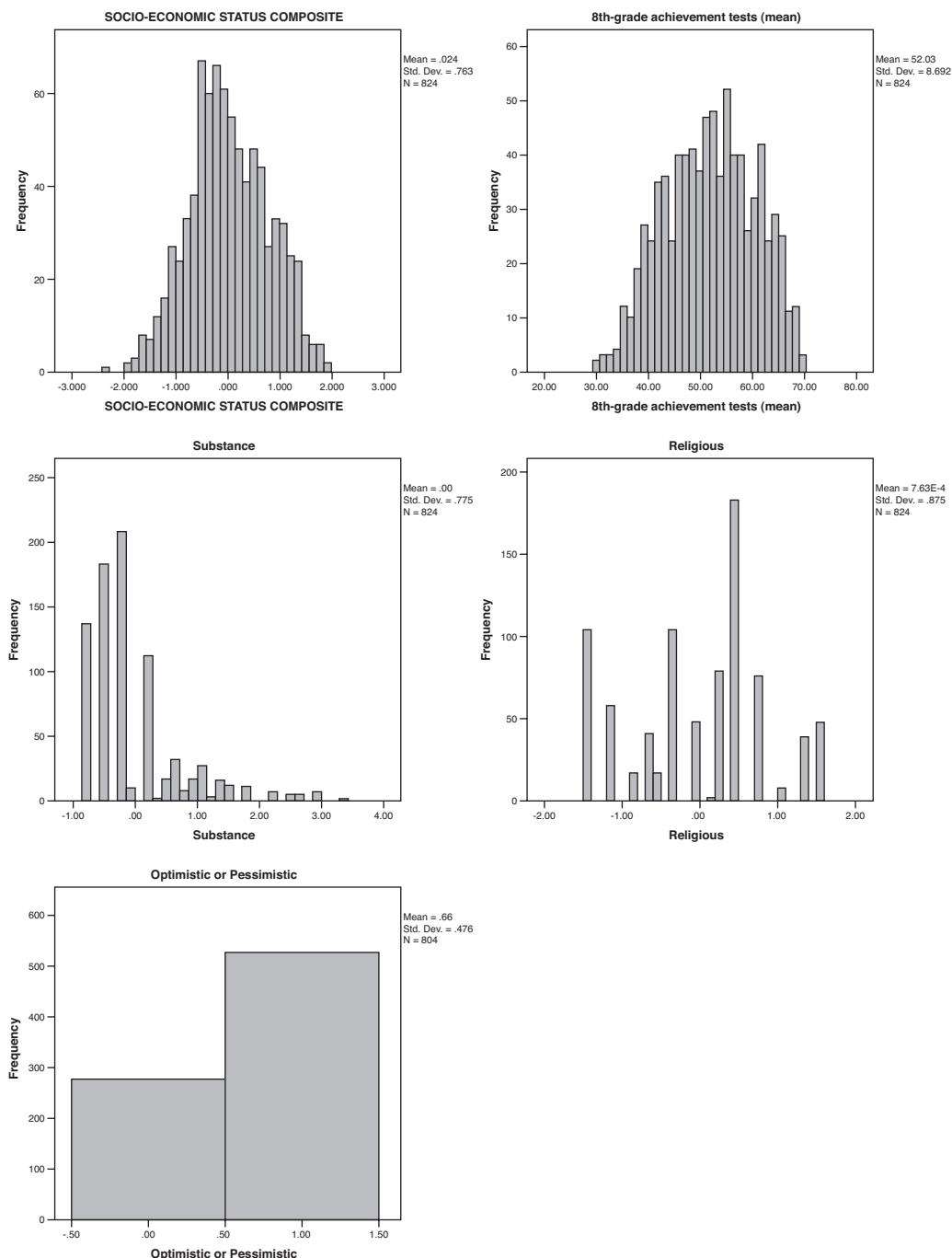


Figure 11.2 Histograms of independent and dependent variables from the logistic regression example.

In Chapter 10 we discussed the assumptions underlying multiple regression and various ways to test those assumptions. Another reason the MR approach is not the best way to deal with the prediction of a categorical dependent variable is that we violate many of those assumptions. Look at the graphs shown in Figure 11.4. The first graph shows a scatterplot of the optimism/pessimism dichotomous variable regressed on the unstandardized predicted

Descriptive Statistics				
	Mean	Std. Deviation	N	
Optim_Press Optimistic or Pessimistic	.6555	.47551	804	
Substance	-.0008	.77350	804	
Religious	.0098	.87918	804	
byses SOCIO-ECONOMIC STATUS COMPOSITE	.02206	.762524	804	
bytests 8th-grade achievement tests (mean)	52.0521	8.62075	804	

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.352 ^a	.124	.120	.44615

a. Predictors: (Constant), bytests 8th-grade achievement tests (mean), Religious, Substance, byses SOCIO-ECONOMIC STATUS COMPOSITE

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22.523	4	5.631	28.288	.000 ^a
	Residual	159.043	799	.199		
	Total	181.566	803			

a. Predictors: (Constant), bytests 8th-grade achievement tests (mean), Religious, Substance, byses SOCIO-ECONOMIC STATUS COMPOSITE

b. Dependent Variable: Optim_Press Optimistic or Pessimistic

Coefficients^a

Model	Unstandardized Coefficients		Beta	t	Sig.
	B	Std. Error			
1 (Constant)	.276	.109		2.539	.011
Substance	-.052	.021	-.084	-2.439	.015
Religious	.098	.019	.181	5.214	.000
byses SOCIO-ECONOMIC STATUS COMPOSITE	.099	.024	.158	4.179	.000
bytests 8th-grade achievement tests (mean)	.007	.002	.131	3.485	.001

a. Dependent Variable: Optim_Press Optimistic or Pessimistic

Figure 11.3 Results of a multiple regression with the dichotomous optimism dependent variable.

value from the MR. Recall from Chapters 3 and 10 that this predicted value is a linear combination of the four independent variables, so this scatterplot represents the regression of the DV against the combined IVs. Note that all data points are clustered at the top or the bottom of the Y axis, and this is because the dependent variable can only take on two possible values: 0 or 1. The scatterplot shows the regression line, as in many previous scatterplots. I hope it is clear that the linear fit line is not the best possible line among these data points; that is, a nonlinear relation could produce a better fit. The second graph shows a scatterplot of the regression standardized residuals against the predicted values, with a Loess fit line. In chapter 10 we saw how to use this fit line to assess the linearity assumption of multiple regression. This loess line certainly shows a considerable deviation from a straight line, suggesting a violation of this important assumption. In fact, a nonlinear fit line, an ogive, would be a better fitting line for these data; such a line is shown in Figure 11.5.

Another assumption for regression discussed in Chapter 10 is that the errors, or residuals, will be normally distributed. Figure 11.6 shows two graphs that can be used to probe this

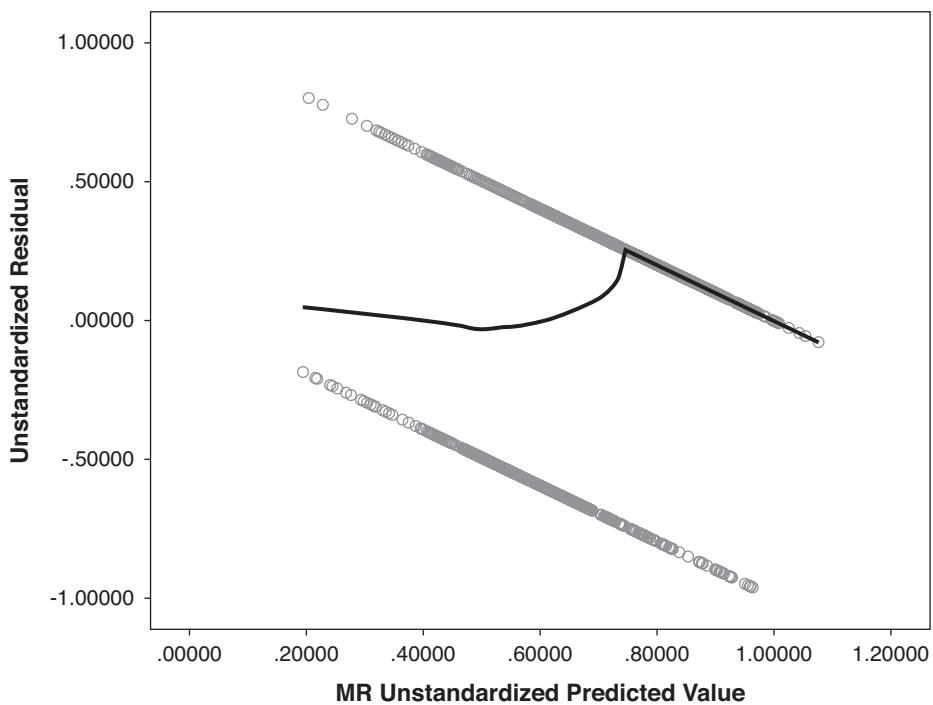
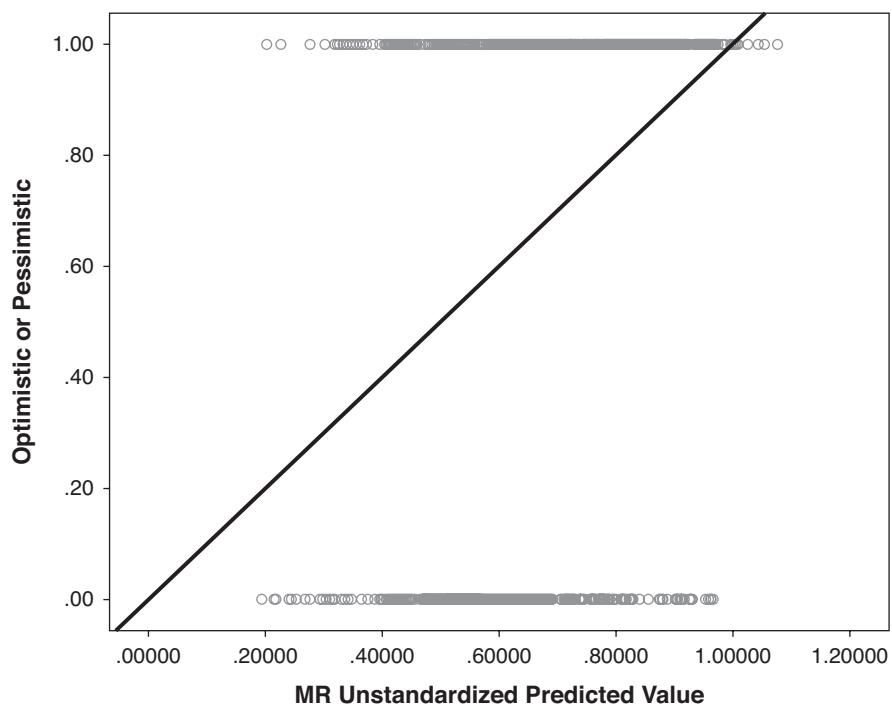


Figure 11.4 Graphs demonstrating the non-linear relation between optimism/pessimism and its predictors, and suggesting that ordinary multiple linear regression is not a good analysis option.

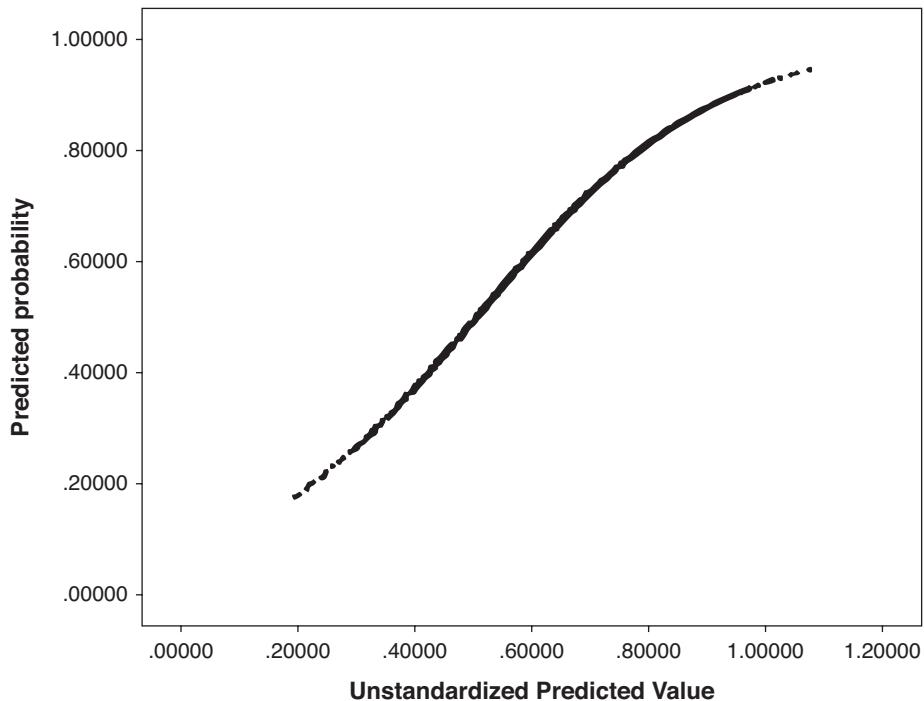


Figure 11.5 A better-fitting line of the relation between optimism/pessimism predicted by the four independent variables.

assumption. The top graph shows a histogram of the residuals from the multiple regression using these data. This multimodal plot clearly departs from a normal distribution. The lower part of the figure shows a P-P plot of the residuals. Normally distributed residuals would cluster around the straight diagonal line. Again, there is a clear departure from normality, and again, it appears that an ordinary MR with these data leads to some clear violations of the assumptions for regression.

Logistic Regression: Transforming the Dependent Variable to Log Odds

Also noted in Chapter 10 was that sometimes such assumption violations could be addressed via a transformation of the variables involved. That is one way of thinking about what happens in a logistic regression (LR), instead of focusing on the dichotomous dependent variable, we focus the odds of that variable. Thus we ask what are the *odds* of being pessimistic or optimistic given a particular set of values for the independent variables? We have not discussed odds in this text, but it is a concept that is likely at least a little familiar to most of us, and especially anyone who likes to gamble. Odds are related to probability, and are calculated as the probability of an event happening divided by the probability of it not happening. In this case, the odds would be calculated as the probability of being optimistic divided by the probability of being pessimistic (for a given set of values for the independent variables).

Odds ratios are easy to interpret, but keep in mind that they are different from, but related to, probability. An odds ratio of 2 means that the odds are doubled, whereas one of .5 means that the odds are halved. In the current example, the use of odds is easier to explain with a single independent variable, and one with a limited number of possible values. The religiosity variable fulfills this requirement, and it was also (at least in the regression) the strongest predictor of optimism. Figure 11.7 shows the crosstabs of the Religious variable with the

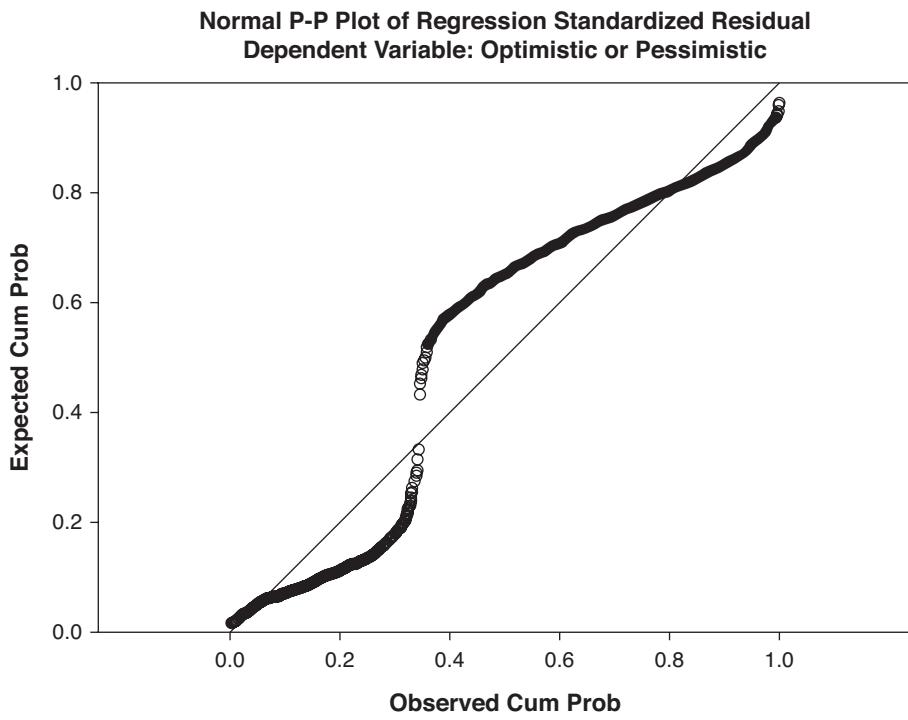
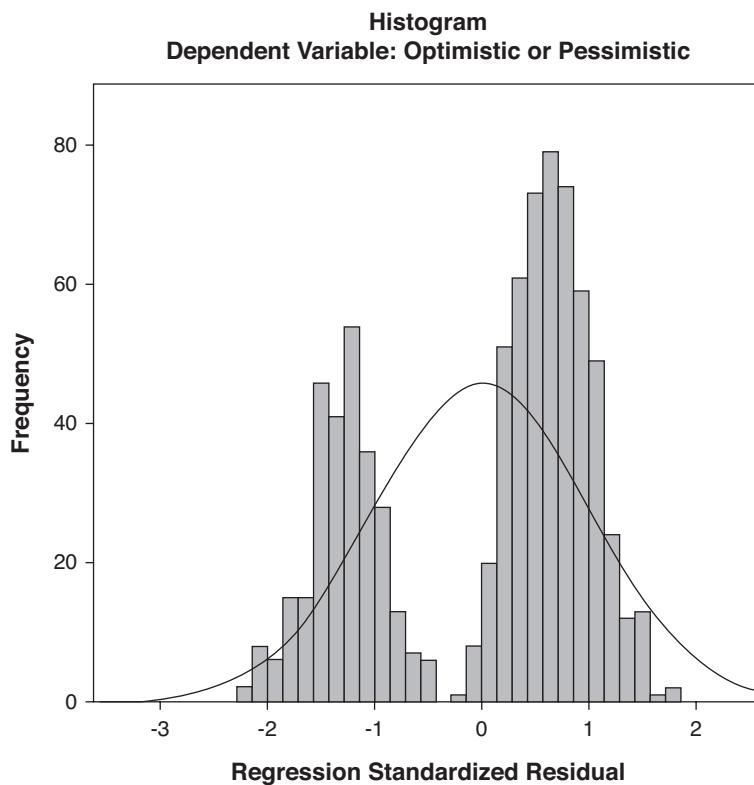


Figure 11.6 The residuals for the multiple regression are not normally distributed.

Religious * Optim_Pess Optimistic or Pessimistic Crosstabulation

			Optim_Pess Optimistic or Pessimistic		Total
			.00 Pessimistic	1.00 Optimistic	
Religious	-1.43	Count	57	44	101
		% within Religious	56.4%	43.6%	100.0%
	-1.15	Count	22	35	57
		% within Religious	38.6%	61.4%	100.0%
	-.87	Count	10	7	17
		% within Religious	58.8%	41.2%	100.0%
	-.62	Count	15	23	38
		% within Religious	39.5%	60.5%	100.0%
	-.59	Count	12	5	17
		% within Religious	70.6%	29.4%	100.0%
	-.34	Count	28	51	79
		% within Religious	35.4%	64.6%	100.0%
	-.31	Count	9	11	20
		% within Religious	45.0%	55.0%	100.0%
	-.06	Count	19	23	42
		% within Religious	45.2%	54.8%	100.0%
	-.04	Count	2	2	4
		% within Religious	50.0%	50.0%	100.0%
	.20	Count	0	2	2
		% within Religious	0.0%	100.0%	100.0%
	.22	Count	19	56	75
		% within Religious	25.3%	74.7%	100.0%
	.48	Count	1	3	4
		% within Religious	25.0%	75.0%	100.0%
	.50	Count	44	133	177
		% within Religious	24.9%	75.1%	100.0%
	.76	Count	0	3	3
		% within Religious	0.0%	100.0%	100.0%
	.78	Count	20	53	73
		% within Religious	27.4%	72.6%	100.0%
	1.03	Count	1	7	8
		% within Religious	12.5%	87.5%	100.0%
	1.31	Count	9	30	39
		% within Religious	23.1%	76.9%	100.0%
	1.59	Count	9	39	48
		% within Religious	18.8%	81.3%	100.0%
Total		Count	277	527	804
		% within Religious	34.5%	65.5%	100.0%

Figure 11.7 Percent of 10th-graders who are pessimistic versus optimistic for different values of religiosity.

Optim_Pess dichotomous outcome. Note that for values of -1.43 on the religiosity variable, the split on the optimism/pessimism variable was 56.4% pessimistic and 43.6% optimistic. Thus for those with a value of -1.43 Religious, the probability of being pessimistic was .564 versus .436 optimistic. The *odds* of being optimistic for this value of Religious would be $.436/.564$ or $.773$ (or $.773$ to 1). For every one chance of being pessimistic for this value of Religious, there is a $.773$ chance of being optimistic. In contrast, the odds of being optimistic for the highest value of Religious (1.59) are $.812/.188$ or 4.32. Tenth-graders who report this level of religiosity are much more likely to be optimistic as opposed to pessimistic.

But one more transformation is required: we need to go from odds to the (natural) logarithm of odds. The primary reason for this is because log odds have better characteristics as a dependent variable than do simple odds (e.g., log odds can range from large negative numbers to large positive numbers, whereas odds have a lower bound of zero). Thus in LR we focus on the natural logarithm of the odds, known as the log odds. Don't worry, however; in the output there is a translation of the results back into odds ratios. Several of the texts I have cited elsewhere in this book provide more information about the calculation of odds, odds ratios, and log odds (Darlington, 1990; Darlington & Hayes, 2017; Thompson, 2006). For our purposes here, just consider that we have made a transformation in the dependent variable; in LR, we have transformed the dependent variable from a dichotomous outcome into the log odds of that outcome.

Conducting the Logistic Regression and Understanding the Output

Of course the actual logistic regression is conducted using a logistic regression command in your statistics program, and you conduct logistic regression in the same general way as multiple regression, by regressing a dichotomous dependent variable on one or more (dichotomous or continuous) independent variables. Figure 11.8 shows some of the SPSS output for such an analysis. The first table shows the number of cases used in the analysis. The next shows how the dependent variable is coded for the analysis. Here, the value we coded as zero (pessimistic) is, in fact, coded as zero for the analysis, whereas the value coded one (optimistic) is treated as one in the analysis. This table is worth checking to make sure values are indeed coded in the way expected; things could be reversed from what you expect, which will likely result in the coefficients being reversed from what you expect as well.

The next section of the output (Figure 11.8) is labeled Block 0: Beginning Block. LR operates sort of like sequential regression, by adding predictors in blocks. The difference is that Block 0 is the logistic regression with no predictors included, just the intercept. Like structural equation modeling (the focus of Part 2 of this book), LR uses maximum likelihood estimation as opposed to least squares estimation (see Chapter 3 for more on least squares). It also uses a chi-squared test at each block of the regression to determine whether each block of the regression improves prediction (as opposed to an F test associated with ΔR^2). Since the purpose of this section is to try to understand what happens in LR from a MR orientation, we will not expand on these aspects of LR here. Instead, think of LR as doing something like sequential MR, and note that we are going to be focused on a different statistic for understanding the statistical significance of the regression. Thus Block 0 sets the baseline for the success of the prediction for subsequent comparisons.

Block 0 also shows the success of the prediction using only the overall probability of being pessimistic versus optimistic. The "Classification Table" calculates a predicted probability for every case for being in group 1 (optimistic). If that probability is greater than .5, the case is predicted for group 1 (optimistic); if less than .5, it is predicted for group 0 (pessimistic). With no predictors in the model, this probability is based only on the overall numbers of students in the sample who are pessimistic versus optimistic. There are 527 out of 804 students

Logistic Regression

Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	804	80.4
	Missing Cases	196	19.6
	Total	1000	100.0
Unselected Cases		0	.0
Total		1000	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.00 Pessimistic	0
1.00 Optimistic	1

Block 0: Beginning Block

Classification Table^{a,b}

Observed		Predicted			Percentage Correct	
		Optim_Press Optimistic or Pessimistic		.00 Pessimistic		
		.00 Pessimistic	1.00 Optimistic			
Step 0	Optim_Press Optimistic or Pessimistic	.00 Pessimistic 1.00 Optimistic		0 0	.0 100.0	
	Overall Percentage				65.5	

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	.643	.074	75.111	1	.000	1.903

Figure 11.8 Selected output for the logistic regression of pessimism vs. optimism on four predictors. Output from the initial block of the LR with no predictors in the equation.

who are optimistic, so the probability of being optimistic (based on no other information) is $527/804 = .655$. This value is greater than .5, so as shown in the table, at Block 0 the LR predicts that everyone is optimistic. In other words, if all you knew was that 66% of students in your sample were optimistic, your best bet for any student would be that they were optimistic (in the absence of other information).

The “Variables in the Equation” is read just like the table of coefficients in MR. There is a b , and it is tested for statistical significance. The b is also converted into another coefficient that may be easier to interpret. We will postpone looking at the details of this table until we have some predictor variables in the equation.

Figure 11.9 shows the first portion of printout from Block 1 of the LR. For this regression I entered all of the predictors in a single block (i.e., a simultaneous LR). It

Block 1: Method = Enter

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	105.133	4	.000
	Block	105.133	4	.000
	Model	105.133	4	.000

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	930.408 ^a	.123	.169

a. Estimation terminated at iteration number 4
because parameter estimates changed by less than .001.

Figure 11.9 Logistic regression, block 1, with all predictors in the equation.

is also possible to enter variables in blocks (sequential LR) or a stepwise fashion using a variety of criteria (not generally recommended for explanatory research; see Chapter 5). The “Omnibus Tests of Model Coefficients” shows the χ^2 (actually the change [decrease] in χ^2 , $\Delta\chi^2$) associated with the overall model (Model) and with this block (Block) of variables. The $\Delta\chi^2$ is tested for statistical significance, $\Delta\chi^2 = 105.133$ with 4 df for the four predictors, and the probability of this decrease in χ^2 happening by chance alone is $< .001$. This table is thus analogous to testing the R^2 for statistical significance in ordinary MR. If a sequential approach were used, the $\Delta\chi^2$ values shown for the overall model (Model) and the Block would differ on block 2; if a stepwise approach were used, the values for step would differ from those associated with Model for block 2 and beyond. With either of these approaches, the statistical significance of the associated $\Delta\chi^2$ would be analogous to the statistical significance associated with ΔR^2 in a sequential or stepwise regression. The $-2 \log$ likelihood value in the “Model Summary” table is the genesis of the χ^2 values shown in the previous table. If you were to conduct a sequential LR and subtract the $-2 \log$ likelihood from one block to the next you would see that these match the χ^2 associated with that block. Why a $-2 \log$ likelihood? Recall that we are testing the natural logarithm of the odds of pessimism/optimism. That’s where the “log” comes from. The negative is there to reverse what would all be negative numbers. The 2 turns this into a χ^2 distribution. Conceptually, from a MR orientation, think of the log likelihood as being analogous to the R^2 (or sums of squares) from regression (except that smaller values are better and we try to reduce it rather than increase it), and the χ^2 test as analogous to the F test of the R^2 . This table also includes two estimates of R^2 . LR does not use least squares estimation and therefore does not produce a measured of variance explained compared to the variance unexplained. But those of us with a background in MR like some sort of estimate of the proportion of explained variance, so two are produced in SPSS. There are others; as can be seen in the table, they do not always produce consistent estimates of R^2 , and thus should be used cautiously.

The table “Variables in the Equation” in the lower part of Figure 11.10 looks quite familiar to those used to regression output, and it is interpretable in a similar way. The b ’s shown in the second column represent the amount of change in the dependent variable for each one unit change in each independent variable. Just as in MR, these coefficients can be used to

Classification Table^a

Observed		Predicted			Percentage Correct	
		Optim_Press Optimistic or Pessimistic		.00 Pessimistic		
		.00 Pessimistic	1.00 Optimistic			
Step 1	Optim_Press Optimistic or Pessimistic	.00 Pessimistic	1.00 Optimistic	87	190	31.4
				62	465	88.2
	Overall Percentage					68.7

a. The cut value is .500

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	Substance	-.249	.102	5.906	1	.015
	Religious	.481	.095	25.477	1	.000
	byses	.497	.122	16.722	1	.000
	bytests	.036	.010	11.638	1	.001
	Constant	-1.137	.543	4.380	1	.036

a. Variable(s) entered on step 1: Substance, Religious, byses, bytests.

Figure 11.10 More output from Block 1 of the logistic regression, including a table of coefficients (lower table) and a table showing how well the LR equation would perform in predicting group membership (upper table).

create an equation for predicting the dependent variable from the independent variables. In this case, it is:

$$\begin{aligned} \text{logodds Optimism} = & -1.137 - .249\text{Substance} + .481\text{Religious} \\ & + .497\text{BYES} + .036\text{BYTests}. \end{aligned}$$

Recall that the dependent variable is in log odds units, so these coefficients and this equation are not as useful as we might wish. The standard errors are used to test the statistical significance of each variable with all the others controlled, using the Wald test and the formula $(\frac{b}{SE_b})^2$; they can also be used to create confidence intervals around the coefficients.

Unlike a *t* test from multiple regression, these Wald tests each have 1 *df*. The probabilities are interpreted in the same fashion, however, and these show that each of our four predictors was statistically significant.

The final column shows the exponentiated version of the *b*, a conversion that makes it more easily interpretable. For this interpretation, the dependent variable optimistic/pessimistic is still in odds format but is no longer logarithmically transformed. Thus interpretations are based on the odds of being optimistic. One-to-one odds would represent an equal chance of being pessimistic or optimistic. Values greater than one indicate that increases in that independent variable increase the odds of going from pessimistic to optimistic, whereas values less than one indicate that increases in that independent variable lead to (or predict) lower odds for the dependent variable. Focus on substance use and religiosity. The value of 1.617 for Religious means that, other things being equal, a one point increase in religiosity results in an increase in the odds of being optimistic by 1.617. Alternatively, we could say that this 1-point increase would increase the odds of being optimistic by 61.7%. Other things being equal, religious faith and attendance seems to result in increased optimism about the future.

The .780 value for Substance is slightly harder to interpret, but the fact that it is less than one means that increases in substance use predict *decreases* in the odds of being optimistic, and thus increases in the chances of being pessimistic. A one point increase in substance use thus results in a 22% ($1 - .780 \times 100$) decrease in the odds of being optimistic. Alternatively, we could use this odds ratio of less than 1 to calculate the probability of being pessimistic rather than optimistic by reversing it ($1/.780 = 1.282$). Then our interpretation would be that a 1-point increase in substance use results in a 28% increase in the odds of being *pessimistic*. To round out our interpretations, the coefficients suggest that BYTests and BYSES also have positive and statistically significant effects on the odds of being Optimistic. A 1-point increase in SES increases the odds of being optimistic by 64%, and a 1-point increase in test scores increases those odds by 3.6%. Keep in mind that these are still unstandardized coefficients, so you need to keep in mind that a “1-point increase” means something different for each of these independent variables. As a review of Figure 11.2 will show, the variables BYSES, Substance, and Religious were all means of z-scores, so a 1-point increase in any of these variables is a much larger change than a 1-point increase in BYTests, which was a mean of *T*-scores. The output shows no analog to β , that is, no standardized regression coefficients, so we can’t really compare the magnitude of the coefficients to one another. See Thompson (2006, p. 413) for suggestions of ways to produce something akin to standardized coefficients in LR by standardizing the independent variables; he also suggested the use of structure coefficients in LR.

The final piece of output we will address is shown at the top of Figure 11.10, the “Classification Table.” This LR regression equation is used to predict group membership (pessimistic versus optimistic) and these predictions are compared to each participant’s actual categorization as optimistic versus pessimistic. As shown in the final column of the Table, 68.7% of those in the sample were correctly classified, although the equation was more accurate in classifying those who were optimistic than those who were pessimistic. This represents a slight improvement in classification with no predictors in the model (Figure 11.8, 65.5%).

Categorizing a Continuous Variable

Let’s revisit the beginning of this example, in which I took a continuous variable, an Optimism composite, and turned it into a categorical Pessimistic/Optimistic one. The reason for doing so was to have an interesting and understandable example. But is this a good idea? In a word, no. Early in the text I argued against turning a continuous variable into a categorical one, and I have reiterated that admonition elsewhere (e.g., Chapter 8). Let me reiterate it yet again: it is generally not a good idea to turn a continuous variable into a categorical one. I have done so here for the sake of an example, but that does not mean it is a good idea.

Note the output shown in Figure 11.11, the results of a multiple regression using the original *continuous* Optimism variable. It is reassuring that the same variables are statistically significant, and that the coefficients are in the same direction (Substance is negative, all others are positive) in this analysis as in the MR for the dichotomous Optimism variable, and as in the LR for the dichotomous Optimism variable. Note, however, the model summary table in Figure 11.9 as compared to Figure 11.3 (the MR for the dichotomous outcome). With the continuous dependent variable, $R = .370$, and the variance explained (R^2) was .137. In contrast, with a dichotomous outcome, $R = .352$ and $R^2 = .124$. Categorizing a continuous variable reduces—throws away! wastes!—its variance and reduces its correlation with any other variable, thus reducing its R and R^2 . This reduction in correlation becomes larger the farther we get from a 50/50 split. Thompson calls such categorization “data mutilation” (2006, pp. 386–390), and I agree (see also Cohen, 1983).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.370 ^a	.137	.133	.54777

a. Predictors: (Constant), bytests 8th-grade achievement tests (mean), Religious, Substance, byses SOCIO-ECONOMIC STATUS COMPOSITE

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error			
1	(Constant)	3.707	.134		27.747	.000
	Substance	-.057	.026		-2.213	.027
	Religious	.139	.023		6.024	.000
	byses SOCIO-ECONOMIC STATUS COMPOSITE	.131	.029		4.540	.000
	bytests 8th-grade achievement tests (mean)	.008	.003		3.258	.001

a. Dependent Variable: Optimism

Figure 11.11 Multiple regression using the original, continuous Optimism dependent variable. Compare these results to those in Figure 11.3 to see the cost of dichotomizing a continuous variable.

In my experience, such dichotomization is a disappointingly common occurrence in logistic regression research: researchers take perfectly good continuous dependent variables and turn them into dichotomous ones. This practice, again in my limited experience, seems especially common in medical and diagnostic research, where, for example, scores on a measure of depressive symptomology are categorized into not depressed/depressed, or infant birth weights into adequate/low birth weight. Such categorization seems reasonable given the seemingly related categorization that is often necessary in applied practice, such as deciding whether a patient is depressed or not. The resulting research on predictors or influences on these outcomes can thus help the physician predict, for example, which patients are likely to have babies with low birth weight, or the psychologist which clients are likely to be depressed. But such predictions would be just as easy, and likely more accurate and more valid, if done using the continuous as opposed to the categorical outcome. If the prediction equation is really being used, why not predict a continuous birth weight from the variables of interest and then, based on that prediction, flag values that are below a certain level? My point is that it generally does not make sense to categorize a continuous variable for analytic purposes. It may make sense to categorize a continuous variable after analysis as an aid in interpretation (also discussed in Chapter 8). Unless there are compelling reasons for doing otherwise, leave your variables in continuous format!

Appropriate Uses of Logistic Regression

My diatribe concerning categorizing a continuous variable should not give the impression that logistic regression is rarely advised. I discourage the categorization of continuous variables, but there are many naturally occurring categorical variables that are legitimate possible outcome variables in our research. LR is indeed an appropriate analytic choice when

we wish to predict or explain such outcomes. We may wonder why some students drop out of a class or of school while others continue, and how we can predict such an outcome. Some adolescents choose to smoke or to use drugs, while others do not. Some people with mental health concerns seek counseling, others seek help from family, whereas others do not seek help at all. Some former prisoners manage to stay out of prison, whereas others do not. All of these examples involve natural categorical outcomes and would be appropriate for LR; you can no doubt think of others in your own area of study.

Logistic Regression Versus Discriminant Analysis

There is another, older method for analyzing categorical dependent variables in MR fashion: discriminant analysis. With a dichotomous outcome, discriminant analysis is mathematically equivalent to ordinary multiple regression (Cohen et al., 2003), although the output looks somewhat different. Logistic regression is the more popular method at the current time, in part because, as shown here, much of what you have learned about multiple regression is directly applicable to logistic regression. Logistic regression also has an advantage over discriminant analysis in that it can include both categorical and continuous variables as independent variables, whereas, strictly speaking, discriminant analysis should include only continuous independent variables. Logistic regression also requires fewer and more reasonable assumptions. Discriminant analysis had been a better choice for categorical variables that included more than two categories, but up-to-date logistic regression programs can also handle polytomous, in addition to dichotomous, categorical variables. The *Sage Quantitative Applications in the Social Science* series includes good introductions to both methods (Klecka, 1980; Menard, 1997), as do Pituch and Stevens (2016). The several references already mentioned provide useful introductions to LR (Darlington, 1990; Darlington & Hayes, 2017; Thompson, 2006), and the text by Hosmer, Lemeshow, and Sturdivant (2013) provides more depth. The UCLA statistics help websites are also very useful (e.g., “FAQ: How do I interpret odds ratios in logistic regression?”).

MULTILEVEL MODELING

One of the assumptions of multiple regression briefly discussed in the last chapter is that the observations are drawn independently from the population. One way this assumption can be violated is for some observations to be related to one another, to overlap or cluster in some way. Think, for example, about the NELS data. The design of the full NELS survey was to select schools from a national list and then to select, at random, approximately 24 students per school for inclusion in the sample. We have treated the NELS participants as if they were unrelated; but if you think about it, you will probably expect that students will be somewhat more similar to other students within their same school than to students from other schools. This similarity probably becomes stronger when we focus on variables (like homework?) that may be controlled, in part, by the schools. What this means is that the NELS observations are not quite as independent as we would like, which, in turn, may deflate the standard errors of the regression coefficients and make variables seem statistically significant when they are not. To use a more striking example, suppose we were to regress a measure of marital satisfaction on variables such as age, educational attainment, and occupational status. Imagine, however, that we have collected data from couples—both members of every couple. Married couples are likely more similar to one another than are two strangers on all these characteristics, and thus these observations are not independent. One method of dealing with such problems is through a method known as hierarchical linear modeling (HLM) or, more generally, multilevel modeling (MLM). MLM is a regression method that can take into account data that are clustered in some way—students in schools, people in couples, and so on. To use the NELS

example, with MLM we could examine the effects of homework on achievement at both an individual level and school level. In addition to dealing with the problem of lack of independence of observations, multilevel models can also provide a richer understanding of how group-level variables can affect individual-level variables.

Effects of SES on Achievement

Let's consider a simple example, focused on the effects of SES on student achievement. We have used SES or some component of SES (e.g., parent education level) as a student level variable fairly consistently in this text. But if you think about schools near you, I'll bet that SES is a school-level variable also. That is, in many communities, there are higher-SES schools and lower-SES schools. And most parents assume that the general achievement level is higher in the high-SES schools (if you don't believe this is true, ask real-estate agents the kinds of questions they get about schools and school districts!). If you quizzed these parents a little further, I'll bet you would find that they believe, or at least hope, that their children will achieve at a higher level if they attend a higher-SES and higher-achieving school. Thus we have several possible hypotheses embedded in this thinking: that school-level SES may affect school-level achievement and that school-level SES may also affect individual-level achievement. How could we test such hypotheses?

The first requirement to test this speculation is that we have data capable of doing so, that is, measures of SES and achievement from *multiple* students within *multiple* schools. Although the NELS data we have been using throughout the text would seem to fit this bill, it does not because it is a random subsample of 1,000 students selected from larger NELS data. As a result, for most schools represented in the subsample we only have one or maybe two students. The larger NELS dataset, however, does indeed fulfill this basic requirement, and included data on an average of 24 students selected at random from each of the 1,000 schools in the data set. Thus I will use a different subsample of this original dataset for this illustration. For it, I selected all students from schools with 30 or more students in the larger NELS data set. This resulted in the selection of 4,630 students from 127 schools. The data are on the website ("nels smaller 3.sav" on www.tzkeith.com), and are limited to the variables to be used in these analyses plus a few others.

Next, consider the variables needed to test our hypotheses. Of course we would need a measure of SES (the BYSES variable we have often used), a measure of achievement (BYTests), and a variable that tells us what school each student attends (the SCH_ID variable). Our first hypothesis was that the average SES of the school would affect the average achievement level. To test this hypothesis we would need to create measures of the average SES and achievement by school. In SPSS this is easily accomplished using the AGGREGATE command from the DATA menu, and it is likely just as easily accomplished in other statistics programs. This command will allow you to put these aggregated variables back in the original student-level dataset, or create a new school-level dataset. This first option—the school level variable inserted into an individual-level data set—is often referred to as disaggregation, and the variable (in MLM jargon) as a "contextual" variable. (Hox, Moerbeek, & van de Schoot 2018, chap. 1), a group-level variable (one often derived from an individual-level variable as in this example) thought to influence some individual-level outcome.

Multiple Regression Analysis

Perhaps you could create a school-level dataset and then regress average achievement on average SES? Your results would suggest that school SES had a very powerful effect on school-level achievement ($R^2 = .743$, $\beta = .862$), a much stronger effect than any of our previous

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	50.906	.112		452.514	.000	50.685	51.126
ses_mean average SES for school	8.159	.201	.504	40.643	.000	7.766	8.553
SES_C SES centered by school	3.744	.185	.251	20.217	.000	3.381	4.107

a. Dependent Variable: bytests 8th-Grade Achievement

Figure 11.12 Multiple regression of school-level and within-school SES on individual student achievement.

analyses examining the effect of individual SES on individual achievement (with the current data, $\beta = .525$). Curious.

Our second hypothesis was that school-level SES should affect individual level achievement. To follow our current thinking, perhaps in the student-level file you could regress student achievement on school-level average SES? And for good measure, why not also include individual-level SES? This analysis would presumably tell you the effect of the school-level SES as well as the student's own SES on students' achievement. The table of coefficients for such an analysis is shown in Figure 11.12. This result would seem to suggest that school-level SES has a strong effect on individual achievement ($\beta = .504$, $b = 8.16$), and that a student's own level of SES also has a statistically significant, although smaller effect ($\beta = .251$, $b = 3.74$). Note a slight deviation from my explanation in this table: instead of using the original (individual-level) SES variable in this regression, I used a centered version (SES_C). To create this variable, each student's school SES (ses_mean) was subtracted from his or her individual SES. This centering was done in an effort to separate the SES variable into school-level and individual-level components. As a result of this centering, SES_mean and SES_C were uncorrelated. If the original SES variable had been used, its correlation with the SES_mean variable would have been fairly high, .675.

These two analyses seem to answer our questions, and conceptually, at least, they will help us understand MLM. There are problems, however, both conceptual and statistical, and these are well-cataloged in most texts on MLM (Hox et al., 2018; Raudenbush & Bryk, 2002). Let me illustrate one such problem using the individual-level analysis.

Separate Regression Lines by School

When we conduct an ordinary regression analysis on individual-level data (ignoring school-level effects), we assume that the school-level regression lines are the same for every school in our analysis. But doesn't it make sense to assume that some schools are simply better than others? And that they will raise all students' achievement, whether they are of high or low SES backgrounds? If so, this should show up as a higher intercept if we were to conduct separate regressions for every school. And doesn't it also make sense to think that some schools should be more effective at breaking the relation between SES and achievement? That is, shouldn't some schools be particularly effective in working with lower-SES students, whereas others are more successful with higher-SES students? If this were the case, we should see this possibility play out as different slopes when we conduct separate regressions by school; schools more effective in breaking the SES to Achievement effect should show less steep slopes in their regression lines.

Figure 11.13 shows such regression lines. For this graph I selected out 428 students in 12 schools to make the graph readable. First note the heaviest dark regression line that starts about 32 on the Y axis. This is the regression line from a standard regression in which

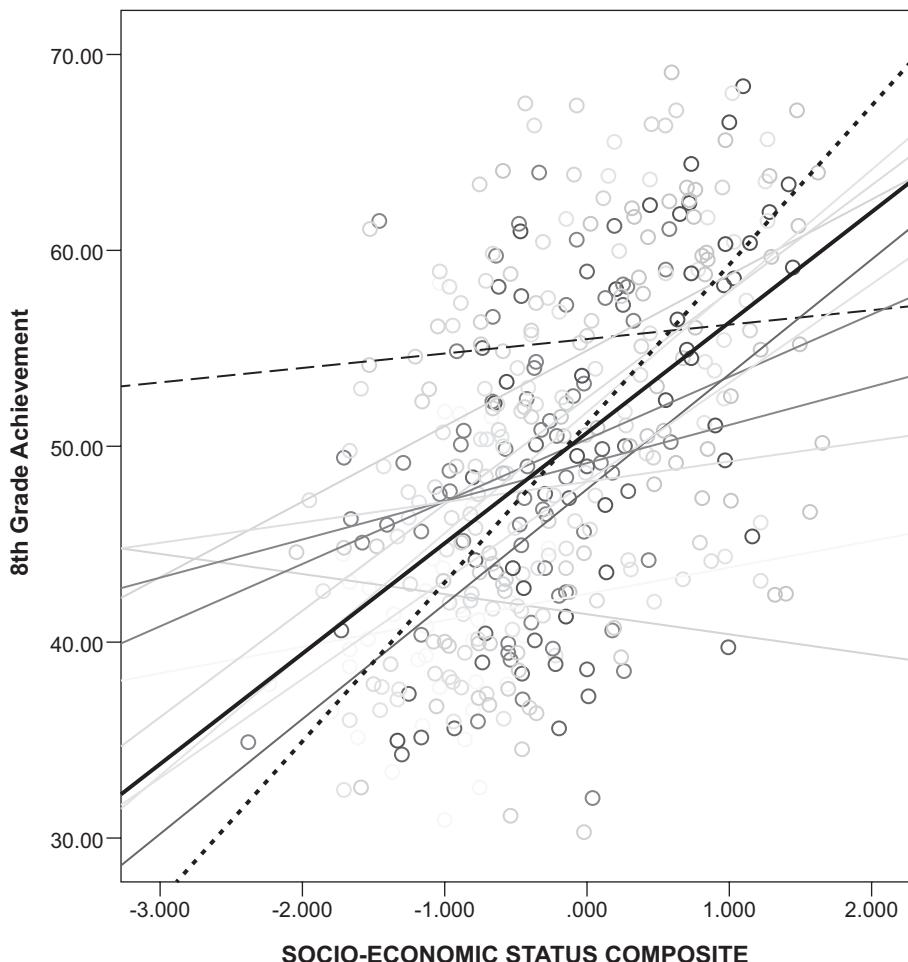


Figure 11.13 The overall regression of 8th-grade achievement test scores on SES is shown by the heavy solid line. The other regression lines are for the regressions of Achievement on SES for individual schools.

Achievement was regressed on SES for all students, without regard for their school. Think of it as the average regression line, in some sense of the word. The other regression lines show the regression of Achievement on SES separately for each school. Note how variable they are; some are indeed quite steep (showing a strong relation between SES and Achievement), whereas some are quite flat, suggesting that SES is much less important for achievement in those schools. And note the considerable differences in the intercepts (the level of Achievement for those with a value of zero on the original SES variable). It looks like some schools are simply better in producing Achievement than others. Note in particular the dotted regression line that starts at around -2.9 on the X axis. This is a very steep line, representing a school in which SES and Achievement are very much related. This is probably the school you would like your children to attend if you were a high-SES parent! Note next the dashed regression line that starts at about 54 on the Y axis. This school seems to show a much weaker relation between SES and Achievement; if I were a low-SES parent, this would be a school I'd want my child to attend. Note finally that the intercepts for these two schools are almost equal (in this graph the intercept—where the lines cross zero on the X axis—is

near the middle of the X axis). Either would work well, it appears, for mid-level-SES family. If nothing else, the graph certainly does seem to suggest the importance of examining the within-school regressions!

Different Slopes for Different Folks?

Think for a minute: where have we looked at graphs like this before? And when did we talk about the importance of centering before? Reward yourself with a treat if you answered “when we talked about interactions.” Yes, what this graph shows is an interaction between school and SES in their effect on achievement. This is the final advantage of MLM, the ability to model interactions between level 2 (in this example, school-level variables) and level 1 (individual-level) variables in their effect on outcomes. And while we can do so in ordinary multiple regression, we can do so better with MLM.

First, take look at how this might be accomplished via regression. Figure 11.14 shows some of the results from a regression of student-level achievement (BYTests) on school-level SES (SES_Mean), individual-level SES (SES_C, centered around each student’s school SES), and a cross-product of the two (sesM_by_sesC). Recall that we previously used cross-products to test for interactions, also known as moderation. The results suggest that both school-level SES and individual-level SES are important in predicting student achievement. The non-significant finding for the sesM_by_sesC cross-product term suggests that these two variables do not interact, however, in their effect on achievement. Said differently, school-level SES does not moderate the effect of individual-level SES on achievement. Or, despite the apparent variability in regression lines, it appears that the best summarization of these data would be a series of regression lines all with the same slope.

As an aside, the graph in Figure 11.13 does not correspond exactly to the regression. Figure 11.15 shows a graph that gets closer to the regression. On the X axis is SES centered by school, and the lines represent school-level SES (as opposed to schools, as was done in Figure 11.13). The results are quite similar, however, including the overall regression line as well as the dashed and dotted lines we focused on previously.

Multilevel Analysis of the Effect of SES on Achievement

Next, let’s turn to a MLM analysis of this same problem. MLM used to require specialized software, and those, including HLM (Hierarchical Linear Modeling, Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011), and MLwiN (Rasbash, Steele, Browne, & Goldstein, 2017) are still excellent, regularly updated options. Unsurprisingly, the major statistical programs have also added MLM features, so that it is possible to conduct MLM via SAS’s PROC MIXED and SPSS’s

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	50.905	.113		452.480	.000	50.685	51.126
ses_mean average SES for school	8.160	.201	.504	40.642	.000	7.766	8.554
SES_C SES centered by school	3.725	.188	.250	19.825	.000	3.357	4.093
sesM_by_sesC	-.230	.382	-.008	-.602	.547	-.979	.519

a. Dependent Variable: bytests 8th-Grade Achievement

Figure 11.14 Multiple regression of individual-level Achievement on school-level and individual-level SES, plus a cross-product (interaction) of the two.

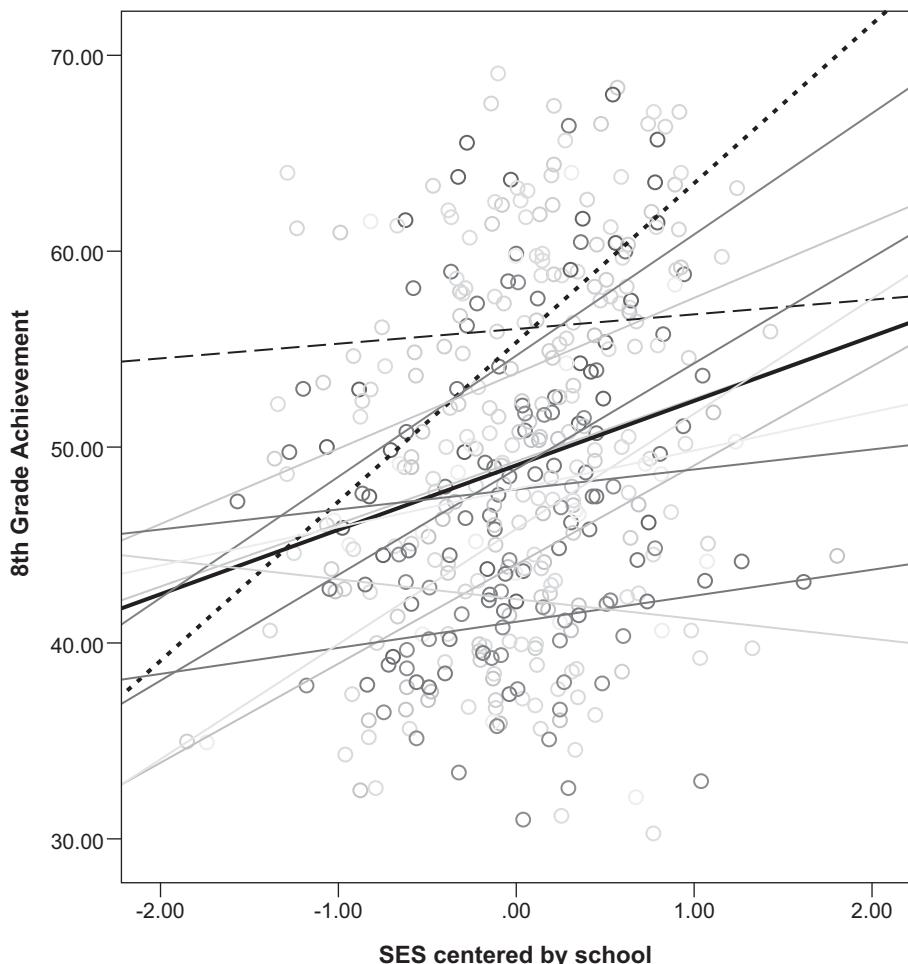


Figure 11.15 Another scatterplot showing both the overall regression of Achievement on SES (heavy solid regression line) and separate regressions by school. This graph uses the SES_c variable. Thus the zero-point on the X axis here represents the mean SES of each school. For the graph in Figure 11.13, the zero represented the mean of SES for all students in the NELS dataset.

Mixed procedures. The Mplus computer program (Muthén & Muthén, 1998–2012) for structural equation modeling (discussed in Part 2 of this book) will also conduct MLM, including latent variable MLM. The output in this chapter is from the SPSS Mixed procedure. Given that this is not a tutorial on how to conduct MLM but rather a bridge to understanding MLM from a regression orientation, I will not present the output in the depth that we have used up to this point.

Figures 11.16 through 11.19 show the results of series of MLM analysis designed to determine the prediction of school-level and individual-level SES on student achievement. The figures also show the syntax used to conduct the analyses. MLM analyses are commonly conducted in this sort of sequential fashion, gradually adding level 2 (in this case, school-level) and level 1 (individual-level) predictors and their interactions. In addition, although not shown in these figures, because MLM analysis, like logistic regression, uses maximum-likelihood estimation, “fit indices” are also shown, including the $-2 \log$ likelihood, as we encountered in logistic regression. As with logistic regression, the $-2 \log$ likelihood can be

```
MIXED bytests
/print=solution testcov
/method=ml
/fixed=intercept
/random=intercept | subject(sch_id).
```

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	50.786663	.459064	126.203	110.631	.000	49.878203	51.695123

a. Dependent Variable: bytests 8th-Grade Achievement

Estimates of Covariance Parameters^a

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Residual Intercept [subject=sch_id]	56.301644	1.211400	46.477	.000	53.976711	58.726719
Variance	24.779558	3.330467	7.440	.000	19.040960	32.247665

a. Dependent Variable: bytests 8th-Grade Achievement

Figure 11.16 Unconditional model (a model with no predictors of BYTests). This model shows the variance in Achievement within schools and between schools.

used to determine whether each new step leads to a better fit, or explanation, of the data by the model. With nested models the $-2 \log$ likelihood values can be subtracted from one another, with the difference equivalent to a χ^2 distribution that can be tested for statistical significance. For now, think of these fit indices as being somewhat similar to the R^2 and ΔR^2 from multiple regression in that they help inform us whether it is worth adding predictors to the model. The fit of models will be discussed extensively in Part 2 of this text, where we will use $\Delta\chi^2$ to compare models. More detail concerning maximum likelihood estimation will also be presented.

Unconditional Model

Figure 11.16 shows the results of a model without any predictors but one in which we specify that we are interested in the school- versus individual-level achievement. This model is often referred to as the unconditional model because the dependent variable is not “conditioned on” (regressed on) any predictors. The reason for this step in the analysis is that it gives us an idea of the degree to which the variance in the dependent variable (BYTests) can be considered between-school variance versus within-school (individual-level) variance. This information is contained in the table labeled “Estimates of Covariance Parameters,” and which includes both covariances (in subsequent models) and variances (think of a variance/covariance matrix). The estimate for the “intercept [subject=school_id]” is the between school variance, and the residual is the remaining, or within-school, variance. This information can be used to calculate the interclass correlation, a ratio of the between group to total variance using the formula $\rho = \frac{V_b}{V_w + V_b}$, or, in this case $\rho = \frac{\text{Intercept}}{\text{Residual} + \text{Intercept}} = \frac{24.780}{56.302 + 24.780} = .306$. This finding means that approximately 31% of the variance in the achievement test scores is between school variance, and that is a lot! A MLM analysis is likely appropriate.

Adding a Level 2 Covariate

Figure 11.17 shows the results of the MLM analysis of school-level achievement regressed on average school-level SES. In MLM lingo, this will often be referred to as adding a level 2 covariate. You can read the table of “Estimates of Fixed Effects” the same way you would a table of regression coefficients in MR. Note that the effect of school-level SES was statistically significant, suggesting that school-level SES is indeed important for the average level of school achievement. Although the MLM results do not include standardized coefficients, we can calculate those using the formula we learned for regression: $\beta = \frac{b \times SD_x}{SD_y}$ (Hox et al., 2018). Although not shown in the Figure, the standard deviations of SES_mean and BYTests were .560 and 9.058, respectively, and $\beta = \frac{8.194 \times .560}{9.058} = .507$, which we would likely classify as a large effect (assuming we used the same criteria that we have used in MR). Let us briefly look at the remainder of the output shown, again with an idea of giving you a way of thinking about MLM interpretation using a MR orientation. The intercept shown in the first table in the figure is the expected mean achievement for a school of average SES (given that SES is centered at zero because it is an average of z-scores). The variances shown in the second table (Estimates of Covariance Parameters) can be compared to the values from the previous analysis with no predictors in the model (Figure 11.16). The residual, the unexplained variance in individual-level achievement, shows the unexplained within-school variance in achievement. This is relatively unchanged from the previous model, because there are no level 1 (individual-level) predictors in the model. In contrast, the school-level residual variance (“Intercept [subject=sch_id]”) is reduced from a value of 24.780 in the previous model to a value of 4.922 in the current model. School-level SES is quite effective in explaining school-level variation in achievement (and recall that school-level variation in achievement accounts for approximately 31% of all the variation in achievement).

```
MIXED bytests with ses_mean
/print=solution testcov
/method=ml
/fixed=intercept ses_mean
/random=intercept | subject(sch_id) .
```

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	50.862076	.228727	124.976	222.370	.000	50.409396	51.314756
ses_mean	8.193784	.419393	121.308	19.537	.000	7.363507	9.024061

a. Dependent Variable: bytests 8th-Grade Achievement

Estimates of Covariance Parameters^a

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Residual Variance	56.305034	1.211488	46.476	.000	53.979932	58.730286
Intercept [subject=sch_id]	4.921789	.827835	5.945	.000	3.539587	6.843739

a. Dependent Variable: bytests 8th-Grade Achievement

Figure 11.17 MLM analysis, regression of school-level achievement on school-level SES.

Adding a Level 1 Covariate

The next step (Figure 11.18) shows the addition to the analysis of a level 1 covariate, that is, the prediction of BYTests by within-school SES (SES_C), in addition to average school SES (SES_mean). Recall that SES_C is the SES variable centered within school. It is, in essence, an individual (within-school) SES measure, with variation in SES across schools removed. As shown in the Estimates of Fixed Effects table, within-school SES (individual-level SES with school-level SES removed) was a statistically significant predictor of individual-level achievement ($b = 3.684$, $p < .001$); achievement increases as students' SES within a school increases. The standardized coefficient for SES_C is calculated as .247: large but not as large as the school-level SES predictor (.510). The Intercept in the top table labeled Fixed Effects (50.852) represents the expected achievement for students whose SES levels are at their school-level mean and who attend a school with average SES.

In the table of Covariance Parameters (lower portion of Figure 11.18), the residual shows the residual variance in BYTests, after accounting for within- and between-level school SES. Compared to the same value in the model shown in Figure 11.16, this value (50.86) represents about a 10% reduction ($50.86/56.30 = .89$) in unexplained variance with the addition of these two predictors. If this were multiple regression we would be discussing this decrease in unexplained variance as an increase in R^2 . This table also includes a set of two variances [UN(1,1) and UN(2,2)] and one covariance [UN(2,1)]. The first variance [UN(1,1)] represents the variance in intercepts, or the variation in achievement means across schools. Think of this as the variation in the height of the regression lines in Figure 11.15. Note that there is considerable (and statistically significant) variation in these intercepts. This finding is not surprising because schools differ in their SES and children within each school also differ in their SES. The second variance [UN(2,2)] represents the slope variance, that is, the variation

```
MIXED bytests with SES_mean SES_C
/CRITERIA=MXITER(500)
/print=solution testcov descriptive
/method=ml
/fixed=intercept SES_mean SES_C
/random=intercept SES_C | subject(sch_id) covtype(UN).
```

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	50.851845	.228039	124.915	222.996	.000	50.400524	51.303166
ses_mean	8.255060	.411689	123.005	20.052	.000	7.440147	9.069973
SES_C	3.683728	.207082	118.132	17.789	.000	3.273653	4.093803

a. Dependent Variable: bytests 8th-Grade Achievement

Estimates of Covariance Parameters^a

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Residual	50.545718	1.101435	45.891	.000	48.432394	52.751255
Intercept + SES_C	5.046839	.823190	6.131	.000	3.665894	6.947986
[subject= sch_id]	1.148305	.540769	2.123	.034	.088418	2.208193
UN (2,2)	1.326492	.637371	2.081	.037	.517258	3.401744

a. Dependent Variable: bytests 8th-Grade Achievement

Figure 11.18 Adding a Level 1 covariate: prediction of Achievement based on school-level SES and within-school SES.

in the slopes in Figure 11.15, or the variation in the influence of student SES on achievement across schools. The results show that there is indeed a statistically significant degree of variation for the slopes across schools (given we are using $p < .05$). Compare this to the coefficient associated with SES_C in the table of Fixed Effects above in this same Figure 11.18. You can think of the fixed effect (3.684) as the average slope across schools when predicting Achievement from within-school SES. The value of the (residual) variance [UN(2,2)] associated with the SES slopes shows that there is significant variation in the slopes across schools, however. In the next step we will add a cross-level interaction (the interaction of school-level and individual-level SES in their effect on Achievement) term to see if that helps explain the variation in slopes across schools. The covariance [UN(2,1)] is not generally interpreted.

Adding the Cross-Product to Test the Interaction of School-Level and Individual-Level SES

The final analysis in this MLM is shown in Figure 11.19, and is, perhaps the most interesting and most closely approximates our multiple regression analysis of these same data. The intercept from the table of Fixed Effects (50.819) represents the expected Achievement for students with average levels of within-school SES and average school-level SES (average for the entire sample and average for their school). The coefficients for SES_mean and SES_C represent the influence (or prediction) of school-level SES on school-level Achievement and the influence of within-school SES on individual Achievement. Both are statistically significant. Table 11.1 shows the standardized coefficients associated with these effects. Both effects are large, although school-level SES appears a better predictor of Achievement than within-school SES. Finally, the coefficient for SES_Mean*SES, representing the interaction of school-level and within-school SES, was not statistically significant. As much as it appears from the scatterplots that the slopes of the regressions across schools are different, this difference in slopes is not statistically significantly different, once school-level and within-school SES are

```
MIXED bytests with SES_mean SES_C
  /CRITERIA=MXITER(500)
  /print=solution testcov
  /method=ml
  /fixed=intercept SES_mean SES_C SES_mean*SES_C
  /random=intercept SES_C | subject(sch_id) covtype(UN) .
```

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	50.851333	.228036	124.934	222.997	.000	50.400020	51.302647
ses_mean	8.212505	.418401	121.619	19.628	.000	7.384211	9.040798
SES_C	3.668397	.208706	124.952	17.577	.000	3.255341	4.081453
ses_mean * SES_C	-.236231	.414750	161.947	-.570	.570	-1.055246	.582784

a. Dependent Variable: bytests 8th-Grade Achievement

Estimates of Covariance Parameters^a

Parameter	Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval		
					Lower Bound	Upper Bound	
Residual	50.543328	1.101331	45.893	.000	48.430203	52.748654	
Intercept + SES_C	UN (1,1)	5.046617	.823085	.6.131	.000	3.665831	6.947494
[subject= sch_id]	UN (2,1)	1.138532	.539929	2.109	.035	.080290	2.196774
	UN (2,2)	1.318330	.635190	2.075	.038	.512746	3.389575

a. Dependent Variable: bytests 8th-Grade Achievement

Figure 11.19 Adding the cross-product of school-level and within-school SES.

Table 11.1 Standardized Coefficients, Calculated From the Unstandardized Coefficients Shown in Figure 11.19 and the SDs of the Variables

Parameter	b	SE	p	β
SES_Mean	8.213	.418	<.001	.507
SES_C	3.668	.210	<.001	.246

controlled. Another way of thinking about this coefficient is that school-level SES explained a non-significant portion of the differences in slopes shown, for example, in Figure 11.15.

How does this figure compare to coefficients obtained using multiple regression (Figure 11.14)? The coefficients, both unstandardized and standardized, are not that different, and in this example at least, they tell the same story: school-level and within-school SES are both important predictors of Achievement. The big difference in the two results is in the standard errors for the coefficients, and these are different primarily because of the sample sizes used in their calculation. This shows up most clearly in the *df* for each analysis. For the multiple regression, the *df* are the same for each coefficient, as is the value for the *df_{residual}*: 4442, calculated without respect to our knowledge that these are students nested within schools. In the MLM, the *df* are calculated separately for each coefficient, and note that they are much smaller than those for the regression. They are also calculated taking into account the nested nature of the data, and are more accurate than those in the MR. As noted in the introduction to this topic, the nested nature of the data affects the standard errors (and thus statistical significance) of effects much more so than the estimates themselves.

The Covariance Parameters table in Figure 11.19 shows that the individual-level residual variance in Achievement [Residual] is still statistically significant, as is the remaining variance in school intercepts [UN(1,1)] and the remaining variance in the school slopes [UN(2,2)]. Note also that there is relatively little reduction in the unexplained variation in slopes [UN(2,2)] from that shown Figure 11.18. Together, these findings suggest that there are indeed different slopes associated with the prediction of Achievement for each school but that school-level SES does not help explain these differences. Presumably there are other variables, not explored, that could help explain this remaining variance (if you would like to pursue this example, try using the variable in the data representing public versus private and Catholic schools as a level 2 predictor, along with a SES_C by Private school interaction term).

Here I have tried, via a simple example, to explain MLM using MR concepts and the jargon we have developed to talk about MR. To help with the transition, I have also introduced some of jargon used by MLM methodologists. MLM discussions will also often refer to fixed versus random effects. Most discussions of MLM refer to the regression equation results (i.e., the intercepts and coefficients from the tables labeled here as “Estimates of Fixed Effects”) as exactly that: fixed effects. In contrast, the information contained in the tables labeled in SPSS as “Estimates of Covariance Parameters” are often referred to as random effects. Think of it this way: the regression results apply to all of the schools in our samples; the value of 3.668 for the SES_C coefficient in Figure 11.19, for example, is the effect of within-school SES averaged across all schools. In contrast, the information in the “Covariance Parameters” are things—intercepts and slopes—that *vary* across schools; they are not fixed, they are random. As we moved from Figure 11.16 through 11.19, we gradually added new predictors to the equation to attempt to explain some of this random variation. Because this short section is presented as a transition to understanding MLM given that you understand MR, I have avoided many important aspects of MLM, such as alternative methods of centering and different types of maximum likelihood estimation, among others.

MLM: Next Steps

I chose this simple example to illustrate MLM for several reasons. First, it is interesting and the underlying reasoning for looking at school-level effects is readily understandable. Second, it is similar to examples used in one of the primary texts for MLM analysis (Raudenbush & Bryk, 2002), although that example uses data from an earlier data set (High School and Beyond, HSB) and also includes a second level-2 covariate, public versus Catholic school. The HSB example from Raudenbush and Bryk is further used to illustrate how to conduct MLM using SAS (Singer, 1998) and SPSS using syntax (Peugh & Enders, 2005). All of these references are excellent resources for those wishing to learn more about MLM. Two other excellent resources are the book by Joop Hox mentioned earlier in this chapter (Hox et al., 2018), and a short text that focuses specifically on the use of SPSS to conduct MLM (Heck, Thomas, & Tabata, 2014). Another useful text by Pituch and Stevens (2016) includes introductory chapters on both MLM and logistic regression.

SUMMARY

This chapter covered two methods that can be considered extensions of multiple regression. Logistic regression is useful when the outcome variable of interest is categorical. The example used four variables (BYTests, BYSES, Substance use, and Religiosity) to predict a categorical pessimism/optimism variable. The example was first analyzed via multiple regression analysis and then via logistic regression. A problem with using MR for this analysis was that such analysis violated many of the assumptions for the method as outlined in Chapter 10. One way of thinking about LR is that it is like MR but with the categorical variable transformed into a metric that avoids these violations. For LR, the categorical dependent variable is transformed into the natural logarithm of the odds of being in one group (optimistic) versus the other (pessimistic).

We saw that the actual output for the logistic regression was similar to that of regression but with some differences. Because LR uses a different method of estimation—maximum likelihood as opposed to least squares—different statistics are used to assess the statistical significance of the regression equation. For LR, the $-2 \log$ likelihood, converted to a $\Delta\chi^2$, was tested for statistical significance and used instead of a R^2 to assess whether prediction of Optimism from the four independent variables was statistically significant, often referred to as the “fit” of the model to the data. The table of coefficients produced by the logistic regression showed whether each variable added statistically significantly to the prediction. The b coefficients from that table were for log odds units, however, and there is no LR analog to the β s from multiple regression. A column showing the exponentiated value of b was our primary focus for interpretation for each independent variable, and these were easily interpretable as odds ratios, that is, the ratio of being optimistic as opposed to pessimistic. So, for example, the value for BYSES of 1.644 could be interpreted as meaning that for each one point increase in SES, the odds of being optimistic are increased by a value of 1.644, or that for each one unit increase in SES the odds of being optimistic (as opposed to pessimistic) increased by 64.4%. The LR output also included a table showing how well the LR equation served in predicting group membership, first with no variables in the equation and then with all predictors. It is possible, of course, to conduct LR in a sequential fashion, adding variables one at a time or in blocks, and assessing changes in model fit with each block, as well as assessing the improvement in prediction with each block. I ended this short presentation with a reiteration of my previous admonition against routinely turning continuous variables into categorical ones for analysis (as I did to illustrate LR). In other words, I suggested “do as I say, not as I do!”

Multilevel modeling (MLM), also known hierarchical linear modeling (HLM), is useful for taking into account the often hierarchical or clustered nature of the data we analyze. Examples of such clustered data include students within schools (as with the NELS data),

individuals within couple groups, or experimental intervention participants within different sites. The MLM example was illustrated using a subset of data from the larger NELS data to examine the effect of school-level versus individual-level SES on academic achievement. In ordinary multiple regression, we would assume that there was one regression equation that applied to everyone in our sample, regardless of school or other possible clustering variables. As a result, our estimates of the intercept and slope for SES in the regression equation would be fixed at the same value, known as a fixed effect. Our examination of scatterplots and regression equations across schools suggested otherwise, however. It suggested that there may be differences in the regression equations of Achievement on SES by school; that is, that the intercepts and slopes of the school-level regressions may differ from school to school. Another way of saying this is that our scatterplots suggested the possibility that a model with random (as opposed to fixed) effects might be more appropriate. This is one major advantage of MLM over ordinary MR: the ability to allow for separate regression equations across the groups on which our participants are clustered. Another advantage of MLM is the ability to separate the effects at different levels, in this case the effect of school-level SES versus individual-level SES on achievement.

Our scatterplots and separate regression lines by school were reminiscent of the follow-ups we conducted in previous chapters for the findings for statistically significant interactions in MR. This, then, is one way making the transition in understanding to MLM from MR: we are testing for possible interactions between variables at different levels in their effects on outcomes. In this case, we were testing for possible differences in slopes and intercepts for different schools: the possible interaction of school-level and individual-level SES on Achievement.

Once this conceptual understanding of what is done in MLM is grasped, it is easier to understand the somewhat different (different from MR) output that results from MLM. Generally with MLM, one adds predictors gradually and, as with LR, examines the fit of the model (using $-2 \log$ likelihood, among others) as a result of these new predictors. In our example, we started by predicting achievement from school-level SES, then added within-school SES (individual-level SES corrected for school level SES), and then added a cross-product of these two predictors. At each step in the analysis, the coefficients associated with the variables in the equation showed up in the table of Fixed Effects, which is similar in look and interpretation to a table of coefficients in MR. As with LR, there are no standardized coefficients presented in MLM, but these can be calculated. Estimates of the residual variation in slopes and intercepts (by school, in this case) showed up in the table of Covariance Parameters. This latter table is useful in determining if there are other variables that could help explain the differences in slopes and intercepts beyond those already in the equation. Taken together, our example suggested that there were indeed differences in intercepts and slopes across schools, and that school-level SES helped explain the differences in intercepts but not in slopes.

For both topics presented in this chapter, my intent was not to provide a detailed explanation of the methodology. Instead, the hope was to provide a way of understanding these related methods using something that you already understand, multiple regression. References were provided for both methods for further study.

EXERCISES

1. Find a research article in an area of interest to you that used logistic regression in the analysis. Was the dependent variable two categories, or more than two? Was it naturally categorical or did the authors categorize a continuous variable? How did they go about adding predictors? Are you able to interpret the results of the LR using the suggestions in this chapter? What aspects of the reported findings are still puzzling to you?

2. Find a research article in an area of interest to you that used MLM. Make sure the research focused on measured variables (a regression orientation) as opposed to latent variables (a structural equation modeling orientation). What program was used to conduct the MLM? What are the level 1 versus level 2 variables? Are there more than two levels? What was the order of model testing (Unconditional model, followed by the addition of what?)? Was it similar to the order used here? Are you able to interpret the results of the MLM using the suggestions in this chapter? What aspects of the reported findings are still puzzling to you?

Part II

Beyond Multiple Regression: Structural Equation Modeling



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

12

Path Modeling

Structural Equation Modeling With Measured Variables

Introduction to Path Analysis	258
<i>A Simple Model</i>	258
<i>Cautions</i>	262
<i>Jargon and Notation</i>	264
A More Complex Example	266
<i>Steps for Conducting Path Analysis</i>	266
<i>Interpretation: Direct Effects</i>	269
<i>Indirect and Total Effects</i>	271
Summary	275
Exercises	278
<i>Notes</i>	279

In this chapter, we continue our journey beyond multiple regression and begin discussing structural equation modeling (SEM). This chapter focuses on the technique of path analysis, which can be considered the simplest form of SEM. Because we used path-type models as a way of displaying and understanding regression models throughout Part 1 of this text, this transition to a formal presentation of path modeling should be a natural extension of our work so far. As you will see, simple path analyses can be solved using multiple regression analysis, although we will soon begin using specialized structural equation modeling software for both simple and complex path models.

In the penultimate chapter of Part 1, we reviewed one of the difficulties with multiple regression analysis, the fact that we can come to different conclusions about the effects of one variable on another depending on which type of multiple regression we use and which statistics from the analysis we interpret. (If you are beginning the book here, I recommend that you read Chapter 10 as a review of multiple regression.) As you will see, this difficulty is obviated in path analysis and structural equation modeling, where it is natural to focus not only on direct effects but also on indirect and total effects (total effects are the sum of direct and indirect effects). We will use both simultaneous and sequential MR in path analysis, an exercise that will clarify the relation between these two methods. In the process, we will focus more explicitly on explanation, and on the issue of cause and effect. I think that path analysis makes many aspects of multiple regression more understandable, and it is often a better choice for the explanatory analysis of nonexperimental data.

Before we begin, let's deal with a little jargon. The general type of analysis discussed in this part of the book, SEM, is also referred to as analysis of covariance structures, or causal analysis. Path analysis, one form of SEM, is the subject of this and the next two chapters; it may also be considered as a component of SEM. Confirmatory factor analysis (CFA) is another component. More complex forms of SEM are often referred to as latent variable SEM, or simply as SEM. SEM is also sometimes referred to as LISREL analysis, which is actually the first computer program for conducting latent variable SEM and stands for *linear structural relations*. We will discuss these and other topics in subsequent chapters, including this and other SEM computer programs. Now we introduce path analysis.

INTRODUCTION TO PATH ANALYSIS

A Simple Model

Let's return to the example we used in Chapter 10, in which we were interested in the effects of Family Background, Ability, Academic Motivation, and Academic Coursework on high school Achievement. For the sake of simplicity, we will focus on only three of the variables: Ability, Motivation, and Achievement. Suppose, then, we are interested in the effects of Motivation on Achievement. Although presumably motivation is manipulable, it is not a variable that you can assign at random, and thus you will probably need to conduct a nonexperimental analysis, as was done in Chapter 10. Intellectual Ability is included in the model to control for this variable. More specifically, we believe that Ability may affect both Motivation and Achievement, and we know that it is important to control for such *common causes* if we are to estimate accurately the effects of one variable on another.

Figure 12.1 illustrates the data we collected. Motivation is a composite of items reflecting academic motivation (student ratings of their interest in school, willingness to work hard in school, and plans for post-high school education); Achievement is a composite of achievement tests in reading, math, science, civics, and writing. We also collected data on Intellectual Ability (a composite of two verbal ability tests), with the notion that ability should be controlled because it may affect both Motivation and Achievement. The curved lines in the figure represent correlations among the three variables. The figure essentially presents the correlation matrix in graphic form. The correlation between Ability and Motivation, for example, is .205. (The data are from the correlation matrix used in Chapter 10.)

Unfortunately, the data as presented in Figure 12.1 do little to inform our question of interest: understanding the effects of Motivation on Achievement. The correlations are statistically

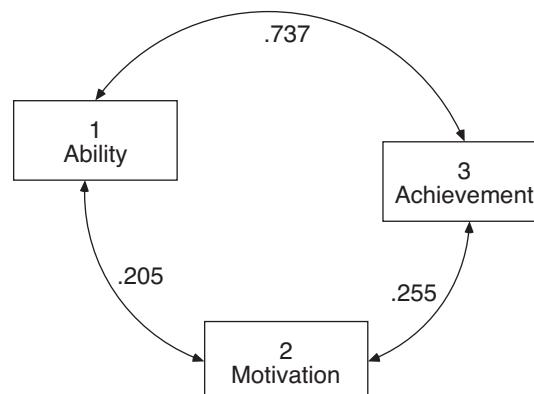


Figure 12.1 Correlations among Ability, Motivation, and Achievement. An “agnostic” model.

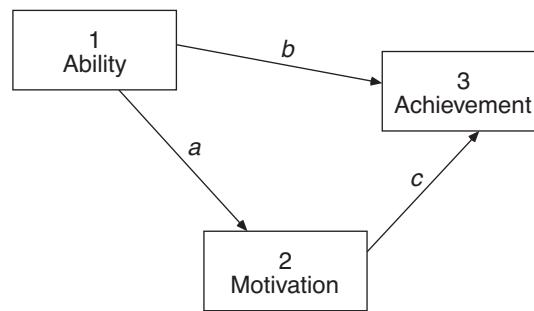


Figure 12.2 Presumed causal structure of the three variables. Note that the assumptions about causal direction were not based on the correlations.

significant, but we have no information on the effects of one on the other. We can think of this figure, then, as an “agnostic” model. In Figure 12.2 we take the first bold step in solving this dilemma by drawing arrows or paths from presumed causes to presumed effects. The purpose of this research was to determine the *effect* of Motivation on Achievement, so it certainly makes sense to draw a path from Motivation to Achievement. Ability was included in the research because we worried that it might *affect* both Motivation and Achievement; therefore, paths drawn from Ability to Motivation and Achievement are the embodiment of this supposition. Our drawing of the paths asserting presumed cause and effect was not so bold after all; it simply made obvious the reasoning underlying our study and the data we collected.

What exactly do these paths mean? They assert what is called a *weak causal ordering*, meaning that the path from Motivation to Achievement does not assert that Motivation directly causes Achievement, but rather that *if* Motivation and Achievement are causally related the cause is in the direction of the arrow, rather than the reverse. Note that we did not use the correlations or the data to make these inferences about causality; instead, our informal causal thinking guided the data we collected and used! Figure 12.2 formalizes our notions of how these three variables are related and thus represents our model of the nature of the relations among these three variables.

The data shown in Figure 12.1 may be used to solve for the paths in the model shown in Figure 12.2. The easiest way to do so is to use the tracing rule: “the correlation between two variables X and Z is equal to the sum of the product of all paths from each possible tracing between X and Z [in Figure 12.2]. These tracings include all possible routes between X and Z , with the exceptions that (1) the same variable is not entered twice per tracing and (2) a variable is not both entered and exited through an arrowhead” (Keith, 1999, p. 82; cf. Kenny, 1979, p. 30). Thus, the correlation between Ability and Achievement (r_{13}) is equal to path b plus the product of path a times path c : $r_{13} = b + ac$. Two other formulas (for the other two correlations) may be derived: $r_{23} = c + ab$ and $r_{12} = a$. You may wonder why the third equation does not include the tracing bc . The reason is that this tracing would violate the second exception (the same variable was entered and exited through an arrowhead).

We now have three equations and three unknowns (the three paths). If you recall high school algebra, you can use it to solve for the three unknowns:¹

$$a = r_{12}$$

$$b = \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2}$$

$$c = \frac{r_{23} - r_{12}r_{13}}{1 - r_{12}^2}$$

(If you don't recall high school algebra, note 1 shows how these three equations were generated.) Substituting the actual correlations in these equations, we calculate

$$a = .205$$

$$b = \frac{.737 - .205 \times .255}{1 - .205^2} = .715$$

$$c = \frac{.255 - .205 \times .737}{1 - .205^2} = .108$$

The solved paths are included in the model in Figure 12.3. The model may be interpreted as demonstrating the effects of Ability and Motivation on Achievement, along with the effects of Ability on Motivation (given several assumptions). The paths shown are the standardized path coefficients and are interpreted in standard deviation units. Thus, the path from Motivation to Achievement of .108 suggests that, given the adequacy of our model, each *SD* increase in Motivation will result in a .108 increase in Achievement.²

If this sounds familiar, it should. This type of interpretation is the same as that for standardized regression coefficients. A closer inspection of the formulas above will show striking similarity to those in Chapter 2 for regression coefficients. In fact, these formulas *are* the formulas for standardized regression coefficients. We don't need to use algebra to solve for the paths; we can use good old multiple regression analysis!

To solve for the paths using multiple regression, regress Achievement on Ability and Motivation. The β 's from this regression are equal to the standardized paths, calculated previously, from Ability and Motivation to Achievement. The path from Ability to Motivation is estimated through the regression of Motivation on Ability. Relevant portions of the output are shown in Figure 12.4. The first table of coefficients is from the first regression and shows estimates for the paths to Achievement; the second table of coefficients is from the second regression and shows the path to Motivation. Compare the results to those shown in Figure 12.3. We could also interpret the *b*'s from these tables as the unstandardized path coefficients.

We can use and interpret that printout and model in the same fashion as we previously did with multiple regression. The model thus suggests that Motivation has a moderate effect (using the rules of thumb from Chapter 4) on Achievement, after taking students' Ability into account.³ Ability, in turn, has a moderate effect on Motivation and a very large effect on Achievement. We can use the rest of the regression output as we have previously. Just as in other forms of MR, the unstandardized regression coefficients—used as estimates of the unstandardized paths—may be more appropriate for interpretation, for example, when the variables are in a meaningful metric. In the present example, the standardized coefficients

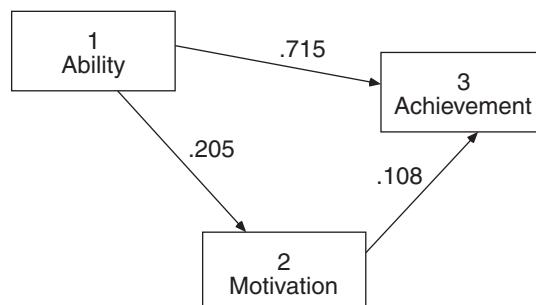


Figure 12.3 We used the data from Figure 12.1 to solve for the paths from Figure 12.2. The paths represent the standardized effect of one variable on another, given the adequacy of the model.

Model Summary

Model	R	R Square	F	df1	df2	Sig. F
1	.745 ^a	.554	620.319	2	997	.000

a. Predictors: (Constant), MOTIVATE, ABILITY

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-3.075	1.627			-6.267	.118
	ABILITY	.477	.014	.715	33.093	.448	.505
	MOTIVATE	.108	.022	.108	5.022	.066	.151

a. Dependent Variable: ACHIEVE

Model Summary

Model	R	R Square	F	df1	df2	Sig. F
1	.205 ^a	.042	43.781	1	998	.000

a. Predictors: (Constant), ABILITY

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	36.333	2.089			32.235	40.432
	ABILITY	.137	.021	.205	6.617	.096	.177

a. Dependent Variable: MOTIVATE

Figure 12.4 Using simultaneous multiple regression to solve the paths.

are probably more interpretable. (You may wonder why the unstandardized and standardized paths from Motivation to Achievement are the same. The reason is because the *SDs* for the two variables are the same.) In addition, we can use the *t*'s and standard errors from the output to determine the statistical significance of the path coefficients, as well as confidence intervals around the paths. The 95% confidence interval around the (unstandardized) path from Motivation to Achievement was .066 to .151.

The model shown in Figure 12.3 is not entirely complete. Conceptually and statistically, it should be clear that the model does not include all influences on Achievement or Motivation. You can no doubt think of many other variables that should affect high school achievement: family background, coursework, homework, and others. And what about effects on Motivation; if Ability only affects Motivation at a level of .205, obviously many influences are unaccounted for. The model shown in Figure 12.5 rectifies these deficiencies by including “disturbances” in the model, symbolized as d1 and d2. Disturbances represent *all other* influences on the outcome variables other than those shown in the model. Thus, the circled variable d2 represents all influences on Achievement other than Ability and Motivation. The disturbances are enclosed in circles or ellipses to signify that they are *unmeasured* variables. We obviously don't measure all variables that affect Achievement and include them in the model; the disturbances, then, are unmeasured, rather than measured variables.

When I say that the disturbances represent all other influences on the outcomes besides the variables in the model, this explanation may ring a bell, as well. You might think that the

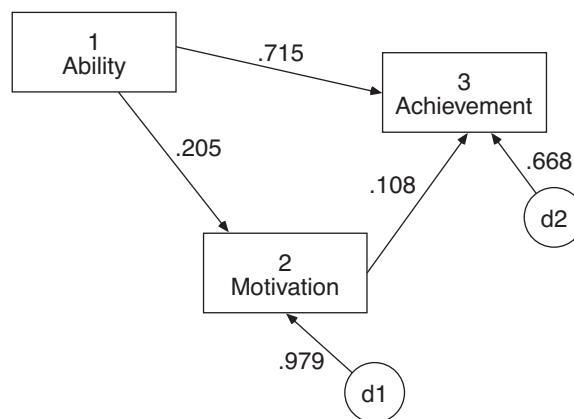


Figure 12.5 The full, standardized, solved model, including disturbances of the presumed effects. Disturbances represent all other, unmeasured variables that affect a variable other than the variables already pointing to it.

disturbances should somehow be related to the residuals, which we at one point described as what was left over or unexplained by the variables in the model. If you had this sense, then reward yourself with a break or a chocolate, because the disturbances are basically the same as the residuals from MR. You have probably encountered instances in research and statistics where two different names are used to describe the same concept; this is another instance of this practice. Although many sources use the term disturbances to describe these other influences (e.g., Bollen, 1989; Kenny, 1979), others continue to use the term *residual*, and others simply refer to these outside influences as *errors*. The paths associated with the disturbances are calculated as the square root of $1 - R^2(\sqrt{1 - R^2})$ from each regression equation. Focus again on Figure 12.4. For the first equation, the regression of Achievement on Ability and Motivation, R^2 was equal to .554, and thus $\sqrt{1 - R^2} = .668$, the value shown for the path from d2 to Achievement. Take a moment to calculate the disturbance for Motivation.

Cautions

With all this talk of cause and effect, you may feel a little queasy. After all, aren't we here breaking the one cardinal rule of elementary statistics: Don't infer causation from correlations? If you are having such misgivings, I first urge you to revisit the short quiz on this same topic in Chapter 1. Second, I point out that, no, we did not infer causality from the correlations. Yes, we had the correlations, but recall that they did not lead to or even enter into our causal inferences. We made the inference of causality when we drew paths from one variable to another, and we drew these paths *without* reference to the correlations. Neither the magnitude nor the sign (positive or negative) of the correlations entered our consideration of cause and effect.

How did we, and how could we, make these inferences of cause and effect? Several lines of evidence can be used to make such inferences and thus to draw the paths. First is *theory*. School learning theories generally include both motivation and ability (or some similar construct) as influences on academic achievement and thus justify the paths from Ability and Motivation to Achievement (Walberg, 1986). And even when formal theory is not available, informal theory can often inform such decisions. Talk to an observant teacher and he or she will tell you that if you can increase a child's level of motivation his or her achievement will likely increase.

Second, we should attend to *time precedence*. As far as we know, causality cannot operate backward in time and so, if we can establish that one variable occurs prior to another in time, it makes it easier to draw the path. This is one reason that longitudinal data are so valued in research; we can feel more confident about inferring cause and effect when our “effect” is measured after our “cause.” Yet even with cross-sectional data it is often possible to determine logical time precedence. In the current example, it is well known that ability is a relatively stable characteristic, for most people, from about the time children start school. Logically, then, Ability, stable from an early age, occurs prior to high school motivation and achievement, and thus it makes sense to draw a path from Ability to both Motivation and Achievement. For an even more striking example, consider if we had the variable Sex in our model. For almost everyone, biological Sex is stable from conception on. Thus, no matter when Sex is measured, we can feel confident placing it prior to variables that logically occur after conception.

Third, you should have a competent understanding of the relevant research. Previous research may well highlight the proper causal inference. Even if it doesn’t—even if you find that other researchers have had to make these same inferences—previous research may help you understand the logic by which others have decided that A affected B rather than B affecting A.

Our fourth and final line of evidence we’ll call logic, although it is probably a combination of logic, observation, understanding, and common sense. Go back to the illustration of what I termed informal theory. Teachers observe children every day in their classes; they are keen observers of the process of learning. If you were to ask a teacher, “Which is more likely, an increase in students’ levels of motivation affecting their learning or an increase in their learning affecting their motivation?” most would pick the former possibility. You can use the same sort of process to make such inferences. Imagine the ways in which A could affect B, and then imagine the ways in which B could affect A. If you are familiar with the phenomena you are considering, if you have observed them carefully, you will often find it easy to imagine the cause going in one direction but may require mental gyrations to imagine it going in the other. This logical exercise, then, will often suggest that one direction of causation is much more plausible than the other.

Again, these lines of evidence are how we make such inferences of cause and effect. Once we have made those inferences, the correlations merely provide fuel for our calculations.

More formally, three conditions are necessary before we can make a valid inference of causality (see Kenny, 1979, or Kline, 2016, for additional discussion of these conditions; for a considerably expanded discussion of the concept of causality, see Pearl, 2009 and Pearl and MacKenzie, 2018). First, there must be a relation between the variables being considered. If two variables are unrelated, then they are also *causally* unrelated. This condition is generally satisfied by the presence of a correlation between the variables (although there are exceptions). Second, and as already discussed, the presumed cause must have time precedence over the presumed effect. Causality does not operate backward in time. Third, the relation between the variables must be a true, rather than a spurious, relation. This is the hardest condition to satisfy and gets to the heart of what we have been calling the problem of omitted common causes. We will delve into this problem more deeply in the next chapter, but for now simply note that this condition means that all common causes are taken into account. Given that these three conditions are satisfied, it is perfectly reasonable to make an inference of cause and effect. What makes nonexperimental research so interesting and challenging is that we can often be very confident that we have satisfied these three conditions but never completely sure. (As it turns out, however, we can never be sure in experimental research either.)

Just to make sure we are all on the same page, let’s be completely clear as to what we mean by cause. When we say one variable “causes” another, we do *not* mean that one variable

directly and immediately results in change in another. When we say, for example, that smoking causes lung cancer, we do not mean that every person who smokes will necessarily and quickly develop lung cancer. What we mean is that if you smoke you will, as a result of smoking, increase your probability of developing lung cancer. The term cause is thus a probabilistic statement.

Jargon and Notation

I've been sneaking some of the jargon of SEM into the chapter as we introduce path analysis. Before we move to an expanded example, let's spend a little time going over such jargon so that it will be familiar. I have already noted that the variables representing other influences from outside the model are often called disturbances in path analysis, although many researchers use the term with which you are already familiar, residuals. In addition, I have noted that variables that we wish to symbolize but which we have not measured (unmeasured variables) are generally enclosed in circles or ovals. In contrast, measured variables, variables that we have measured in our data, are generally enclosed in rectangles. Paths or arrows represent influences from presumed cause to presumed effect, whereas curved, double-headed arrows represent correlations without an inference of causality.

Recursive and Nonrecursive Models

The models shown in Figures 12.2 and 12.3 are called *recursive* models, meaning that paths, and presumed causes, go in one direction only. It is also possible to have feedback loops in a model, to specify that two variables affect each other in a reciprocal fashion. Such models are termed *nonrecursive*; an example is shown in Figure 12.6, where Variable 2 is assumed to both affect (path *c*) and be affected by Variable 3 (path *d*). You cannot solve for the equations for nonrecursive models using the tracing rule, although you can generate the correct equations using the first law (see Note 2). Likewise, nonrecursive models cannot be estimated through multiple regression (you can estimate such models with MR, but the results will be incorrect). It is possible to estimate nonrecursive models using specialized SEM software or through a method called two-stage least squares regression, although such estimation is often tedious (and, as we will see momentarily, *this* model could not be estimated). It is tempting, especially for those new to SEM, to solve difficult questions of presumed cause and effect by deciding that such effects are reciprocal. Can't decide whether Motivation affects Achievement or Achievement affects Motivation? Draw paths in both directions! Generally, however, this is equivocation rather than decision. Nonrecursive models may require additional

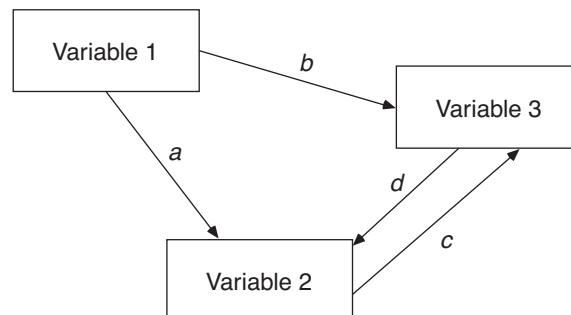


Figure 12.6 Nonrecursive model. The model is also underidentified and cannot be solved without additional assumptions.

constraints to avoid underidentification (see below) and, in my experience, often end up suggesting that the effect is indeed in the direction we would have guessed had we done the difficult work of making such decisions. I am not suggesting that you develop a cavalier attitude toward making decisions about the correct direction of causality; it often requires tough work and deep thought. Instead, I am arguing that you should not try to avoid this work by defaulting to nonrecursive models. Save such models for those instances when you have real, substantive questions about causal direction or when effects really appear to go in both directions. Some authors (e.g., Kenny, 1979) refer to recursive models as hierarchical models and nonrecursive models as nonhierarchical, but such usage may be confusing because sequential regression is also often termed hierarchical regression.

Identification

The model shown in Figure 12.3 is also a *just-identified* model. In a simplistic sense, what this means is that we had just enough information to estimate the model. Focus again on the Figures 12.1 through 12.3. We had three unknowns (the three paths in Figure 12.2), and we solved for these three paths using the three correlations from Figure 12.1. We had just enough information to solve for the paths. In addition to being a nonrecursive model, the model shown in Figure 12.6 is an *underidentified* model. For this model, we still have three correlations, but we now have four paths that we need to estimate. Unless we make some additional assumptions (e.g., assuming that paths d and c are equal), we cannot solve for the paths in this model.

The model shown in Figure 12.7, in contrast, is *overidentified*. For this model, we have more correlations than paths. The result is that we could, in fact, develop two separate sets of equations to solve for paths a and b . Consider the three equations generated from the tracing rule:

$$r_{13} = b \quad r_{12} = a \quad r_{23} = ab.$$

Using these equations to solve for a (and substituting for b), for example, we could generate the equations $a = r_{12}$ and $a = r_{23}/r_{13}$. And for b , $b = r_{13}$ and $a = r_{23}/r_{12}$. At first blush, the possibility of calculating two different estimates of each path might seem a problem. But consider for a minute what it would mean if our two estimates of the same path were very close to one another versus considerably divergent? Wouldn't you be more likely to believe a model in which you could estimate a path several different ways and always get the same result? We won't explore this topic in any greater depth right now but will return to it later. In the meantime, simply recognize that overidentified models are not problematic, but, rather, overidentification may help us evaluate the quality of our models.

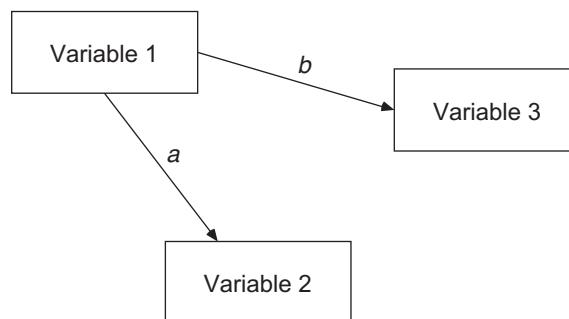


Figure 12.7 Overidentified model. The paths can be estimated more than one way.

This discussion has been a necessary simplification of the topic of identification, which can be much more complex than has been presented here. For example, it is possible for portions of a model to be overidentified and other portions to be underidentified. The primary rule presented for determining identification—comparing the number of correlations to the number of unknown paths—is really more of a necessary but insufficient condition for identification. Nevertheless, this rule generally works well for the simple path models of the type presented in this and the next chapter. For a more detailed discussion of the topic of identification with simple or complex models, see Bollen (1989).

Exogenous and Endogenous Variables

In SEM, the presumed causes (e.g., Ability in Figure 12.3) in a model are often referred to as *exogenous* variables. In medicine or biology, exogenous means “having a cause external to the body” (Morris, 1969, p. 461). An exogenous variable has causes outside the model or not considered by the model. Or, more simply, exogenous variables are ones that have no arrows pointing toward them. In contrast, variables that are affected by other variables in the model, variables that have arrows pointed toward them, are termed *endogenous* variables (meaning, loosely, from within). In Figure 12.3, Motivation and Achievement are endogenous variables.

Measured and Unmeasured Variables

In the discussion of disturbances, I noted that we generally symbolize unmeasured variables in path models by enclosing them in circles or ellipses. Unmeasured variables are variables that we wish to include in a path model, but we have no measures of these variables in our data. For now, the only unmeasured variables we will deal with are disturbances, but in later chapters we will focus on other types of unmeasured variables. Unmeasured variables are also known as *latent* variables or *factors*.

Variables enclosed in rectangles are measured variables for which we have actual measures in our data. These include all sorts of items, scales, and composites. Indeed, all the variables we have discussed so far in this book, with the exception of disturbances and residuals, are measured variables. Measured variables are also known as *manifest* or *observed* variables.

A MORE COMPLEX EXAMPLE

Now that you have a handle on the basics of path analysis, let’s expand our example to a more realistic level. We will now focus on the effects of Family Background characteristics, Ability, Motivation, and Academic Coursework on High School Achievement. These are, then, the same data and the same example from Chapter 10, but in path analytic form. The comparison of the results of the path analysis to the results for the different forms of multiple regression will be instructive and help illustrate important concepts about both methods.

Steps for Conducting Path Analysis

Here are the steps involved in conducting a path analysis (Kenny, 1979; Kline, 2016).

Develop the Model

The first step in path analysis is to develop and draw the model based on formal and informal theory, previous research, time precedence, and logic. Figure 12.8 shows my model, or theory, of how these variables are related to one another. School learning theories consistently include variables reflecting Ability (e.g., ability, aptitude, previous achievement), Motivation

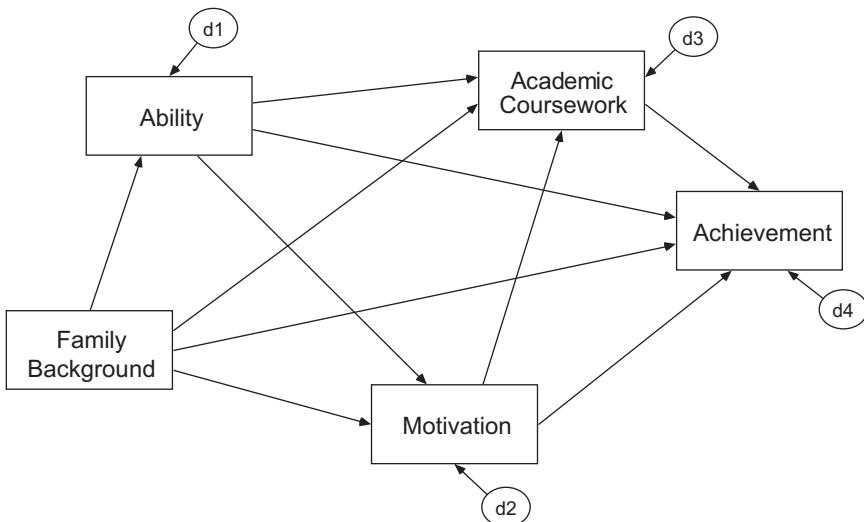


Figure 12.8 Model of the effects of Family Background, Ability, Motivation, and Academic Coursework on Achievement.

(internal motivation, perseverance), and Coursework (quantity of instruction, time spent learning, opportunity to learn) as influences on learning and achievement (e.g., Walberg, 1986). School learning theory, therefore, supports our drawing paths from Ability, Motivation, and Coursework to Achievement. You can probably easily justify these paths in other ways, as well.

Family Background is basically a background variable. By this I mean that it is included in the model because it seems needed to make the model valid (i.e., I think it may be a common cause of some of the variables and Achievement), but I'm not really interested in its effects on any of the other variables in the model. The fact that I consider this a background variable is not, however, justification for placing it first in the model. The likelihood that Family Background occurs before the other variables in time can be used to draw such paths, however, and you may find that the notion of *background variable* often is related to time. In the present case, Family Background is a parent variable, and most of its components—parents' level of education, occupational status—were likely in place, for many families, before children were even born. Even in cases in which parents were still in school or not yet employed when their children were born, time precedence would seem to flow from Family Background to the other variables in the model. Think about it: is it more likely that parents' SES will affect their child's ability (or motivation, etc.) or that a child's ability will affect his or her parents' SES? I suppose the second option is possible (children's ability affecting parents' SES), but it requires some mental gyrations to come up with plausible scenarios. Such reasoning may be used to draw paths from Family Background to each of the other variables in the model.

Time precedence, along with previous research, may also be used to justify the paths from Ability to each subsequent variable in the model. Ability, intelligence, or academic aptitude is relatively stable from an early elementary level on, and there is ample evidence that Ability affects many aspects of life and schooling, from Motivation to Achievement (Jensen, 1998).

This leaves the path from Motivation to Coursework. Imagine two high school students of equal ability and background. It is not hard to imagine one student taking a relatively easy mix of courses in high school and the other taking courses like pre-calculus, physics,

and advanced English. Academic Motivation—the desire to work hard and persevere in school, the expectation that schooling and what is learned in school will be important for the future—is likely a key difference between these students. Many of you can probably think of such examples in your own family, siblings or children who were highly motivated taking tough courses versus others just getting by. In essence, it makes a great deal of sense to posit that students with high levels of academic motivation will, other things being equal, take a tougher mix of academic courses than will students with lower levels of motivation. (Keith and Cool, 1992, further bolstered this time precedence by measuring Motivation 2 years prior to Coursework.)

This reasoning justifies the directions of the paths in the model, but what about the *variables* in the model? In particular, are there variables that should be included in the model that have not been included? That is, have I neglected an important common cause? Are there variables in the model that are unnecessary? I will postpone in-depth discussion of these issues until the next chapter. For now, I simply note that theory and previous research can help answer these questions, as well.

Check the Identification Status of the Model

Make sure that the model is either just-identified or overidentified so that the model may be estimated. The model shown in Figure 12.8 is just-identified. The correlation matrix includes 10 correlations, and there are 10 paths to be solved. The model appears to be just-identified and can probably be estimated.

Measure the Variables in the Model

We next need to decide how to measure the variables in the model. This may mean selecting tests and items designed to measure the constructs of interest and then administering these measures to a sample of participants. When using existing data, such as the NELS data, this may mean seeing if items that measure the variables of interest have already been administered to a sample of participants. In the present case, the variables in the model were already measured in the High School and Beyond data set; the authors selected items and composites to measure these constructs.

Estimate the Model

Our next step is to estimate the model. We are currently discussing how to estimate such models using multiple regression analysis; in subsequent chapters we will learn how to estimate such models using SEM software. To estimate the paths to Achievement using MR, we regress Achievement on Family Background, Ability, Motivation, and Academic Coursework. Partial results of this regression are shown in Figure 12.9. The b 's and β 's from the regression are the estimates of the unstandardized and standardized path coefficients, respectively, from each variable to Achievement. The R^2 is used to calculate the path from the disturbance (d4) to Achievement: $\sqrt{1 - R^2} = \sqrt{1 - .629} = .609$.

The paths to Academic Coursework are estimated by regressing Courses on Family Background, Ability, and Motivation, and the path from d3 to Coursework is estimated from the R^2 from that regression ($R^2 = .348$). Results from this regression are shown in Figure 12.10. The paths to Motivation are estimated from the regression of Motivation on Family Background and Ability, and the path from Family Background to Ability is estimated by the regression of Ability on Family Background. The relevant regression results are shown in Figure 12.11.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.793 ^a	.629	.627	6.103451

a. Predictors: (Constant), COURSES, FAM_BACK, MOTIVATE, ABILITY

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	6.434	1.692	3.803	.000	3.114	9.753
	FAM_BACK	.695	.218	.069	3.194	.001	.268
	ABILITY	.367	.016	.551	23.698	.000	.337
	MOTIVATE	1.26E-02	.021	.013	.603	.547	-.028
	COURSES	1.550	.120	.310	12.963	.000	.054
						1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 12.9 Using simultaneous regression to estimate the paths to Achievement.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.590 ^a	.348	.346	1.617391

a. Predictors: (Constant), MOTIVATE, FAM_BACK, ABILITY

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant)	-3.661	.433	-8.454	.000	-4.511	-2.811
	FAM_BACK	.330	.057	.165	5.827	.000	.219
	ABILITY	4.99E-02	.004	.374	13.168	.000	.042
	MOTIVATE	5.34E-02	.005	.267	10.138	.000	.043
							.064

a. Dependent Variable: COURSES

Figure 12.10 Estimating the paths to Academic Coursework through simultaneous multiple regression.

Figure 12.12 shows the path model with all the standardized path coefficients added. You should compare the model to the regression results to help you understand where each path came from, including those from the disturbances.

Interpretation: Direct Effects

So, what do these findings tell us? If you focus first on the paths to Achievement, you will see these findings and their interpretation are the same as those from the simultaneous multiple regression of Achievement on these four variables in Chapter 10. Ability and Academic Coursework each had a strong effect on Achievement (.551 and .310, respectively), whereas Family Background had a small, but statistically significant effect (.069). As in the simultaneous

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.235 ^a	.055	.053	9.729581

a. Predictors: (Constant), ABILITY, FAM_BACK

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant) 39.850	2.279		17.488	.000	35.379	44.322
	FAM_BACK 1.265	.339	.127	3.735	.000	.601	1.930
	ABILITY .101	.023	.152	4.495	.000	.057	.146

a. Dependent Variable: MOTIVATE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.417 ^a	.174	.173	13.640426

a. Predictors: (Constant), FAM_BACK

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1	(Constant) 100.000	.431		231.831	.000	99.154	100.846
	FAM_BACK 6.255	.432	.417	14.494	.000	5.408	7.102

a. Dependent Variable: ABILITY

Figure 12.11 Estimating the paths to Motivation and Ability.

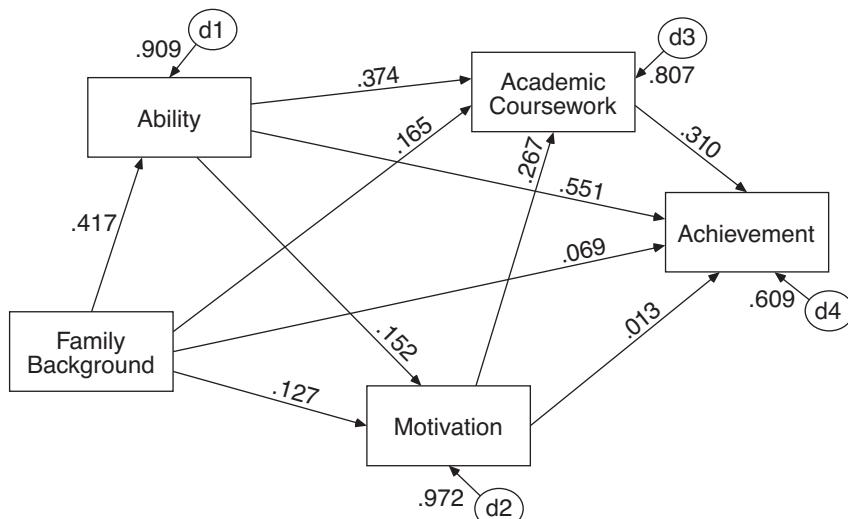


Figure 12.12 Solved model explaining Achievement, showing all standardized paths and disturbances.

regression of these same data in Chapter 10, the effect of Motivation on Achievement was small and not statistically significant.

The path model includes much more than this single simultaneous regression, however, because it also includes information about the effects *on* Coursework, Motivation, and Ability. Which of these variables affect the courses students take in high school? As hypothesized (and given the adequacy of the model), students' level of Academic Motivation had a strong effect on Coursework (.267); students who are more motivated take a more academic mix of courses than do students with lower levels of motivation. The largest effect on Coursework was from Ability (.374); more able students also take more academic courses in high school. Finally, Family Background also had a moderate effect on Coursework (.165), meaning that students from more advantaged backgrounds are more likely to take academic courses in high school than are students from less advantaged backgrounds.

The solved model also speaks to the extent to which Family Background and Ability affect Motivation; higher levels of both Ability and Family Background lead to higher levels of Academic Motivation. In addition, students from more advantaged backgrounds also show higher levels of Ability.

As an aside, notice the paths from the disturbances to each of the endogenous variables. As a general rule, these get smaller the farther to the left in the model. Don't read too much into this phenomenon. Achievement has four paths pointing toward it, four variables explaining it in the model, whereas Ability has only one explanatory variable (Family Background) pointing toward it. Other things being equal, it is natural that our model should explain more of the variance of Achievement than Ability, and thus the paths from the disturbances from Achievement should be smaller.

Indirect and Total Effects

The model (Figure 12.12) includes other information, beyond what we would get in the usual MR, as well (e.g., Chapter 10). The results of this analysis suggest that Motivation affects Coursework, which in turn affects Achievement. This makes sense: more motivated students take more academic courses in high school, and this coursework, in turn, improves their achievement. Thus, although Motivation has little direct effect on Achievement, it does have an indirect effect, through Coursework. In fact, we can easily calculate this indirect effect: multiply the path from Motivation to Coursework times the path from Coursework to Achievement ($.267 \times .310 = .083$), which is the indirect effect of Motivation on Achievement through Coursework. We can also add the direct and indirect effects to determine the *total* effect of Motivation on Achievement ($.083 + .013 = .096$).⁴

It is slightly more complex to calculate the indirect and total effects of Ability or Family Background, because the farther back you go in the model, the more possible indirect effects there are. To calculate the indirect effect of Ability on Achievement, for example, you would need to calculate the indirect effect through Coursework ($.374 \times .310 = .116$), Motivation ($.152 \times .013 = .002$), and both Motivation and Coursework ($.152 \times .267 \times .310 = .013$). These indirect effects are then summed to calculate the total indirect effect, .131, and added to the direct effect (.551) to calculate the total effect, .682. Table 12.1 shows the standardized direct, indirect, and total effects for each variable on Achievement. Calculate the indirect and total effects of Family Background on Achievement to see if your results match mine. Note also that there are no indirect effects for Coursework on Achievement. This is, of course, because our model includes no intervening variables between Coursework and Achievement. If it did, there would be indirect effects for Coursework as well.

Table 12.1 Standardized Direct, Indirect, and Total Effects of School Learning Variables on High School Achievement

Variable	Direct Effect	Indirect Effect	Total Effect
Academic Coursework	.310	—	.310
Motivation	.013	.083	.096
Ability	.551	.131	.682
Family Background	.069	.348	.417

Using Sequential Regression to Estimate Total and Indirect Effects

Recall that in Part 1 of this book we focused on differences in findings from simultaneous (or forced entry) and sequential (hierarchical) regression. I noted at the time that the reason for this difference is that simultaneous regression focuses on direct effects, whereas sequential regression focuses on total effects. We have seen in this chapter that the b 's and β 's from simultaneous regression may be used as estimates of the direct effects in path analysis. Figure 12.13 shows some of the output for the sequential regression of Achievement on the variables in the school learning model, reproduced from Chapter 10. The figure shows the table of coefficients, with the variables entered into the equation based on their order of appearance in the model; that is, the first (exogenous) variable (Family Background) was entered first, followed by Ability, and so on. Focus on the standardized coefficients, β 's, as each variable is added to the model; these coefficients are in italic boldface in the figure. Compare these coefficients to the total effects shown in Table 12.1 and you will see that they are the same, within errors of rounding. Thus, sequential regression may be used to estimate the *total effects* of each variable on the outcome for a path model. To do so, regress the endogenous variable of interest on each presumed cause in the order of their appearance in the model. The β for the variable entered at each step is the estimate of the variable's *total standardized effect* on the endogenous variable. Likewise, the b for the variable entered at each step is the estimate of the variable's total unstandardized effect. If you are interested in the statistical significance of the total effects, however, you need to correct the degrees of freedom, using the value with all variables in the model. That is, look up the statistical significance of the t 's using 995 df (total $N-k-1$), rather than the df from each equation. Using this method, we can calculate the indirect effects via simple subtraction: we subtract the direct effect from the total effect to estimate the total indirect effects of each variable on the outcome. Try this subtractive method to calculate the indirect effects in Table 12.1.

In Chapter 9 we focused on the topic of mediation in some depth. Recall that indirect effects are synonymous with mediation. Thus we can use the methods discussed in Chapter 9 to calculate the standard errors, confidence intervals, and statistical significance of indirect effects in this model. See also Kris Preacher's Web page on mediation mentioned in that chapter: www.quantpsy.org/sobel. Alternatively, as we will see in subsequent chapters, you can estimate the model with a SEM program, which will calculate standard errors of direct, indirect, and total effects.

Note we could also calculate the total effects of each variable on each of the other endogenous variables in the model (in addition to their effects on Achievement). To estimate the total effects of each variable on Coursework, for example, we sequentially regress Coursework on Family Background, followed by Ability, and followed by Motivation. The coefficient for the variable entered at each step equals its total effect on Coursework. The coefficients for the final step in the multiple regression equal the direct effects for each variable on Coursework.

Model	Coefficients ^a						
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.000	.288	.417	173.873 14.494	.000 .000	49.436 3.605
	FAM_BACK	4.170	.288				50.564 4.735
2	(Constant)	4.557	1.559	.133	2.923	.004	1.498 .873
	FAM_BACK	1.328	.232		5.729	.000	7.617 1.782
	ABILITY	.454	.015		.682	.000	.424 .485
3	(Constant)	.759	1.766	.667	.430	.667	-2.706 .753
	FAM_BACK	1.207	.231		.121	.000	4.224 1.661
	ABILITY	.445	.015		.667	.000	.414 .475
	MOTIVATE	9.53E-02	.021		.095	.000	.053 .137
4	(Constant)	6.434	1.692	.000	3.803	.000	3.114 .268
	FAM_BACK	.695	.218		.069	.001	9.753 1.122
	ABILITY	.367	.016		.551	.000	.337 .398
	MOTIVATE	1.26E-02	.021		.013	.603	-.028 .054
	COURSES	1.550	.120		.310	.000	1.315 1.785

a. Dependent Variable: ACHIEVE

Figure 12.13 Using sequential multiple regression to estimate the total effects of each variable on Achievement. The indirect effects are then calculated through subtraction (total–direct).

We can calculate the indirect effects by subtracting the direct from the total effect for each variable.

Even with only five variables in the path model, it soon becomes tedious to solve for indirect and total effects directly, that is, by multiplying and summing paths. There are several possible shortcuts for doing such calculations. The one we have illustrated here—using sequential regression to estimate total effects and then calculating indirect effects by subtraction—is one of the easiest and has the advantage of illuminating the previously puzzling relation between sequential and simultaneous regression. The reason simultaneous and sequential regression tell different stories is because they focus on different questions; simultaneous regression focuses on direct effects, whereas sequential regression focuses on total effects. I hope the method also illustrates the importance of proper order of entry in sequential regression. If you wish to interpret sequential regression results in a causal fashion, you must enter the variables in their proper causal order.

It should be clear that this method of estimating total and indirect effects *does* work, but it may not be clear *why* it works. Recall that for the next to last variable in the causal chain (Coursework) the direct effects were equal to the total effects. The reason, of course, is there are no intervening or mediating variables between Coursework and Achievement and thus no possible indirect effect. The total and direct effects for Coursework on Achievement are the same. All the effect of one variable on another, then, is a direct effect *when there are no intervening variables*. It then stands to reason that one way of calculating total effects is to remove intervening variables.

In essence, what we have done with our sequential regression is to temporarily remove the intervening variables. Focus on Figures 12.14 through 12.16. The first step in the sequential regression, in which Achievement was regressed on Family Background, estimates the model shown in Figure 12.14. In this model, all intervening variables between Family Background and Achievement are removed. The total effect of Family Background remains the same whether there are no intervening variables or whether there are three, or even 30, intervening variables; the total effects *are always the same*. Therefore, when we estimated *this* model, with the intervening variables removed, the direct effects and total effects are the same. The regression coefficient from this regression (.417) can then be used as an estimate of the total effect for the full model with all intervening variables. Figure 12.15 removes the intervening

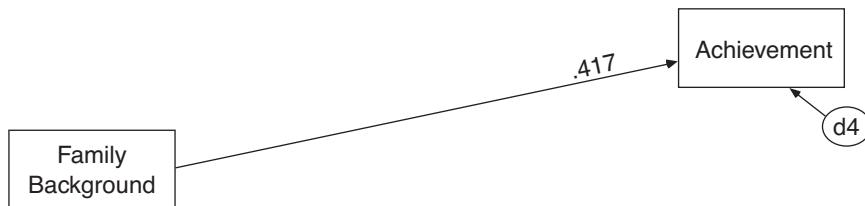


Figure 12.14 “Model” used to estimate the total effect of Family Background on Achievement.

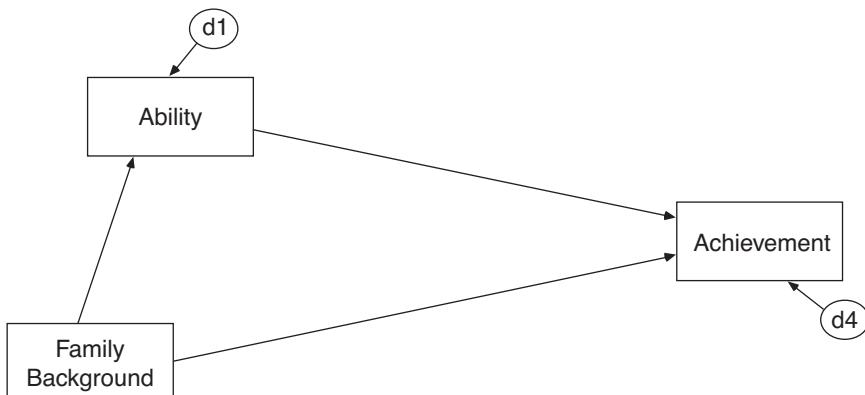


Figure 12.15 Estimating the total effect of Ability on Achievement. The total effect is estimated by the β (or b) for the variable added at this stage of the sequential regression.

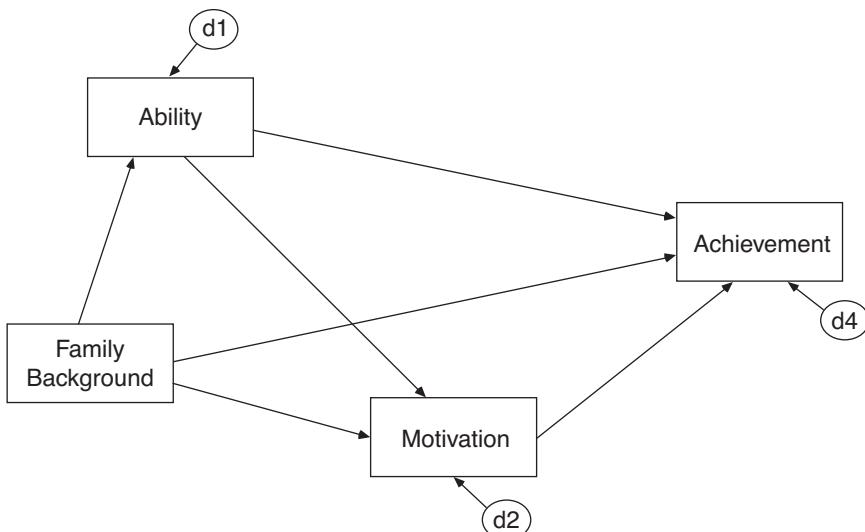


Figure 12.16 Estimating the total effect of Motivation on Achievement.

variables between Ability and Achievement. The second step in the sequential regression, in which Achievement is regressed on Family Background and Ability, operationalizes the model in Figure 12.15, and because there are no intervening variables between Ability and Achievement, the regression coefficient for Ability estimates the total effect of Ability on

Achievement. Finally, the model shown in Figure 12.16, the third step in the sequential regression, provides the estimate of the total effect of Motivation on Achievement.

Interpretation

Let's take a few minutes to interpret these findings and, at the same time, further understand the relation between simultaneous and sequential regression. Focus on Motivation in Figure 12.12. The path model and Table 12.1 suggest that Motivation's effects on Achievement are primarily indirect, not direct. Motivation influences Achievement by influencing the courses students take in high school. Highly motivated students take more academically oriented courses, and these courses, in turn, improve their Achievement. Academic coursework mediates the effect of Motivation on Achievement. In contrast, Ability's effects on Achievement are primarily direct. A portion of the effect of Ability is indirect, through Motivation and Coursework—more able students are more highly motivated and take more academic coursework, on average, than less able students—but the majority of the effect is direct: more able students also have higher academic Achievement. Again, the simultaneous regressions focused on direct effects and the sequential regressions focused on total effects.

I hope this discussion has illustrated some of the heuristic beauty of path models. They allow us to focus on both direct and indirect effects. Indirect effects, also known as mediating effects, are often vital for understanding how an influence comes about. *How does Motivation affect Achievement?* One important way is by influencing the courses students choose to take in high school. More motivated students take more academic coursework, and this coursework raises achievement. We often miss understanding these indirect effects when we analyze our data with ordinary MR without path models. When you conduct path analysis, make sure to calculate and interpret all three types of effects. When you find a direct effect and wonder how it comes about, try incorporating several plausible mediating variables in a path model to see if you can understand how these effects happen. Suppose you find, for example, that physical abuse affects children's later social status. You may wonder whether these children's social behaviors (e.g., aggression) mediate, and thus partially explain, this effect. That is, are abused children more likely to be aggressive, with the aggression leading to a reduction in their subsequent social status (Salzinger, Feldman, Ng-Mak, Mojica, & Stockhamer, 2001)?

Path analysis has other advantages over multiple regression. A figure often makes it more obvious than does a table of regression coefficients exactly what are the presumed causes and the presumed effects. I think that the obviousness of the figural, causal assumptions in path analysis makes it more likely that the researchers will consider causal assumptions, as well as the basis for making these assumptions (theory and previous research). If nothing else, the drawing of the path model is at least an informal theory of cause and effect. As already discussed, path analysis makes use of the different stories told by simultaneous and sequential regression. For these reasons, I believe that path analysis (and SEM) is often the best method of analysis for nonexperimental research.

SUMMARY

We have covered a lot of material in this chapter, and I hope the chapter has both covered new ground and made clear some loose ends from our adventures in MR. This chapter formally introduced path analysis, which is the simplest form of structural equation modeling, or SEM.

We introduced the chapter with a simple model involving Ability, Motivation, and Achievement. Our initial, agnostic model simply showed the correlations among these three

variables, a less-than-satisfying solution since it did not inform our research question of interest, which was understanding the influence of Motivation on Achievement. Thinking through our research interest and using a combination of theory, logic, and previous research, we were able to make some general causal statements: (1) if Motivation and Achievement are causally related, Motivation affects Achievement, rather than the reverse, and (2) Ability may affect both Motivation and Achievement. These statements, a weak causal ordering, were translated into a path model in which Ability was assumed to affect both Motivation and Achievement and Motivation was assumed to affect Achievement. The correlations, notably, were not used to draw the paths. We now had three unknowns (three paths) and three pieces of data (the correlations), and through the use of algebra we were able to generate equations for and solve for the paths.

Although we can solve for the paths using algebra, for simple recursive models the paths are equal to the standardized or unstandardized coefficients from a series of simultaneous regressions. For the three-variable model, we regressed Achievement on Ability and Motivation, with the β 's providing estimates of the standardized paths from Ability and Motivation (and the b 's estimating the unstandardized paths). A second regression of Motivation on Ability provided the estimate of the path from Ability to Motivation. The influences of the disturbances (or residuals) were estimated by $\sqrt{1 - R^2}$ from each regression equation. Disturbances represent all other influences on a variable besides the variables in the model, and were symbolized by variables enclosed in circles or ovals.

What evidence was used to make the inferences of causality? It was not the correlations. Instead, we focused on formal and informal theory, time precedence, an understanding of the phenomenon being studied, and logic. At a more formal level, three conditions are required to make a valid inference of cause and effects: there must be a functional relation between the variables, the cause must precede the effect in time (either actually or logically), and the relation must be nonspurious.

We dealt with some jargon you are likely to encounter in path analysis. Measured variables, those measured in your research, are symbolized by rectangles. Unmeasured, or latent variables, are symbolized by circles or ovals. Disturbances represent unmeasured variables not considered in the model; disturbances may also be referred to as residuals or errors. Recursive models have arrows flowing in only one direction, whereas nonrecursive models have feedback loops, or arrows pointing in two directions. Just-identified models are those for which we have just enough information to solve for the paths, and overidentified models are those for which we have more information than we need and can thus estimate some of the paths in more than one way. Underidentified models are those for which we have more paths than we have information to estimate the paths; they are therefore not solvable without the addition of extra constraints. Exogenous variables are presumed causes, variables with no paths pointing towards them. Endogenous variables are presumed effects; they have paths pointing to them in the model. Most of this jargon is summarized in Figure 12.17.

We conducted a path analysis using the data from Chapter 10, where the data were used to highlight the differences in findings from simultaneous and sequential regression. We developed a model of the effects of Family Background, Ability, Motivation, and Coursework on Achievement based on theory, time precedence, previous research, and logic. Paths and disturbances were estimated via a series of simultaneous multiple regressions. Given the accuracy of the model, the results suggested that Ability and Coursework had strong effects on Achievement, Family Background had a small effect, and Motivation had no appreciable effect. Further inspection of the model showed that Motivation had a strong effect on the Coursework students take in high school, so Motivation should have an indirect effect on Achievement through Coursework. We were able to calculate these indirect effects by multiplying together the two paths. We added this indirect effect to the direct effect to estimate the

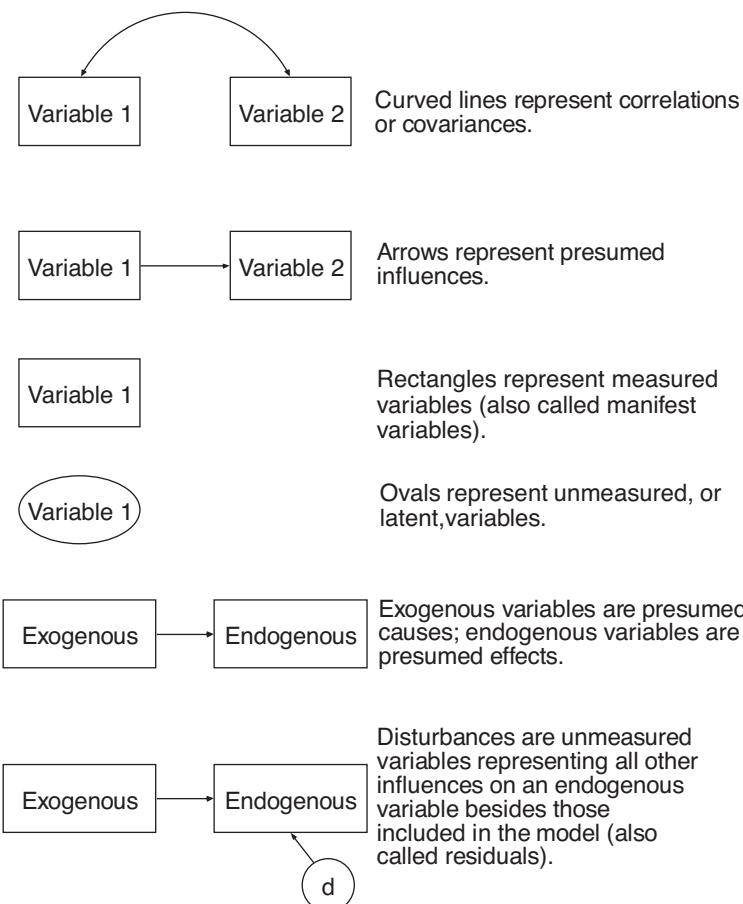


Figure 12.17 Quick summary of some of the jargon of path analysis.

total effect of Motivation on Achievement. When we focused on the total effect, Motivation did indeed have an influence on Achievement and one that makes sense: more motivated students, it appears, take more advanced coursework, and this coursework, in turn, improves their achievement.

An easier way to estimate total effects is through sequential regression. To do so, we regressed Achievement on Family Background and then added Ability, then Motivation, and then Coursework. The β associated with each variable, when entered, represents its total standardized effect. Thus, when Motivation was added to the model, its β was .095, its total effect. This procedure works because the total effects are the same whether or not there are intervening variables between the variable of interest and the outcome. If we remove the intervening variables, the total effects are equal to the direct effects. We then estimated the indirect effects by subtracting the direct from the total effects.

In addition to illustrating the basics of path analysis, this chapter tied together a major loose end of Part 1, the apparent inconsistency between simultaneous and sequential regression results. I argued that path analysis is particularly useful because it allows us to focus on both direct and indirect effects and that indirect effects are useful in explaining how an effect works. Intervening or mediating variables can thus be added to models to help understand how an effect comes about. Path models are also useful because they make explicit what is

too often left vague: the researcher's theory of how variables are causally related. In my opinion, path analysis is the best use of MR for explanatory, nonexperimental research.

EXERCISES

- Table 12.2 shows the means, standard deviations, and correlations among the variables used in this chapter's example. Reanalyze the five-variable path model. (For users of SPSS, the file "motivate 5 var path.sps" on the Web site (www.tzkeith.com) shows how to analyze such a matrix using this program.) Calculate all paths and disturbances to create a table of direct, indirect, and total effects. Make sure your results match mine.

Table 12.2 Means, Standard Deviations, and Correlations among the School Learning Variables

	<i>Family Background</i>	<i>Ability</i>	<i>Motivation</i>	<i>Coursework</i>	<i>Achievement</i>
N	1000	1000	1000	1000	1000
Mean	0	100	50	4	50
SD	1	15	10	2	10
Family Background	1				
Ability	.417	1			
Motivation	.190	.205	1		
Coursework	.372	.498	.375	1	
Achievement	.417	.737	.255	.615	1

- Construct a path model using the variables Family Background, 8th-grade GPA, 10th-grade Self-Esteem, 10th-grade Locus of Control, and 10th-grade Social Studies achievement test scores. How did you make the decisions on which variable affected which? Which of these decisions were the most difficult? What sources could you use to better inform your decisions?
- What is the identification status of your model from Exercise 2: just-identified, overidentified, or underidentified? If your model is underidentified, see if you can make it into a just-identified model so that you can estimate it.
- Select the variables BYSES, BYGrads, F1Cncpt2, F1Locus2, and F1TxHStd from the NELS data. Check the variables (e.g., descriptive statistics) to make sure you understand the scales of the variables. Also make sure that any values that should be coded as missing values are so coded.
- Estimate your model using the variables from NELS (Exercise 4). Calculate the direct effects and disturbances, and put them into your model. Calculate total effects and create a table of direct, indirect, and total effects. Interpret the model; focus on direct, indirect, and total effects.
- Compare your model and interpretation with others in your class. How many classmates drew the model in the same way you did? How many drew it differently? What difference did these different models make in results and interpretation?
- Curtis Hansen tested a path model of the influences on accidents among chemical industry workers (1989). A simulated version of a portion of the data are on the website (www.tzkeith.com) under Chapter 12 (e.g., "Hansen accident data.sav"; the file is also available in other formats). A guiding question for our analysis might be: what are the

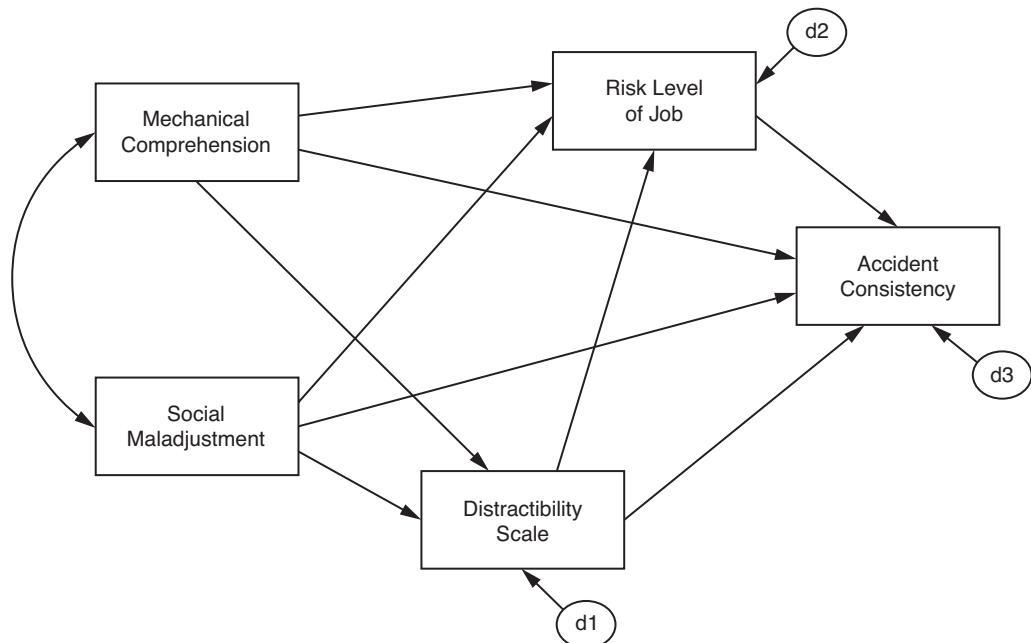


Figure 12.18 Path model of the presumed effects of abilities, personality characteristics, and job characteristics on number and consistency of accidents in an industrial setting.

relative effects of abilities, personality characteristics, and job characteristics on workers' accident rates? Figure 12.18 shows a model designed to answer this question.

Mechanical Comprehension (Mechanic in the data file) was a measure of workers' understanding of mechanical reasoning. Social Maladjustment (Maladjust) was a 50-item scale derived from the MMPI and designed (by Hansen) to assess general social maladjustment. These two variables are exogenous variables in the model. The Distractibility Scale (Distractibility), also derived from the MMPI, was designed to assess distractibility, and especially "neurotic-anxious" (Hansen, 1989, p. 83) characteristics that should lead to distractibility. The Risk Level of the Job (Risk) was a rating of the "responsibility and accident potential" (Hansen, p. 84) of each possible job on a 1 to 35 scale. The final endogenous outcome variable was Accident Consistency (Accident), a measure of the number of accidents for a worker plus the number of years in which each worker had an accident.

Estimate the model shown in the figure using multiple regression analysis. What is the identification status of the model? Calculate the direct effects and disturbances and put them in your model. Calculate total effects on Accident Consistency and create a table of direct, indirect, and total effects. Interpret the results. What were the important effects on accident consistency? Were there meaningful indirect effects? If so, interpret them. Which variable(s) had the strongest total effect on accident consistency?

Notes

- 1 Here's more detail in solving the paths using algebra. The three equations were

$$r_{13} = b + ac,$$

$$r_{23} = c + ab, \text{ and}$$

$$r_{12} = a. \text{ We can rearrange these equations to solve for paths } a, b, \text{ and } c:$$

$$\begin{aligned} b &= r_{13} - ac, \\ c &= r_{23} - ab, \text{ and} \\ a &= r_{12}. \end{aligned}$$

We will solve the equation for b by substituting the third and second equations (for a and c , respectively) into the first equation:

$$\begin{aligned} b &= r_{13} - r_{12}(r_{23} - r_{12}b) \\ &= r_{13} - (r_{12}r_{23} - r_{12}^2b) \\ &= r_{13} - r_{12}r_{23} + r_{12}^2b \\ b - r_{12}^2b &= r_{13} - r_{12}r_{23} \\ b(1 - r_{12}^2) &= r_{13} - r_{12}r_{23} \\ b &= \frac{r_{13} - r_{12}r_{23}}{1 - r_{12}^2} \end{aligned}$$

See if you can use the same approach to solve for c .

- 2 The other method of developing equations to solve for paths is called the first law of path analysis (Kenny, 1979, p. 28). The correlation between Y (a presumed effect) and X (r_{xy}) is equal to the sum of the product of each path (p) from all causes of Y times the correlation of those variables with X: $r_{yx} = \sum p_{yz}r_{xz}$. Using the first law, the correlation between Motivation and Achievement is $r_{32} = br_{12} + cr_{22}$, which reduces to $r_{32} = br_{12} + c$ (description and equation adapted from Kenny, 1979, p. 28). The advantage of the first law is that it can be used to generate equations for any type of model, whereas the tracing rule works only with simple recursive models.
- 3 These rules are that standardized coefficients above .05 could be considered small; those above .10, moderate; and those above .25, large. These rules apply primarily to manipulable influences on school learning.
- 4 Total effects are sometimes referred to as total causal effects. It is also possible to subtract the total causal effects from the original correlation to determine the noncausal (or spurious) portion of the correlation.

13

Path Analysis

Assumption and Dangers

Assumptions	281
The Danger of Common Causes	282
<i>A Research Example</i>	284
<i>Common Causes, Not All Causes</i>	286
<i>Intervening (Mediating) Variables</i>	288
Other Possible Dangers	289
<i>Paths in the Wrong Direction</i>	289
<i>Unreliability and Invalidity</i>	291
Dealing With Danger	291
Review: Steps In a Path Analysis	292
Summary	293
Exercises	294
<i>Notes</i>	295

Path analysis is not magic; it does not prove causality. It does not make a silk purse out of a sow's ear; it cannot turn poor data into valid causal conclusions. Like multiple regression, there are assumptions underlying path analysis and the use of multiple regression to estimate paths. Like multiple regression, path analysis is open to abuse. This chapter will discuss these assumptions and the dangers of path analysis; it will also discuss how to avoid the dangers of the method.

ASSUMPTIONS

Because we have so far been using multiple regression to estimate path models, it should not be surprising that the basic assumptions of multiple regression also apply to path analysis. As discussed in Chapter 10, these include the following:

1. The dependent variable is a linear function of the independent variables. In addition, the causal direction in the model must be correct.
2. Each person (or other observation) should be drawn independently from the population.
3. The errors are normally distributed and relatively constant for all values of the independent variable.

Multiple regression analysis assumes that the errors are uncorrelated with the independent variables or, in the jargon of path analysis, the disturbances are uncorrelated with the exogenous variables. Therefore, the causal mechanism underlying our path analysis (or multiple regression) model needs to conform to these same constraints in order for the regression coefficients to provide accurate estimates of the effects of one variable on another. This assumption also implies several additional assumptions; to the extent that the following conditions are violated, the paths (regression coefficients) may be inaccurate and misleading estimates of the effects.

1. There is no reverse causation; that is, the model is recursive.
2. The exogenous variables are perfectly measured, that is, they are completely reliable and valid measures.
3. “A state of equilibrium has been reached” (Kenny, 1979, p. 51). This assumption means that the causal process has had a chance to work.
4. No common cause of the presumed cause and the presumed effect has been neglected; the model includes all such common causes (Kenny, 1979).

If these sound a lot like the assumptions from Chapter 10, you are perceptive; they are virtually the same but rewritten in path analytic lingo. These assumptions are also required any time we wish to interpret regression coefficients in a causal, or explanatory, fashion.

The first assumption (of the second set) is really twofold. It first means that we have paths drawn in the correct direction. We have already discussed how this is done and will continue to discuss this critical issue in this and later chapters. This assumption also means, as indicated, that the model is recursive, with no feedback loops or variables both causing and affecting other variables. There are methods for estimating such models, but ordinary multiple regression is not a valid method for nonrecursive models.

The second assumption is one we can only approximate. We all know there is no such thing as perfect measurement, especially in the social sciences. When we begin discussing latent variable SEM, we will see how serious our violation of this assumption is and what can be done about it. For now, I will simply note that if scores on our exogenous variables are reasonably reliable and valid little harm is done, meaning our estimates of effects are not overly biased.

The third assumption is that the causal process has had a chance to work. If motivation affects achievement, this process presumably takes a certain amount of time, and this time must have elapsed. This assumption applies to all causal research. Consider an experiment in which children are given some treatment and subsequently measured on a dependent variable. If you make these measurements too soon, not allowing the treatment to work, you will miss spotting any real effects your treatment may have. The amount of time needed depends on the process being studied.

The final assumption is the most crucial, and it is one we have returned to over and over in this book. We will now explore it in more depth, because the danger of omitted common causes is the biggest threat to the causal conclusions we reach from path analysis, in particular, and nonexperimental research in general. Again, I remind you that these assumptions apply to *any* explanatory use of MR.

THE DANGER OF COMMON CAUSES

Suppose I were to go into my local elementary schools and ask every student to read the Gettysburg Address, and I scored each student on the number of words he or she read correctly within 2 minutes. Suppose that I also measured each child’s shoe size. If we correlate these two variables (reading skill and shoe size), we likely will find a substantial correlation

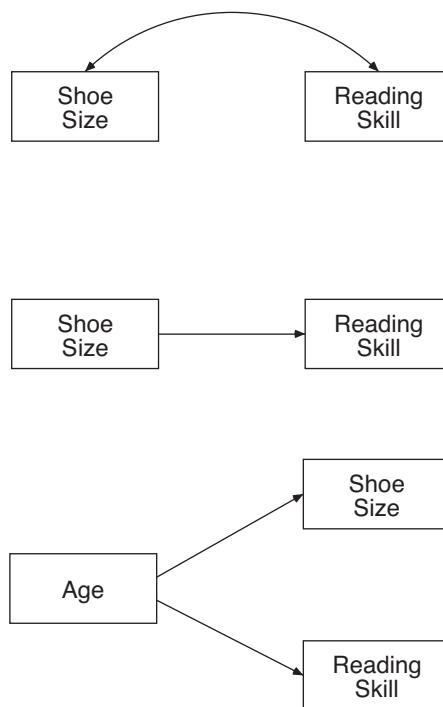


Figure 13.1 Spurious correlation in path form. Although shoe size and reading skill are correlated, shoe size does not cause reading skill, nor does reading skill cause shoe size. There is a third variable, age or growth, that affects both reading skill and shoe size. This common cause of shoe size and reading skill is why the two variables are correlated.

between them. This correlation is illustrated in the top of Figure 13.1 by the curved line between the two variables. It is foolish, however, to conclude that shoe size affects reading skill (as is done in the middle portion of the figure), and it is equally foolish to conclude that reading skill affects shoe size. The reason is that there is a third variable—age or growth—that affects both shoe size and reading skill, as symbolized by the bottom portion of Figure 13.1. Older students, on average, are larger (and thus have larger shoes) and read better than do younger students. The bottom of the figure illustrates the true causal relation among these variables; shoe size and reading skill are correlated only because the two are affected by age. The correlation between shoe size and reading skill is the essence of what we call a spurious correlation. The term *spurious correlation* means that two variables are not related by one variable affecting the other but are the result of a third variable affecting both (cf. Simon, 1954).

This example also illustrates the essence of the problem we have been referring to as that of a neglected common cause. If we set up a path analysis of the reading–shoe size data in which we assumed shoe size affected reading skill (as in the middle of Figure 13.1), the results would not tell us we were foolish, but instead would suggest that shoe size had a substantial impact on reading skill. The reason, again, is that we neglected to control for age, the common cause in our analysis. If we controlled for age, we would see the apparent effect of shoe size on reading skill diminish to zero. The model is crucial; for the estimates to be accurate, we must control for important common causes of our presumed cause and presumed effect. This problem is referred to as omitted common causes, spurious correlation, model misspecification, or the third-variable problem.

A Research Example

A more realistic example will further illustrate the problem. There is ample evidence that involvement by parents in education improves students' learning (Christenson, Rounds, & Gorney, 1992), but estimates of the effects of parent involvement on learning vary widely across studies. Figure 13.2 shows a plausible model of the effects of Parent Involvement on 10th-grade GPA. For this model, Parent Involvement was defined as a combination of parents' educational aspirations for their children and communication between parents and their children about school. Background variables—potential common causes of Parent Involvement and 10th-Grade GPA—include students' Ethnic background (Underrepresented Ethnic Minority)¹, their Family Background characteristics, and their previous school performance (Previous Achievement). Let's concentrate on this final variable. Previous Achievement should certainly affect students' current academic performance, since it forms a basis for all future learning. But should students' previous academic performance also affect the degree to which parents are involved in students' schooling? I think it should; it should affect both parent involvement, in general, and more specifically parents' aspirations for their children's future educational attainment (one of the components of parent involvement). We could turn to previous research and determine that students' previous performance or aptitude indeed affects their parents' level of involvement. In other words, Previous Achievement, or aptitude, appears to be a likely common cause of both Parent Involvement and current GPA.

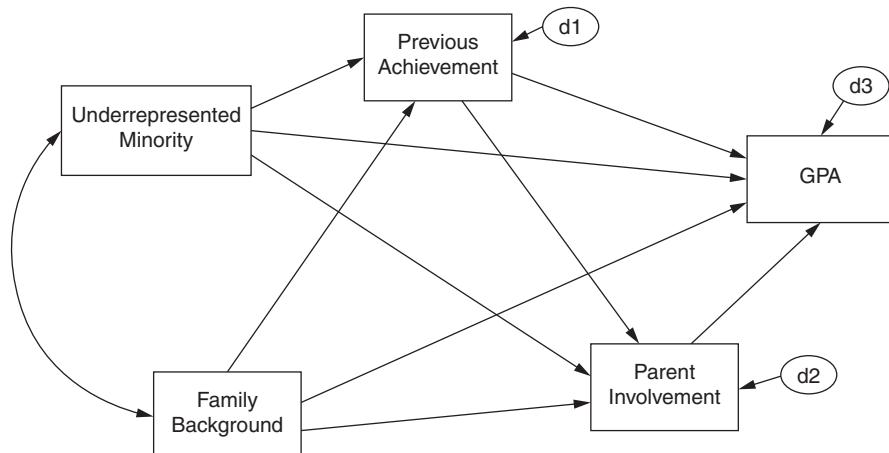


Figure 13.2 Model of the effects of Parent Involvement on high school GPA. The model is just-identified and recursive.

I estimated the model using the NELS data; the standardized results are shown in Figure 13.3. Parent Involvement appears to have a moderate effect on student GPA ($\beta = .16$). The results show that our supposition about Previous Achievement was also correct: Given the adequacy of the model, the results suggest that Previous Achievement had a large effect on both GPA (.42) and Parent Involvement (.34). Previous Achievement thus appears to be an important common cause of Parent Involvement and current Grades.

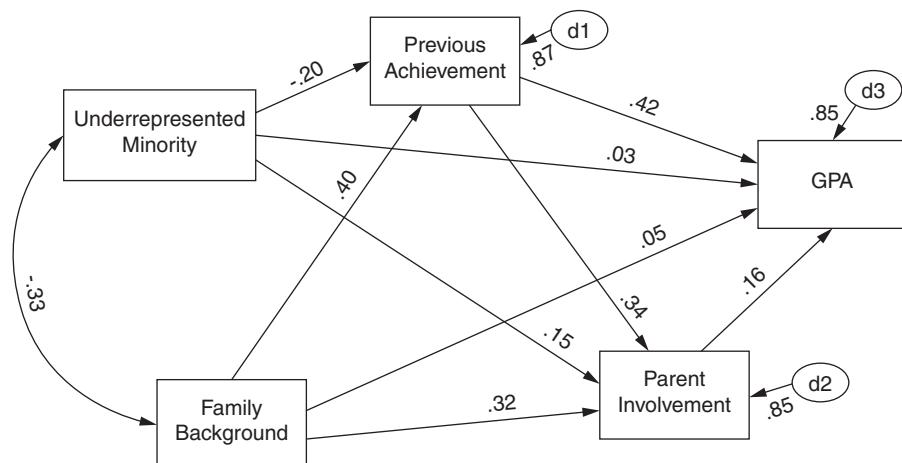


Figure 13.3 Parent Involvement model estimated through multiple regression analysis. Note the effect of Previous Achievement on Parent Involvement and GPA.

What would happen if we were not attuned to the importance of students' previous school performance? What if we had not built Previous Achievement into our model? What if we had neglected this important common cause? The results of such neglect are shown in Figure 13.4. In this model, Previous Achievement was not included; this important common cause was not controlled. The result is that the model substantially overestimates the effect of Parent Involvement on GPA: the effect in this model is .29, as opposed to .16 in the previous model. With the omission of this important common cause, we overestimated the effect of Parent Involvement on GPA.

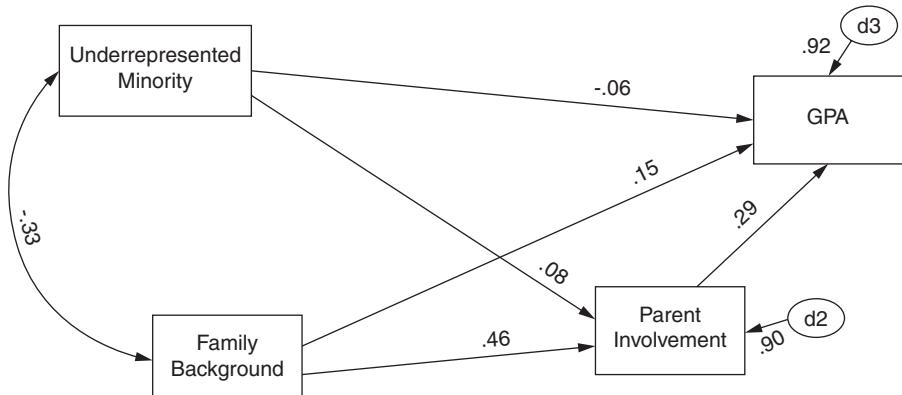


Figure 13.4 Previous Achievement, a common cause of Parent Involvement and GPA, is not included in this model. Notice the inflation of the path from Involvement to GPA.

This example illustrates the importance of including known common causes in path models. The example also illustrates the most frequent consequence of neglecting these common causes: When a common cause is omitted from a model, we often end up overestimating the magnitude of the effect of one variable on another.² Finally, the example illustrates one possible reason for the variability in findings concerning the effect of parent

involvement on school performance: not all research has controlled for previous achievement (and there are other possible explanations, as well). Research on parent involvement is not the only area in which researchers have likely overestimated effects by ignoring important common causes. For example, Page and Keith (1981) showed how Coleman and colleagues (Coleman, Hoffer, & Kilgore, 1981) had overestimated the effects of private schooling on student achievement by ignoring student ability as a potential common cause of achievement and private school attendance. In fact, if you are suspicious of the findings of nonexperimental research, you should probably first look for neglected common causes as the reason for misleading findings.

Note that there was nothing in the analysis summarized in Figure 13.4 that told us we had missed an important common cause. The analysis did not explode; no alarm bells went off. How then do you know that you have included all relevant common causes in your research? A good understanding of relevant theory and previous research are the keys to avoiding this deadly sin, just as they are for drawing the model in the first place.

Common Causes, Not All Causes

Unfortunately, many neophytes to path analysis (and nonexperimental research in general), terrified of neglecting a common cause of a presumed cause and a presumed effect, include every variable they can think of that might be such a common cause. Others misunderstand the admonition about *common* causes and try to include all possible causes of *either* the presumed cause or the presumed effect. Both approaches lead to overloaded and less powerful analyses (by reducing degrees of freedom in the regression), ones that are more likely to confuse than inform (and see Darlington, 1990, chap. 8, or Darlington & Hayes, 2017, chap. 17, for additional dangers with including too many variables).

I demonstrated in Chapter 9 (and also note 2 in Chapter 4) that the inclusion of a non-common cause in a regression does not change the estimates of regression coefficients (it can, under certain circumstances, increase the power of a statistical test, however; Darlington & Hayes, 2017, chap. 17). Here we will demonstrate this truism again using the current example. Focus again on Figure 13.3. For this model, we do not need to include all causes of Parent Involvement in the model, nor do we need to include all causes of GPA in the model. This is fortunate, because there must be hundreds of variables that affect GPA alone! All we need to include in the model are *common* causes of Parent Involvement and GPA. Note the effect of Underrepresented Minority background on Parent Involvement and GPA. URM affects Parent Involvement; other things being equal, underrepresented ethnic minority students report greater involvement than do students who are white or of Asian descent (URM students are coded 1 and white/Asian students coded 0). But once the other variables in the model are controlled, URM had no meaningful effect on GPA ($\beta = -.03$). Despite its inclusion in the model, it appears that URM is not a common cause of Parent Involvement and GPA. If my argument is correct, if variables need not be included in a model unless they are common causes, then the exclusion of URM from the model should have little effect on our estimate of the magnitude of influence of Parent Involvement on GPA. As shown in Figure 13.5, the exclusion of URM had only a minor effect on this estimate, which changed from .160 to .165 (rounded to .17). We could exclude URM from this model without seriously affecting the estimate of the influence of Parent Involvement on GPA. To reiterate, models must include common causes of the presumed cause and the presumed effect if they are to be valid, but they need not include *all* causes.³

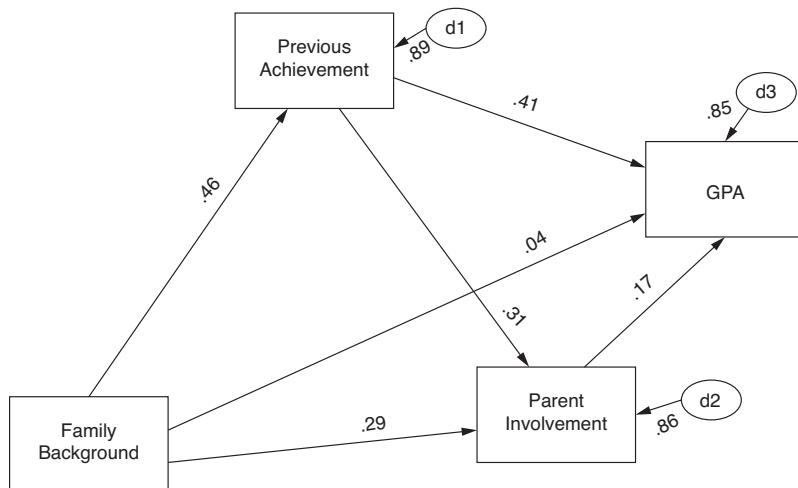


Figure 13.5 In this model, Underrepresented Minority was excluded. But URM was not a meaningful common cause; it affected only Parent Involvement, not GPA, in Figure 13.3. Thus, its exclusion in this model has little effect on the estimate of the effect of Parent Involvement on GPA.

True Experiments and Common Causes

The elimination of the danger of omitted common causes is the reason that true experiments allow such a powerful inference of cause and effect. As a general rule, experimental research, in which participants are *assigned at random* to experimental versus control groups, has a higher degree of internal validity than does nonexperimental research, meaning that it is generally less dangerous to make an inference of cause and effect with a true experiment. Figure 13.6 helps illustrate the reason for this power. Suppose you conduct an experiment in which you assign, at random, children with behavior disorders to two types of treatments: group therapy or behavior modification, with some measure of behavior improvement as the dependent variable. Figure 13.6 illustrates this experiment in path analytic form, with the path from the dummy variable Group Therapy versus Behavior Modification to Behavior providing the estimate of the relative effectiveness of the two treatments. But a multitude of variables affect children's behavior, from parents to friends to teachers, and many more. Why don't we have to consider

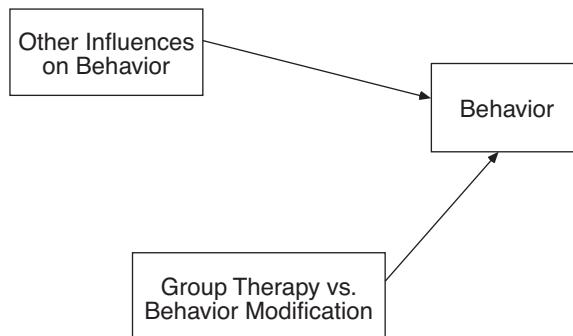


Figure 13.6 A true experiment in path form. Due to random assignment to groups (Group Therapy versus Behavior Modification), the variables that affect Behavior (the effect) do not affect the cause (treatment group). Random assignment rules out common causes, which is why we don't need to control for the multitude of other influences on Behavior in our analysis.

these variables when we conduct our analysis of the experiment? The reason we don't have to consider these Other Influences on Behavior is because they are not *common* causes of assignment to treatment groups and Behavior. Although these Other Influences affect Behavior, they did not affect assignment to the Therapy versus Behavior Modification groups because assignment to the treatment groups was random, based on the flip of a coin. This, then, is why true experiments are so powerful. True experiments still require an inference of cause and effect, but we can make that inference so powerfully because the act of random assignment effectively excludes all possible common causes of the presumed cause and the presumed effect. Random assignment assures that no other variables affect the presumed cause.

Intervening (Mediating) Variables

Given the admonition that models must include all common causes of the presumed cause and presumed effect, you may wonder how this applies to intervening or mediating variables. Do you also need to include all variables that *mediate* the effect of one variable on another? The answer is no; mediating variables are interesting because they help explain how an effect comes about, but they are not necessary for the model to be valid; in short, they are gravy. It is good that mediating variables are not required to make the model valid, because you could always include another layer of mediating variables. In the present example, you might wonder if Homework and Screen time mediate the effects of Parent Involvement on GPA (cf. Keith et al., 1993). That is, do parents influence their adolescents' learning, in part, by influencing them to complete more homework and spend less time on video games and TV watching? Suppose you found that these variables indeed mediated the influence of Parent Involvement on GPA; you might then wonder if the effects of Homework were mediated by time on task, and so on. Even with a seemingly direct relation, say the effect of smoking on lung cancer, we could posit and test indirect effects—the effect of smoking on buildup of carcinogens in the lungs, the effect of these chemicals on the individual cells, and so on. Again, it is not necessary to include indirect effects for models to be valid, but such indirect effects can help you understand *how* those effects come about.

In our current example, suppose that our central interest was the effect of Previous Achievement on GPA. If we were to conduct an analysis examining the direct effect of Previous Achievement on GPA without the intervening variable of Parent Involvement, the standardized direct (and total effect) would be .472. If you conduct the calculations for the indirect and total effects, you will find that the total effect of Previous Achievement on GPA for Figure 13.3 is also .472. When mediating or intervening variables are included in the model, the total effects do not change (although direct effects do); indirect effects are unnecessary for model validity.

I stress again, however, that although unnecessary for valid models, indirect effects are often very illuminating. Our current example suggests that Parent Involvement has a positive effect on GPA. But how does that effect come about? Previous research that tested for possible mediation by homework and TV viewing suggests that homework, in fact, partially mediates the effect of parent involvement on learning but that TV viewing does not (Keith et al., 1993). Parents who are more involved encourage, cajole, or force their children to do more homework, and this homework, in turn, raises their achievement. Although parents who are involved also influence their adolescents to spend less time watching TV, TV viewing appears to have little effect on achievement. Thus, leisure TV viewing does not appear to mediate the effect of parent involvement on achievement. As you become more expert in a particular area of research, you will likely find yourself asking questions about indirect or mediating effects. Indeed, even for those conducting experiments, indirect effects may often be of interest. Suppose you find that your experimental treatment (e.g., a new versus an established type of consultation) is effective; you may

next reasonably wonder why. Is it because the new consultation method improved problem identification, or speeded the time to intervention, or made evaluation more complete? Another advantage of mediating variables is that they can help strengthen the causal inferences embedded in path models. Logically, if you can explain both which variables affect an outcome and the mechanism *by which that effect occurs*, your causal claims are more believable. If we can demonstrate the indirect effect of smoking on lung cancer through the buildup of carcinogens in the lungs, it strengthens the case for smoking, as opposed to other characteristics of smokers, being a cause of lung cancer (Pearl, 2009). For additional information on testing mediating variables, see Baron and Kenny (1986), Hayes (2018); or MacKinnon (2008); see also the earlier discussion of mediation in Chapter 9.

OTHER POSSIBLE DANGERS

Paths in the Wrong Direction

Another possible danger in path analysis (and nonexperimental research in general) is that you may draw a path in the wrong direction. The implications of this danger depend on where this mistake takes place.

Figure 13.7 shows a model in which I erroneously assumed that 10th-grade GPA affected 8th-grade Parent Involvement. This model is clearly impossible, because it violates one of our primary assumptions, that cause cannot happen backward in time. The GPA variable occurs in 10th grade (although it is actually a measure of 9th- and 10th-grade GPA), whereas the Parent Involvement variable occurs in 8th grade. The model is clearly impossible. There is, however, nothing in the multiple regression analyses and nothing in the figure that would alert you to the fact that your model is incorrect. Indeed, the model leads you to completely erroneous conclusions about the moderate effect of 10th-grade GPA on 8th-grade Parent Involvement. Obviously, if the arrow between the two variables of prime interest is drawn in the wrong direction, the results will be completely and totally misleading.

In contrast, Figure 13.8 shows a model in which the path between Previous Achievement and Parent Involvement is drawn in the wrong direction. Previous Achievement was included in the model as a potential common cause of Parent Involvement and GPA, so the model in Figure 13.8 no longer controls for this variable as a *common cause*. Again, there is nothing

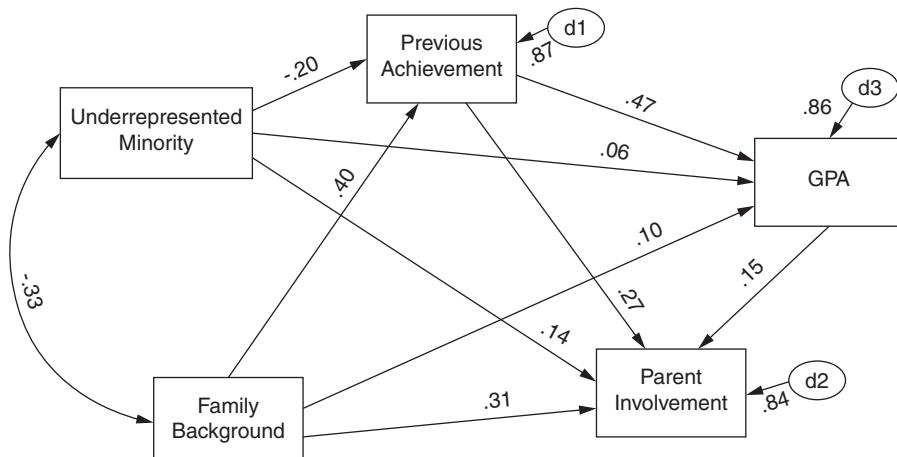


Figure 13.7 In this model the path between GPA and Parent Involvement is drawn in the wrong direction. There is nothing in the results to indicate that it is wrong.

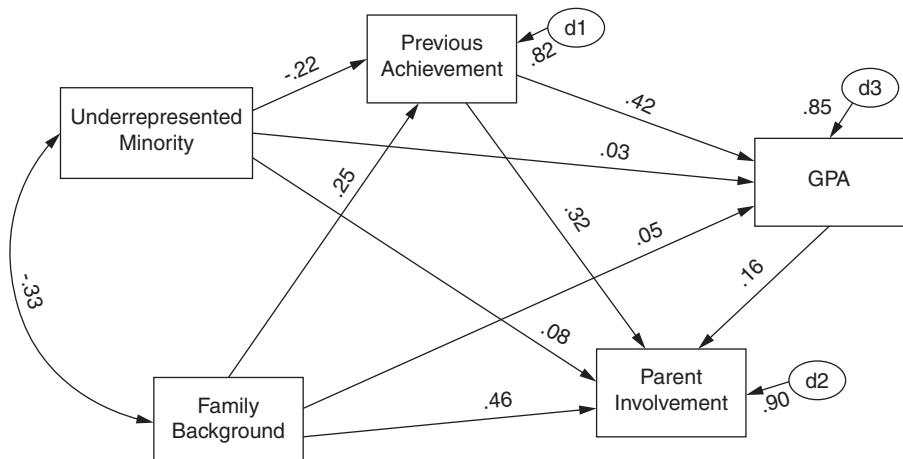


Figure 13.8 In this model the path between Parent Involvement and Previous Achievement is drawn in the wrong direction. The direct effects for these two variables remain the same, but the indirect and total effects differ from the “correct” model in Figure 13.3.

in any of the analyses to suggest that our model is incorrect. In this case, however, with the mistaken path being between our primary causal variable and a “control” variable, the findings are not quite as misleading. In fact, the direct effects of each variable in the model are the same as they were with the “correct” model shown in Figure 13.3. This makes sense when you realize that all paths to GPA are estimated via simultaneous MR, and for the models shown in Figures 13.3 and 13.8, both simultaneous regressions regressed GPA on each of the four variables in the model. What are incorrect in Figure 13.8 are the *total* effects. In Figure 13.3, Previous Achievement has an indirect effect on GPA through Parent Involvement, and thus its total effect is .472, compared to a direct effect of .417. In Figure 13.8, Previous Achievement is the next to last variable in the causal chain, so it has no indirect effect on GPA; its direct and total effects are both .417. For Parent Involvement, the reverse is true. In the “correct” model (Figure 13.3), Involvement had no indirect effect on GPA, so its direct and total effects were both equal to .160. In Figure 13.8, Involvement has an indirect effect on GPA through Previous Achievement, and thus we overestimate its total effect as .294. You should calculate the indirect and total effects yourself to make sure your estimates agree with mine, and that you understand the difference between these two models.

If the variables with paths drawn in the wrong direction are two of the less central variables, there should be little or no effect on the estimates of the primary variables of interest. For example, suppose the current example included a path from Underrepresented Minority to Family Background, rather than a correlation. Suppose further that we erred by drawing that path in the wrong direction (from Family Background to URM). This mistake will have no effect on the estimates of the direct, indirect, or total effects of Parent Involvement on GPA.

To summarize, if the effect, the final endogenous variable, is in the wrong position, estimates of all effects will be erroneous. If the primary causal variable has paths drawn in the wrong direction (but not the primary *effect* of interest), estimates of direct effects may still be accurate, but indirect and total effects will likely be incorrect. If background variables have paths drawn in the wrong direction, this error will likely not affect estimates of effects from the main cause variable to the main effect. These comments apply to just-identified models estimated through multiple regression but are not too far off for other, more complex models.

Reciprocal Causal Relations?

Given the problems resulting from paths drawn in the wrong direction, you may be tempted to be open-minded and proclaim that the variables are causally related in a reciprocal fashion in which not only does *a* affect *b* but *b* also affects *a*. Don't succumb to this temptation at this stage of your development! Although it is indeed possible to estimate such nonrecursive models, you cannot do so using multiple regression. You can estimate nonrecursive models using the SEM programs discussed in subsequent chapters, but such models are neither easy nor their results always illuminating. In my experience, reciprocal effects are also less common than you might think. Reserve the use of nonrecursive models for those cases in which you really think reciprocal effects may exist or for which you have legitimate, substantive questions about causal direction, not those for which you are simply unsure.

An even worse solution to this dilemma is to try to conduct the regression-path analysis both ways to see which "works best." You have already seen that the results of simple path analyses do not tell you when you have a path in the wrong direction. Likewise, the results of the analyses do not inform you as to which direction is best. Once again, theory, previous research, and logic are the appropriate tools for making such judgments.

I should note that, although the results of just-identified path analyses estimated through multiple regression cannot inform decisions about causal direction, properly overidentified models estimated through an SEM program may indeed be able to help with such decisions. In addition, well thought out nonrecursive models estimated via SEM programs can also be very informative about the nature and process of how one variable affects another. We will discuss these issues in later chapters.

Unreliability and Invalidity

One assumption underlying the causal interpretation of regression and path coefficients is that the exogenous variables are measured with near perfect reliability and validity. With our current model, URM may come close to meeting this assumption, but the variable Family Background, a composite of Parent Education, Parent Occupational Status, and Family Income, certainly does not. We obviously regularly violate this assumption but will postpone until later chapters a discussion of the effects of this violation and possible solutions.

DEALING WITH DANGER

The two primary dangers of path analysis are (1) that you have neglected to include in your model an important common cause of the variable you think of as your primary cause and the variable you think of as your primary effect and (2) that you have drawn paths in the wrong direction; that is, you have confused cause and effect. In the jargon of SEM, these are generally termed specification errors, or errors in the model. Of these two, I consider the first the most common and insidious. In most cases, it should be pretty obvious when you draw a path in the wrong direction. What can you do to avoid these errors?

My first response is to say, "Welcome to the dangerous world of structural equation modeling; join us SEMers on the wild side!" More seriously, I again remind you that these same dangers apply to *any* nonexperimental research, no matter how that research is analyzed. One advantage of path analysis and structural equation modeling, in my opinion, is the requirement of a theory, generally expressed figuratively in a path model, prior to analysis. It is much easier to spot missing common causes and causal assumptions in the wrong direction when examining a path model than it is when reading the description of, say, a MR analysis. Furthermore, these dangers apply to *all* research, experimental or nonexperimental, in which we wish to infer cause and effect. A true experiment allows a powerful inference of cause and effect by knocking one leg

out from under the danger of common causes, but the farther we stray from the true experimental ideal of random assignment to treatment groups, the more real this danger becomes. Indeed, many concerns with quasi experimental research (e.g., research using matched groups rather than random assignment) boil down to concerns over unmeasured common causes. With a true experiment we also actively manipulate the independent variable, the presumed cause, thus making true experiments less likely to confuse causal direction, as well.

We have seen that the analyses themselves do not guard against these errors; they do not tell us when our models are wrong or when we have neglected an important common cause. How, then, to avoid these specification errors? I come back to the same refrain: understand relevant theory; be familiar with the research literature; spend time puzzling over your model, especially thinking about potential common causes and potential problems in direction; and draw your model carefully.

These same concerns and dangers apply when you are a consumer and reader of others' research. As you read others' nonexperimental research, you should ask yourself whether the researchers neglected to include any important common causes of their presumed cause and presumed effect. If so, the results of the research will be misleading and likely overestimate (or underestimate) the effect of one variable on another. Armchair analysis is not sufficient, however; it is not valid to simply say, "Well, I think variable Z is a probable common cause of variables X and Y ," and expect to have your concerns taken seriously. You should be able to demonstrate, through theory, previous research, or analysis, that variable Z is indeed a likely and important common cause. Likewise, as you read nonexperimental research, you should be attuned to whether any of the causal assumptions are reversed. Again, you should be able to demonstrate this incorrect causal direction through theory, research, logic, or your own analyses.

We will revisit the danger of measurement error and its effects. For the time being, you should simply strive to make sure that all your variables, and especially scores on your exogenous variables, are as reliable and valid as possible.

REVIEW: STEPS IN A PATH ANALYSIS

Let's review the steps involved in path analysis now that we've carefully considered the dangers.

1. First, spend some time thinking about the problem; how might these variables of interest be causally related?
2. Draw a tentative model.
3. Study relevant theory and research. Which variables must be included in the analysis? You must include the relevant common causes, but not every variable under the sun. The relevant theory and research, along with careful thought, will also help you resolve questions of causal direction. "The study of structural equation models can be divided into two parts: the easy part and the hard part" (Duncan, 1975, p. 149). This step is the hard part of SEM.
4. Revise the model. It should be lean, but include all necessary variables.
5. Collect a sample and measure the variables in the model, or find a data set in which the variables are already measured. Use variables that have strong estimates of reliability and validity.
6. Check the identification status of the model. Make sure the model is just-identified or overidentified.
7. Estimate the model.
8. Fill in the model estimates (paths and disturbances) in your figure. Are the paths more or less as expected? That is, are the paths you expected to be positive in fact positive; those you expected to be negative, negative, and those that you expected to be close to zero in fact close to zero? Meeting such expectations allows more confidence in your model.
9. Write up the results and publish them.

Some writers recommend *theory trimming* in between my steps 8 and 9. Theory trimming means deleting statistically nonsignificant paths and re-estimating the model. I do not recommend this step, especially when using multiple regression to solve for the paths. We will return to this issue in the next chapter.

SUMMARY

The chapter began by reiterating the basic assumptions of multiple regression: linearity, independence, and homoscedasticity. For regression coefficients to provide accurate estimates of effects, the disturbances should be uncorrelated with the exogenous variables. This assumption will likely be fulfilled if there is no reverse causation, the exogenous variables are perfectly measured, equilibrium has been achieved, and there are no omitted common causes in the model.

These assumptions led to a discussion of the dangers of path analysis. When a common cause (a variable that affects both a presumed cause and a presumed effect) is omitted from a model, this omission changes the estimate of the influence of one variable on another. The most common result is that we end up overestimating the effect, although underestimation is also possible. The dreaded *spurious correlation* is a result of an omitted common cause, and thus omitted common causes are the primary reason for the admonition about inferring causation from correlations. I illustrated the effects of omitting a common cause through a research example testing the effects of Parent Involvement on 10th-grade GPA. When Previous Achievement, a common cause of Involvement and GPA, was omitted from the model, we overestimated the effect of Parent Involvement on GPA. Omitted common causes may be a reason for variability in research findings in nonexperimental research.

The warning to include common causes should not be interpreted as a mandate to include all causes of the presumed cause and the presumed effect. Only variables that affect both the presumed cause and presumed effect must be included. We illustrated the difference between a cause and a common cause by deleting Ethnic background from the model. URM affected Parent Involvement but not GPA, and thus was not a *common cause* of the two variables. As a result, when URM was removed from the model, the estimate of the effect of Involvement on GPA barely changed. The main reason that true experiments allow such a powerful inference of causality is because, through the act of random assignment, such research rules out possible common causes of the independent (cause) and dependent (effect) variable, even though experiments do not rule out all causes of the dependent variable.

The warning to include common causes also does not extend to mediating or intervening variables. When an intervening variable is included in the model, the total effects remain the same, but a portion of the direct effect of X on Y becomes indirect effect through the mediating variable. Intervening variables help explain *how* an effect comes about but do not need to be included for the model to be valid.

Estimates of effects are also incorrect when paths are drawn in the wrong direction, although the extent of the problem depends on the paths involved. If the incorrect path is from the effect to the cause, the results will obviously be incorrect and completely misleading. If the incorrect path involves the primary causal variable and one of the other causal variables in the model, this error will affect the total effects but not the direct effects. If the incorrectly drawn path involves some of the background variables in the model, this error should have little effect on the estimates of primary interest (although it will make attentive readers less trusting of your results!). We will revisit and address this danger in subsequent chapters.

How, then, can you be sure that your model is correct? Have a good understanding of relevant theory and previous research. Think about the variables in your model, how they are

related to one another. If necessary, bolster causal assumptions (e.g., *a* affects *b*, rather than *b* affects *a*) through the use of longitudinal data. Think about possible common causes and investigate them in the research literature. If necessary, test common causes in the research itself. In fact, most of what you should do to ensure the adequacy of your model boils down to the same advice for drawing a model in the first place: theory, previous research, and logic.

I also noted that, as a reader or reviewer of others' nonexperimental research, it is not enough to guess about neglected common causes; you should be able to demonstrate such criticisms through theory, previous research, or independent analysis. Finally, I noted again that these dangers apply to all nonexperimental research, no matter how it is analyzed. One advantage of path models is that the figural display of one's model (in essence a mini theory) often makes errors and assumptions more obvious and therefore more likely to be corrected. We postponed dealing with the violation of the assumption of perfect measurement until later chapters.

EXERCISES

1. Conduct each of the parent involvement analyses reported in this chapter, using the NELS data. The variables, as listed in NELS, are: Underrepresented Minority = URM; Family Background = BySES; Previous Achievement = ByTests; Parent Involvement = Par_Inv; and GPA = FfuGrad. Compare your results to mine.
 - a. Make sure you understand what happens when a common cause is omitted versus a simple cause of only one of the variables of interest (Figures 13.3 through 13.4). Is Family Background a common cause or a simple cause of Parent Involvement and GPA? Try deleting it from the model; what happens to the path from Involvement to GPA?
 - b. Analyze a model without Parent Involvement. Calculate direct, total, and indirect effects for each variable on GPA. Do the same for the model shown in Figure 13.3. Compare the tables of direct, indirect, and total effects.
 - c. Analyze a model like Figure 13.3, but in which a path is drawn from URM to Family Background. Now analyze a model in which the path is drawn from Family Background to URM. Which model is correct? How did you make this decision? What effect, if any, did this change in direction have on the estimate of the effect of Parent Involvement on GPA?
2. Find an article that uses path analysis or explanatory multiple regression on a research topic with which you are familiar and interested. If the authors' model is not drawn in the article, see if you can draw it from their description. How do the authors justify their causal assumptions or their paths? Do you agree, or do you think some of the paths are drawn in the wrong direction? Do you think there are any obvious common causes that have not been included in the model? Can you demonstrate that there are common causes that have been neglected? If the authors included a correlation matrix with their article, see if you can reproduce their results. Draw the estimated model.
3. In Chapter 12, you constructed and tested a path model using the variables Family Background (BYSES), 8th-grade GPA (BYGrads), 10th-grade Self-Esteem (F1Concept2), 10th-grade Locus of Control (F1Locus2), and 10th-Grade Social Studies Achievement (F1TxHStd). Refer to or redo the analysis. For the sake of consistency, make sure you have Social Studies Achievement as the final endogenous variable.
 - a. Notice the direct effects of Self-Esteem and Locus of Control on Social Studies Achievement. Focus on the effect of GPA on Self-Esteem and Locus of Control. Is 8th-grade GPA a common cause of these variables and Social Studies Achievement? Now remove the 8th-grade GPA variable from the model. What happens to direct

- effects of Self-Esteem and Locus of Control on Social Studies Achievement? Explain the difference in effects from the original model.
- b. Did you draw a path from Self-Esteem to Locus of Control or Locus of Control to Self-Esteem? Calculate the direct, indirect, and total effects of these two variables on Social Studies Achievement. Whichever way you drew the path, now reverse the direction and re-estimate the model. Recalculate the direct, indirect, and total effects of these variables on Social Studies. Explain the differences you find.

Notes

- 1 Underrepresented ethnic minority, or URM, is coded so that students from African American, Hispanic, and Native backgrounds are coded 1 and students of Asian and Caucasian descent are coded 0. URM is a categorization sometimes used in college settings, such as admissions and science, technology and engineering fields. See, for example, <http://www.nacme.org/underrepresented-minorities>. As noted in Chapter 7, if our primary interest were in the effect of ethnic background on achievement, then we would likely be better served using a series of categorical variables, such as Asian v White, Hispanic v White, and so on. Here, however, URM is included as a background variable that we control rather than as a variable of primary interest.
- 2 If the common cause has positive effects on both the presumed cause and the presumed effect, its neglect will lead to an overestimate of the effect of the presumed cause on the presumed effect. If a common cause has a negative effect on either variable, its omission will lead to an underestimate, and if it has a negative effect on both, its omission will result in an overestimate.
- 3 There may be other advantages for including a variable in the model that is not a common cause. For example, inclusion of noncommon causes results in overidentified models, the advantages of which we will discuss in the following chapter.

14

Analyzing Path Models Using SEM Programs

SEM Programs	296
<i>Amos and Mplus</i>	297
Reanalysis of the Parent Involvement Path Model	298
<i>Estimating the Parent Involvement Model via Amos</i>	299
Advantages of SEM Programs	303
<i>Overidentified Models</i>	303
<i>Comparing Competing Models</i>	312
More Complex Models	315
<i>Equivalent and Nonequivalent Models</i>	315
<i>Nonrecursive Models</i>	322
<i>Longitudinal Models</i>	323
Advice: MR Versus SEM Programs	325
Advice: Measures of Fit	326
<i>Evaluating a Single Model</i>	326
<i>Comparing Competing Models</i>	328
Summary	328
Exercises	330
<i>Notes</i>	332

To this point I have used multiple regression for the analysis of path models (as well as multiple regression models). It is also possible to use dedicated structural equation modeling (SEM) programs for such analysis. We make that switch in this chapter. As you will see, the results of simple path analyses are identical using SEM or MR analysis, but SEM programs can analyze more complex models and have real advantages when analyzing overidentified models.

SEM PROGRAMS

Numerous SEM programs are available, all of which are capable of analyzing everything from simple path models through latent variable structural equation models. LISREL (*Linear Structural Relations*; Jöreskog & Sörbom, 1996; Mels, 2006) was the first such program and is still widely used. For additional information, go to www.ssicentral.com. Other common programs include EQS (Bentler, 1995; www.mvsoft.com) and Mplus (Muthén & Muthén, 1998–2017; www.statmodel.com). Each such program has its own advantages; Mplus, for example,

has sophisticated routines for analyzing categorical variables and is perhaps the most flexible such program. They generally cost \$500 to \$600 for those in academia, and many of them have try-out or trial versions and reduced pricing for students. For users of R (a free statistical programming language), there are at least three free SEM add-ons for that I know of, OpenMx (<https://openmx.ssri.psu.edu/>; Boker et al., 2012), sem (<https://cran.r-project.org/web/packages/sem/>; Fox, 2006), and lavaan (<http://lavaan.ugent.be/>; Beaujean, 2014; Rosseel, 2012). Ω nyx is a free, stand-alone SEM program which, like Amos, uses a graphical interface (<http://onyx.brandmaier.de/>). The general purpose statistics programs SAS (www.sas.com/en_us/software/stat.html) and STATA (www.stata.com/) also have SEM capabilities.

Amos and Mplus

My favorite teaching program is one called Amos (Analysis of Moment Structures; Arbuckle, 2013; www.spss.com/amos), although I also use Mplus on a regular basis. Amos uses a graphic approach and is probably the most intuitive and easiest SEM program to use. It produces attractive path diagrams (all the path models you have seen so far were produced by Amos) and can be used both to draw a path diagram and analyze it. As of this writing, student pricing for Amos is around \$50 per year as an SPSS for Windows add-on (there is no Mac version). The user's guide for the most recent version is also available as a pdf document on the spss website (under product support) and you can download the program as a free try-out for 30 days. Of course, you can analyze these problems using any SEM program, so if you have another program available you may want to use it. As noted, there are also student or demo versions available of many of the commercial SEM programs.

There are numerous examples of Amos and Mplus input and output—at least one per chapter—on the website (www.tzkeith.com). Statistical programs are revised constantly, so check the website also for more up-to-date information than is contained here in the text. Whatever program you use, you should download or purchase the user's manual, which provides the basics for the use of the program. There are numerous other sources of information about various SEM programs, as well. If you use Amos, for example, I recommend you download a tutorial from https://stat.utexas.edu/images/SSC/documents/SoftwareTutorials/AMOS_Tutorial.pdf. This site also has tutorials for other SEM and general statistics programs, and there are, of course, a growing number of such resources on the web. Although I will generally use Amos or Mplus to estimate subsequent models, the information presented applies to SEM programs in general.

Basics of SEM Programs

Everything you have learned about path analysis so far will transfer to Amos and other SEM programs. Figure 14.1 shows a basic SEM (Amos) version of the parent involvement model first presented in Chapter 13. As in all previous examples, rectangles represent measured variables, and ovals represent unmeasured or latent variables (in this example, the disturbances). Straight arrows represent paths, or presumed influences, and curved, double-headed arrows represent correlations (or, with unstandardized coefficients, covariances). The one new aspect of Figure 14.1 is the value of 1 beside the paths from the disturbances to the endogenous variables. These paths simply set the scale of measurement for the disturbances. Unmeasured variables have no natural scale. When we set the path from the disturbances, which are unmeasured variables, to the measured variables to 1.0, we are merely telling the SEM program that the disturbance should have the same scale as the measured variable. (In reality, any number could be used: .80, 2.0, but 1.0 is the most common and most straightforward.) We will use the same rule of thumb when we begin using other latent variables: we will set one path from each latent variable to 1.0 to set the scale of the latent variable. At

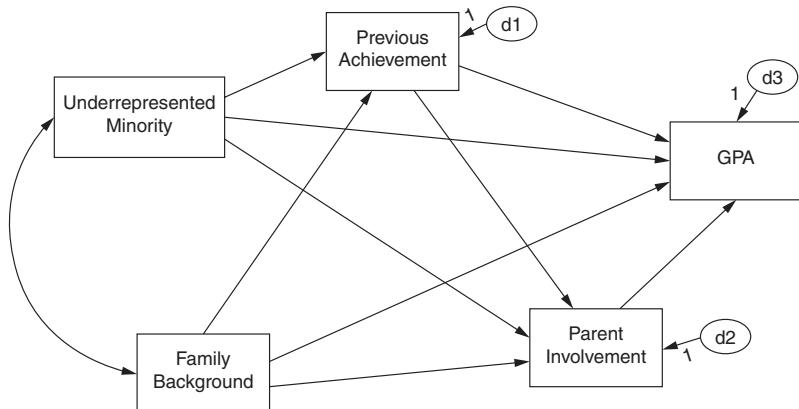


Figure 14.1 Parent Involvement model from Chapter 13, as drawn in the Amos SEM program.

a practical level, the model would be underidentified without this constraint (or some other way of setting the scale of the disturbances). Depending on which program you use, these paths from disturbances to endogenous variables may be set to 1 automatically and invisibly (for example, this happens by default in MPlus).

We could also set the scale by fixing the variance of the disturbance to 1.0; all substantive results would be the same. In fact, this is exactly what we did with multiple regression, even though we did not realize that we were doing so. When we use multiple regression to estimate the paths, the variances of the disturbances are set to 1.0, and the program estimates the paths from the disturbances to the endogenous variables (when we set the path to 1.0, the program estimates the variance of the disturbance). We can choose either method with Amos; I here set the paths to 1.0 because that is the most common method.

REANALYSIS OF THE PARENT INVOLVEMENT PATH MODEL

The model shown in Figure 14.1 provides the basic input for analysis by Amos (the model is on the Web site as “PI Example 1.amw”); add data and you can conduct the analysis. Most SEM programs, including Amos, can use the correlation matrix and standard deviations as input for the analysis. The matrix for this example is saved as both an SPSS (PI matrix, listwise.sav) and an Excel file (PI matrix, listwise.xls). The matrix is also shown in Table 14.1; the

Table 14.1 Means, Standard Deviations, Sample Sizes, and Correlations among the Variables for the Parent Involvement Path Example

Variable	URM	BYSES	BYTests	Par_Inv	FFUGrad
URM	1.000				
BYSES	-.333	1.000			
BYTests	-.330	.461	1.000		
Par_Inv	-.075	.432	.445	1.000	
FFUGrad	-.131	.299	.499	.364	1.000
<i>M</i>	.207	.047	52.323	.059	5.760
<i>SD</i>	.406	.766	8.584	.794	1.450
<i>N</i>	811	811	811	811	811

variable names are as in the NELS raw data. The SPSS commands I used to create the matrix using the NELS data are in the file “create corr matrix in spss.sps.”

Estimating the Parent Involvement Model via Amos

With the model and the data, we can estimate the model via Amos (or any other SEM program). The standardized output for this model is shown in Figure 14.2. Compare the results with your results from Chapter 13; with the exception of the lack of a number associated with the paths from disturbances to endogenous variables, the results should be identical. Figure 14.3 shows the unstandardized output for the model. Recall that we set the paths from disturbances to endogenous variables to 1.0 and estimated the variances of the disturbances. The numbers next to the disturbances are the estimates of their variances. The numbers above the two exogenous variables are their variances.

Again, the results should match those from your regression analysis in Chapter 13.

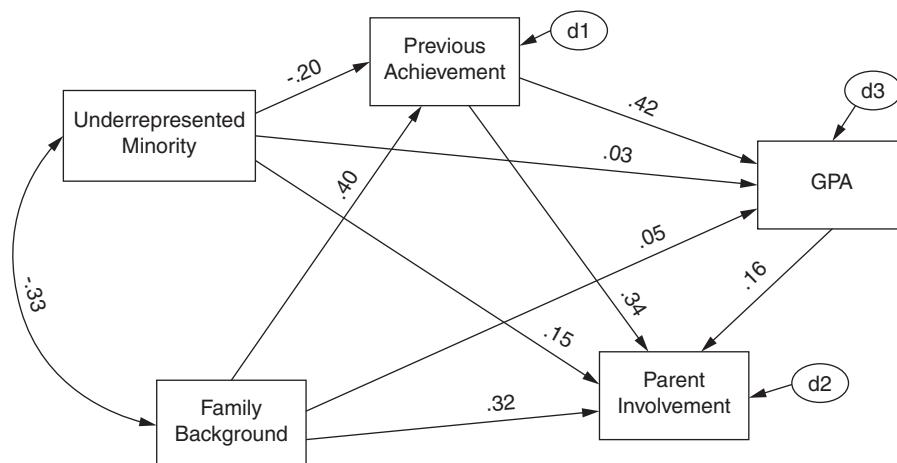


Figure 14.2 Parent Involvement model estimated via Amos. The standardized results are the same as those in Chapter 13 when the model was estimated via multiple regression.

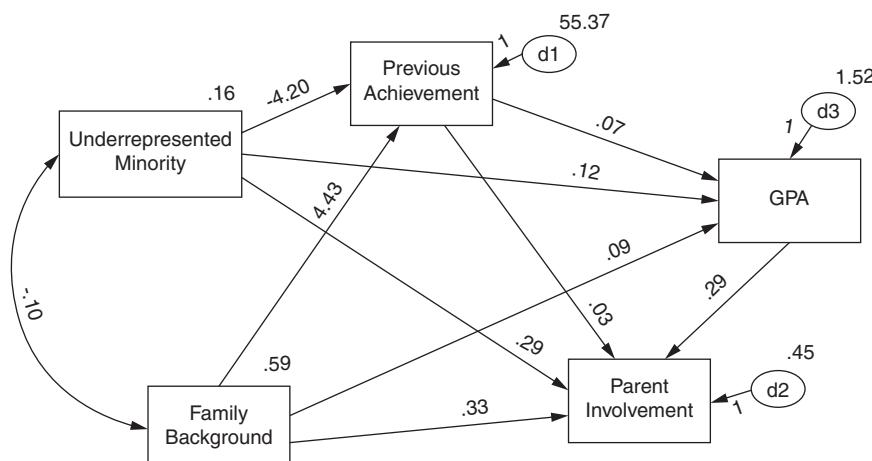


Figure 14.3 Unstandardized estimates for the Parent Involvement model.

Of course, you will get more detailed output than just these diagrams from your SEM program. Figure 14.4 shows one portion of the printout; this and all subsequent printouts show Amos output, but you will get something similar with any of the SEM programs. The top portion of the output (Regression Weights) shows the unstandardized path coefficients, listed under the column Estimate. For example, the first row shows that the unstandardized path from BYSES (Family Background) to BYTests (Previous Achievement) is 4.431 (it is possible to have Amos list the longer variable labels in addition to the variable names, but just the names are shown in this output). The S.E. column shows the standard errors of the

Regression Weights

			Estimate	S.E.	C.R.	P	Label
bytests	<---	byses	4.431	.362	12.229	***	
bytests	<---	URM	-4.195	.684	-6.131	***	
par_inv	<---	bytests	.032	.003	10.034	***	
par_inv	<---	URM	.286	.063	4.525	***	
par_inv	<---	byses	.333	.036	9.345	***	
ffugrad	<---	bytests	.070	.006	11.406	***	
ffugrad	<---	par_inv	.292	.064	4.528	***	
ffugrad	<---	byses	.093	.069	1.354	.176	
ffugrad	<---	URM	.124	.117	1.057	.290	

Standardized Regression Weights

		Estimate	
bytests	<---	.395	
bytests	<---	URM	-.198
par_inv	<---	bytests	.345
par_inv	<---	URM	.146
par_inv	<---	byses	.321
ffugrad	<---	bytests	.417
ffugrad	<---	par_inv	.160
ffugrad	<---	byses	.049
ffugrad	<---	URM	.035

Covariances

		Estimate	S.E.	C.R.	P	Label
URM	<->	byses	-.103	.011	-8.996	***

Correlations

		Estimate	
URM	<->	byses	-.333

Variances

	Estimate	S.E.	C.R.	P	Label
URM	.164	.008	20.125	***	
byses	.586	.029	20.125	***	
d1	55.369	2.751	20.125	***	
d2	.452	.022	20.125	***	
d3	1.521	.076	20.125	***	

Figure 14.4 Output from the SEM program (Amos) showing unstandardized coefficients (regression weights), their standard errors, and critical ratios, along with standardized coefficients.

coefficients, the column labeled C.R. (for critical ratio) shows the z 's for each coefficient. (Recall that $t = \text{coefficient}/SE_{\text{coefficient}}$ and that with large samples t 's greater than approximately 2 are statistically significant. The values are better considered z statistics, but they are essentially the same with the sample sizes we are using.) The column labeled P shows the probability associated with each path, with values less than .001 indicated by ***. The next portion of the figure (Standardized Regression Weights) shows the standardized paths. Again, the output should match the SPSS output from Chapter 13. This portion is followed by the covariance and correlation between the two exogenous variables, the variances of the two exogenous variables and the variances of the disturbances of the three endogenous variables.

SEM programs will also produce tables of direct, indirect, and total effects for both the standardized and unstandardized solution. The tables for the current example are shown in Figure 14.5. The tables are read from column to row; thus the total unstandardized effect of

Total Effects

	byses	URM	bytests	par_inv
bytests	4.431	-4.195	.000	.000
par_inv	.474	.153	.032	.000
ffugrad	.544	-.127	.080	.292

Standardized Total Effects

	byses	URM	bytests	par_inv
bytests	.395	-.198	.000	.000
par_inv	.458	.078	.345	.000
ffugrad	.287	-.035	.472	.160

Direct Effects

	byses	URM	bytests	par_inv
bytests	4.431	-4.195	.000	.000
par_inv	.333	.286	.032	.000
ffugrad	.093	.124	.070	.292

Standardized Direct Effects

	byses	URM	bytests	par_inv
bytests	.395	-.198	.000	.000
par_inv	.321	.146	.345	.000
ffugrad	.049	.035	.417	.160

Indirect Effects

	byses	URM	bytests	par_inv
bytests	.000	.000	.000	.000
par_inv	.141	-.134	.000	.000
ffugrad	.450	-.251	.009	.000

Standardized Indirect Effects

	byses	URM	bytests	par_inv
bytests	.000	.000	.000	.000
par_inv	.136	-.068	.000	.000
ffugrad	.238	-.070	.055	.000

Figure 14.5 Total, indirect, and direct effects of variables on each other in the Parent Involvement model.

Family Background (BYES) on GPA (ffugrad), as shown in the bottom left of the first table, is .544. Take some time to compare these results with those from the previous chapter.

It is also possible to evaluate the statistical significance of the indirect and total effects; in Amos this is done through a bootstrapping procedure. (Recall from Chapter 9 in our discussion of mediation that *bootstrapping* is a procedure in which one takes repeated, smaller random samples of an existing sample. With bootstrapping, it is possible to develop empirical estimates of standard errors of any parameter, even, for example, standard errors of standard errors. Recall also that bootstrapping is generally recognized as a preferred method of calculating standard errors, confidence intervals, and the statistical significance of indirect effects. Figure 14.6, for example, shows the indirect effects for the variables in the Parent Involvement model, followed by their standard errors. You can use this information to calculate the *z* values for each indirect effect to determine its statistical significance; this

Indirect Effects

	byses	URM	bytests	par_inv
bytests	.000	.000	.000	.000
par_inv	.141	-.134	.000	.000
ffugrad	.450	-.251	.009	.000

Standardized Indirect Effects

	byses	URM	bytests	par_inv
bytests	.000	.000	.000	.000
par_inv	.136	-.068	.000	.000
ffugrad	.238	-.070	.055	.000

Indirect Effects - Standard Errors

	byses	URM	bytests	par_inv
bytests	.000	.000	.000	.000
par_inv	.018	.026	.000	.000
ffugrad	.045	.066	.002	.000

Standardized Indirect Effects - Standard Errors

	byses	URM	bytests	par_inv
bytests	.000	.000	.000	.000
par_inv	.017	.013	.000	.000
ffugrad	.023	.018	.013	.000

Indirect Effects - Two Tailed Significance (BC)

	byses	URM	bytests	par_inv
bytests
par_inv	.001	.001
ffugrad	.001	.001	.001	...

Standardized Indirect Effects

	byses	URM	bytests	par_inv
bytests
par_inv	.001	.001
ffugrad	.001	.001	.001	...

Figure 14.6 Indirect effects (both unstandardized and standardized) for the Parent Involvement model, their standard errors, and statistical significance.

is done in the bottom of the figure. Thus, SEM programs allow a direct test of the statistical significance of mediation (see in Chapter 9 the section on Mediation). Amos also provides the *standardized indirect and total effects* and their standard errors and statistical significance (the standardized indirect effects are shown in the figure).

ADVANTAGES OF SEM PROGRAMS

Overidentified Models

Figure 14.7 shows a potential model of the effect of Homework on GPA. The data are from NELS (the larger NELS data, not those on the Web site). For this model, Minority, Family Background, and Previous Achievement were measured in eighth grade and are defined in the way we have in the past (Minority, or ethnic minority background = ethnic minority =1, white=0, Family Background = BYSES, Previous Achievement = BYTests). Homework was based on student reports of time spent on homework in each academic area, measured in both eighth and tenth grades; it may be considered a measure of average homework over time. Grades are students' GPAs (English, Math, Science, and Social Studies) from 10th grade.

Note that several potential paths are not drawn: there are no paths from Minority and Family Background to Grades. Just as it means something to draw a path, it means something to not draw a path and, in fact, it is often a *stronger statement* than drawing a path. When we draw a path, we are stating that one variable may have some effect on another. What the *lack of path* from Family Background to Grades means is that I believe the path from Background to Grades is a value of zero. Indeed, not drawing a path is the same as drawing a path and fixing or constraining that path to a value of zero. This model also makes explicit the notion that the only way Minority and Family Background affect Grades is through Homework and Previous Achievement, that they have no direct effect on Grades, only indirect effects through other variables in the model. I developed this hypothesis in the usual way, based on previous research and logic. Indeed, you will even find support for the exclusion of paths

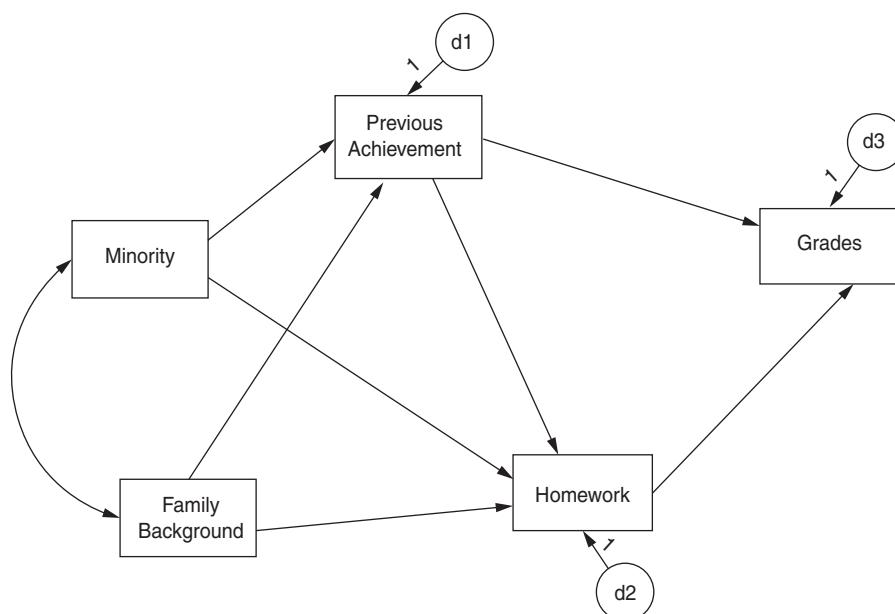


Figure 14.7 Overidentified model testing the effects of Homework on students' Grades in High School.

from Minority and Family Background to Grades based on our Parent Involvement models, which showed only small direct effects for similar variables on Grades.

You know from Chapter 12 that this is an overidentified model, meaning that we have more information than we need to solve for the paths. Note that there are 10 correlations among variables, but we are solving for only eight parameters (seven paths and one correlation). Recall also that if we were solving for the paths using algebra we could come up with multiple formulas for solving some of the paths. I argued in Chapter 12 that this approach may have advantages, because similarity in path estimates calculated two different ways can give us additional confidence in our model, whereas dissimilarity might make us wonder about the veracity of the model.

One advantage of SEM programs is that they provide this type of feedback about overidentified models. The method is not as described above; the programs do not estimate the paths several different ways and allow you to compare the different estimates. Instead, the programs compare matrices and provide measures of the fit of the model to the data. We'll see how this process works when we analyze the model in Figure 14.7.

The data (correlation matrix, standard deviations, means, and N) are contained in both an Excel and an SPSS file ("homework overid 2018.xls" and "homework overid 2018.sav"); the data are also shown in Table 14.2.¹ The file includes standard deviations, sample sizes, and correlations. Means are included but are not required or analyzed. The model shown in the figure was used as input to Amos and is in the file "homework path 1.amw" on the accompanying Web site.

Figure 14.8 shows the solved, standardized path model. Using our rules of thumb, it appears Homework has a moderate effect (.15) on 10th-grade GPA. Previous Achievement had a strong effect on Homework, suggesting a "rich get richer" sort of effect: students who achieve at a high level do more homework, and this homework, in turn, improves their subsequent school performance. Family Background also had a moderate effect on Homework, but Ethnic background (Minority) had no substantive effect.

How can we use the overidentification status of the model to assess the model? Recall how we solved for the paths in our first example of path analysis: through the use of algebra, the tracing rule, and the correlations among the variables. Amos is essentially doing the same thing here: the model specification and the correlation matrix (actually the covariance matrix, but we will address this point later) were used as input, and the program used these pieces of information to solve for the paths. If we can solve for the paths using the

Table 14.2 Contents of the Excel file for the homework path example

<i>rowtype_</i>	<i>varname_</i>	<i>Minority</i>	<i>FamBack</i>	<i>PreAch</i>	<i>Homework</i>	<i>Grades</i>
n		1000	1000	1000	1000	1000
corr	Minority	1				
corr	FamBack	-.3041	1.0000			
corr	PreAch	-.3228	.4793	1.0000		
corr	Homework	-.0832	.2632	.2884	1.0000	
corr	Grades	-.1315	.2751	.4890	.2813	1.0000
stddev		.4186	.8311	8.8978	.8063	1.4790
mean		.2718	.0025	52.0039	2.5650	5.7508

Note: The matrix is in the format required for analysis in Amos. These include the *rowtype_* and *varname_* columns, and the n, corr, and stddev rows (the mean row is not required at this stage of our adventures).

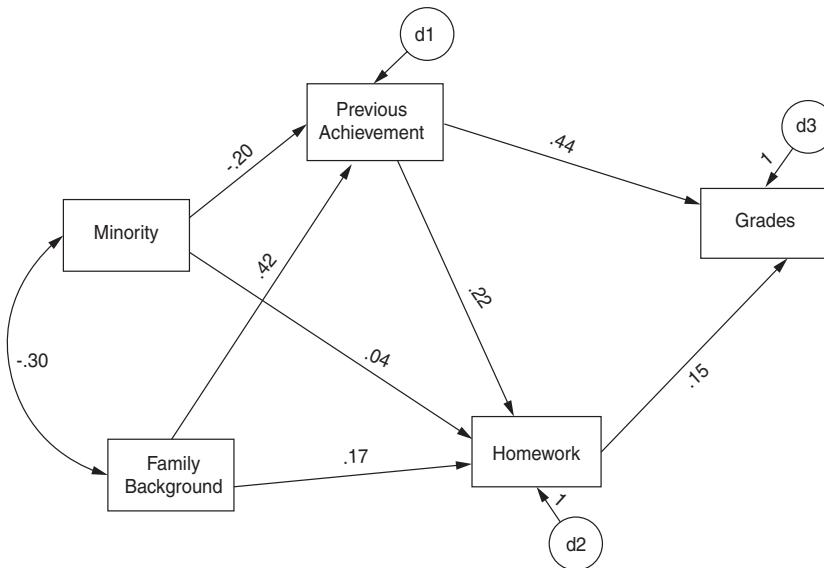


Figure 14.8 Standardized output for the Homework model.

Sample Correlations

	Minority	FamBack	PreAch	Homework	Grades
Minority	1.000				
FamBack	-.304	1.000			
PreAch	-.323	.479	1.000		
Homework	-.083	.263	.288	1.000	
Grades	-.132	.275	.489	.281	1.000

Implied (for all variables) Correlations

	Minority	FamBack	PreAch	Homework	Grades
Minority	1.000				
FamBack	-.304	1.000			
PreAch	-.323	.479	1.000		
Homework	-.083	.263	.288	1.000	
Grades	-.156	.254	.489	.281	1.000

Figure 14.9 The sample (input) correlation matrix compared to the matrix implied by the Homework model.

correlations, why can't we do the reverse: solve for the correlations using the paths? In fact, we can do exactly that. You could, and SEM programs do, use the solved path model (e.g., Figure 14.8) to calculate an expected, or predicted, correlation matrix, the matrix implied by the model.² With overidentified models, this implied matrix (also known as the predicted matrix) and the input matrix will differ to some degree. The actual correlation matrix and the implied correlation matrix from the Amos output are shown in Figure 14.9. Notice that most of the correlations are the same, but that the values in the lower left—the correlations of Grades with Minority and Family Background—differ slightly between the actual and the implied matrices. SEM programs use this degree of similarity or nonsimilarity between the two matrices to assess and measure the fit of the model to the data. This information is also useful for diagnosing and correcting model problems.

Correlations Versus Covariances

Before going any further, it is time to augment our thinking about correlation matrices with the additional considerations of covariance matrices. Most SEM programs are set up to analyze covariance rather than correlation matrices. For some SEM problems you will get the same substantive answer no matter which type of matrix you analyze, but for others you should analyze covariance matrices (see Cudeck, 1989, or Steiger, 2001, for further discussion about this issue). An easy solution is simply to get in the habit of analyzing covariance, rather than correlation, matrices. (An alternative is to use a program, such as SEPATH, a part of the Statistica package, designed specifically to analyze correlation matrices.)

Recall from Chapter 1 that we can easily calculate covariances from correlations if we know the variances or standard deviations of the variables, because $Cov_{xy} = r_{xy} \times SD_x SD_y$. Indeed, this is what Amos did; we input the correlations and standard deviations, and the program generated the covariance matrix from that input. The covariance matrix is shown at the top portion of Figure 14.10. The covariances are shown below the diagonal, and the

Sample Covariances

	Minority	FamBack	PreAch	Homework	Grades
Minority	.175				
FamBack	-.106	.690			
PreAch	-1.201	3.541	79.092		
Homework	-.028	.176	2.067	.649	
Grades	-.081	.338	6.429	.335	2.185

Implied (for all variables) Covariances

	Minority	FamBack	PreAch	Homework	Grades
Minority	.175				
FamBack	-.106	.690			
PreAch	-1.201	3.541	79.081		
Homework	-.028	.176	2.067	.649	
Grades	-.097	.311	6.428	.335	2.185

Residual Covariances

	Minority	FamBack	PreAch	Homework	Grades
Minority	.000				
FamBack	.000	.000			
PreAch	.000	.000	.010		
Homework	.000	.000	.000	.000	
Grades	.015	.027	.001	.000	.000

Standardized Residual Covariances

	Minority	FamBack	PreAch	Homework	Grades
Minority	.000				
FamBack	.000	.000			
PreAch	.000	.000	.003		
Homework	.000	.000	.001	.000	
Grades	.776	.662	.002	.000	.001

Figure 14.10 Sample and implied covariance matrices and residual and standardized residual matrices for the Homework model.

variances are shown in the diagonal. Another way of thinking of covariance versus correlation matrices is to recall that correlation matrices are standardized covariance matrices, with all variables converted to z scores.

Model Fit and Degrees of Freedom

SEM programs, then, generally compare the actual covariance matrix to the implied covariance matrix. Some of the relevant output from Amos is shown in Figure 14.10: the actual covariance matrix, the implied matrix, and the residual covariance matrix. The residual covariance matrix is the result of subtracting the implied matrix from the actual matrix; intuitively, large differences between these matrices and large residuals should signal problems with the model. More helpful are the *standardized* residuals, in which the residuals have been converted to a common, standardized metric. These are standardized like z -scores. Thus, one rule of thumb is that standardized residuals larger than 2 signal an area of misfit in the model. The values of standardized residuals are dependent on sample size, however, so large samples will produce many more large values than will small samples. Thus, a better rule of thumb is to focus on the largest values in this matrix, especially when the more global fit statistics suggest a lack of fit.

Table 14.3 shows another useful, related matrix: the differences between the actual correlations and those implied by the model (the Sample Correlations from Figure 14.9 minus the Implied Correlations). As we will see, this matrix can also be useful for isolating and understanding problems with models. Kline (2016), for example, advises to examine further variables with such “correlation residuals” greater than .10 or less than –.10. This matrix is not produced in all SEM programs (it is not produced in Amos or Mplus, for example), but it is easily generated in Excel by subtracting the values in the implied correlation matrix from those in the actual correlation matrix. As shown in the table, the negative correlation between Minority and Grades that is implied by the Homework was slightly larger than the actual correlation; in contrast, the actual (positive) correlation between Family Background and Grades was slightly larger than the correlation implied by the Homework model.

Although these residual matrices are not particularly useful in the present example in which only two paths have been constrained, as we begin to focus on more complex and latent variable models the standardized residual covariances and the residual correlations will be useful for determining *where* there is misfit in our models. I focused on the actual, implied, and residual matrices now, however, because this difference between the actual and implied covariance matrix is the source of other measures of the fit of the model.

Table 14.3 Differences between the Actual Correlations and those Implied by the Homework Model

	<i>Minority</i>	<i>FamBack</i>	<i>PreAch</i>	<i>Homework</i>	<i>Grades</i>
<i>Minority</i>					
<i>FamBack</i>	0				
<i>PreAch</i>	0	0			
<i>Homework</i>	0	0	0		
<i>Grades</i>	.025	.022	0	0	

We can and will quantify the *degree* to which a model is overidentified. The current model has two paths that could have been drawn to make the model just-identified (paths from Minority and Family Background to Grades). The model thus has two degrees of freedom. More exactly, we can calculate the degrees of freedom using the following steps:

1. Calculate the number of variances and covariances in the matrix using the formula $[p \times (p + 1)]/2$ where p is equal to the number of variables in the model. For the current model, there are 15 variances and covariances: $[5 \times (5 + 1)]/2 = 15$.
2. Count the number of parameters that are estimated in the model. Don't forget covariances between exogenous variables, variances of the exogenous variables, and variances of the disturbances. For the current model, we estimated seven paths, one covariance between the exogenous variables, the variances of the two exogenous variables, and the variances of the three disturbances, for a total of 13 estimated parameters.
3. The degrees of freedom are calculated by subtracting the number of estimated parameters from the number of variances and covariances. The present model has two degrees of freedom ($15 - 13 = 2$).

The degrees of freedom for a model provide information about the degree to which the overall model is overidentified. The degrees of freedom also provide a handy index of the parsimony of the model. In science, we value parsimony: if two explanations for a phenomenon are equally good (or, in SEM, fit equally well), we generally prefer the simplest or more parsimonious explanation. Degrees of freedom are an index of the parsimony of a path model: the more degrees of freedom, the more values constrained (to zero or some other value) prior to estimation, and thus the greater the parsimony.

The difference between the actual and implied matrices provides evidence of the degree to which the model is a good explanation of the data. This difference is used to generate a multitude of fit statistics or fit indexes for overidentified models. There are literally dozens of such fit indexes, with different indexes focusing on slightly different aspects of fit. We will focus on a few common such indexes here; there are also numerous sources for more information about fit indexes (e.g., Hoyle, 1995; Hu & Bentler, 1998, 1999; Marsh, Hau, & Wen, 2007; see also David Kenny's web pages for excellent and up-to-date advice on fit statistics: <http://davidakenny.net/cm/fit.htm>).

Chi-square (χ^2) is the most commonly reported measure of fit.³ Chi-square has the advantage of allowing a statistical test of the fit of the model; it can be used with the degrees of freedom to determine the probability that the model is "correct" (to be explained later). Interestingly, in SEM we want a small χ^2 and one that is not statistically significant. For our current example, $\chi^2 = 2.166$, with 2 *df* and a probability of .338. What does this mean? It means that the actual and the implied covariance matrix are not statistically significantly different from one another, and thus the model and the data are consistent with one another. If the model and the data are consistent, the model *could have* generated the data and thus may provide a good approximation of how the phenomenon being studied works. In other words, the model may approximate reality, it may be "correct." Given all the "mays" and "coulds" in this explanation, you may be disappointed; this is hardly the kind of evidence of the quality of the model you were hoping for! Sorry; fit statistics do *not* prove that a model is true and do *not* prove causality. If the fit indexes are good, they suggest that a model may provide a reasonable, tentative explanation of the data. I'll simply note that this is better than nothing and more feedback than we've had in previous chapters about the quality of our explanations of our data.

Figure 14.11 shows the fit indexes output by Amos; other SEM programs will provide an equally intimidating listing of indexes of fit, many of which will be the same (although some may be labeled differently). Focus on the first few rows and columns. The model that is being estimated (i.e., the model in Figure 14.7) is labeled the “Default model”. The first column of numbers shows the number of parameters (NPAR) that are estimated in the model (remember we calculated 13 parameters being estimated), and the second shows the χ^2 (labeled CMIN, a value of 2.166). These are followed by the degrees of freedom (2) and the probability associated with the χ^2 and df (.338).

The rows labeled “Saturated model” pertain to a just-identified model. A just-identified model will estimate 15 parameters and thus have zero df . With a just-identified model, the implied covariance matrix will be identical to the actual matrix, and thus χ^2 associated with a just-identified model is equal to zero. In other words, a just-identified model will provide a perfect fit to the data. Why not, then, continue to estimate just-identified models, as we have done previously? The reason, again, is that we value parsimony. An overidentified model is more parsimonious than a just-identified model; our present overidentified model fits as well as a just-identified model (another interpretation of the statistically not significant χ^2). Because this model fits as well but is more parsimonious, it is a preferable model from a scientific standpoint.

The rows labeled “Independence model” refer to a model in which the variables in the model are assumed to be independent of one another. This model, also called a *null* model, could be represented by the five variables, with no paths or correlations drawn (and thus for this model all we would estimate would be the five variable variances). It could also be represented by constraining all paths and correlations in the current model to zero. Again, the null model assumes the variables are unrelated. The saturated and independence models essentially provide two endpoints with which we can compare our theoretical model. The saturated model provides a best fitting model and the independence model a very poorly fitting model. Some of the other fit indexes make use of these endpoints.

Other Measures of Fit

χ^2 seems to fill our need for assessing model fit: if it is not statistically significant, we have evidence that our model may explain reality, and if it is statistically significant, our model does not explain the data. Why do we need other fit indexes? Unfortunately, χ^2 has some problems as a measure of fit. First, χ^2 is related to sample size; indeed, χ^2 is calculated as $N - 1$ times the minimum value of the fit function (FMIN on the Amos output). Thus, given the same matrix and a sample size of 10,000 instead of 1000, the χ^2 would be approximately 10 times larger than the current value of 2.166. A χ^2 of 21.66 (actually 21.68, because $N - 1$ rather than N is used in the calculations), again with 2 df , will be statistically significant ($p < .001$), and thus we would reach the conclusion that the model did not fit the data, an opposite conclusion from the one we reached with the sample size of 1000. (A reminder: with SEM the df depend on the number of model constraints, not the sample size.) This weakness of χ^2 is one reason alternative measures of fit have been developed. Most SEM users report χ^2 but also report other fit statistics as well.

Several fit indexes compare the fit of the existing model with that of the null, or independence model. The comparative fit index (CFI) and the Tucker–Lewis index (TLI, also known as the nonnormed fit index, or NNFI) are two common such indexes. The CFI provides a population estimate of the improvement in fit over the null model (although null models are the most common comparison, the CFI can also be calculated with more restricted but substantively meaningful models). The TLI provides a slight adjustment for parsimony and

Model Fit Summary

CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	13	2.167	2	.338	1.083
Saturated model	15	.000	0		
Independence model	5	817.868	10	.000	81.787

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	.008	.999	.994	.133
Saturated model	.000	1.000		
Independence model	1.998	.715	.572	.477

Baseline Comparisons

Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	.997	.987	1.000	.999	1.000
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

Parsimony-Adjusted Measures

Model	PRATIO	PNFI	PCFI
Default model	.200	.199	.200
Saturated model	.000	.000	.000
Independence model	1.000	.000	.000

NCP

Model	NCP	LO 90	HI 90
Default model	.167	.000	8.213
Saturated model	.000	.000	.000
Independence model	807.868	717.735	905.396

FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	.002	.000	.000	.008
Saturated model	.000	.000	.000	.000
Independence model	.819	.809	.718	.906

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.009	.000	.064	.854
Independence model	.284	.268	.301	.000

AIC

Model	AIC	BCC	BIC	CAIC
Default model	28.167	28.324	91.967	104.967
Saturated model	30.000	30.181	103.616	118.616
Independence model	827.868	827.929	852.407	857.407

Figure 14.11 Fit indexes for the Homework model.

ECVI

Model	ECVI	LO 90	HI 90	MECVI
Default model	.028	.028	.036	.028
Saturated model	.030	.030	.030	.030
Independence model	.829	.738	.926	.829

HOELTER

Model	HOELTER .05	HOELTER .01
Default model	2763	4247
Independence model	23	29

Figure 14.11 (Continued)

is relatively independent of sample size (Tanaka, 1993). For both indexes, values approaching 1.0 suggest a better fit; common rules of thumb suggest that values over .95 represent a good fit of the model to the data, and values over .90 represent an adequate fit (cf. Hayduk, 1996, p. 219; Hu & Bentler, 1999).

Another problem with χ^2 and its associated probability is that p is the probability that a model fits perfectly in the population, even though most researchers argue that a model is designed only to approximate reality. The root mean square error of approximation (RMSEA) is designed to assess the *approximate* fit of a model and may thus provide a more reasonable standard for evaluating models. RMSEAs below .05 suggest a “close fit of the model in relation to the degrees of freedom” (Browne & Cudeck, 1993, p. 144), in other words a good approximation. Browne and Cudeck further speculated that models with RMSEAs below .08 represented a reasonable fit, with those above .10 representing a poor fit. Research with the RMSEA supports these rules of thumb (i.e., values below .05 suggesting a good fit; Hu & Bentler, 1999), as well as its use as an overall measure of model fit (Fan, Thompson, & Wang, 1999). Other advantages of RMSEA include the ability to calculate confidence intervals around RMSEA, the ability to use RMSEA “to test a null hypothesis of *poor* fit” (Loehlin & Beaujean, 2017, p. 71), and the ability to conduct power calculations using RMSEA (MacCallum, Browne, & Sugawara, 1996). Conceptually, you can think of RMSEA as representing the degree of misfit per degree of freedom.

One final, useful measure of fit (or misfit) is the standardized root mean square residual (SRMR). We approached the topic of fit by discussing the difference between the actual covariance matrix used to estimate a model and the covariance matrix implied by the model. If you average these differences, you get the root mean square residual. (To keep the negative values from canceling out the positive values, you’d need to first square the values and then take the square root of the final average number.) The SRMR is the standardized version of the root mean square residual. Because correlations are standardized versions of covariances, the SRMR is conceptually equivalent to the average difference between the actual correlations among measured variables and those predicted by the model. Hu and Bentler’s (1998, 1999) simulation research suggests SRMR as among the best of the fit indexes, with values below about .08 suggesting a good fit of the model to the data (this value may be a little high; .06 may be a more reasonable value for SRMR). The SRMR is not produced automatically in Amos but is easily obtained (select the “Plugins” menu, then Standardized RMR).

I currently use RMSEA for my primary measures of the fit of a single model, supplemented by SRMR and CFI or TLI, or sometimes other indexes. As we will soon see, it is also possible, indeed desirable, to compare the fit of competing models; we will use different fit

indexes for this purpose. Note, however, that thinking and research about fit indexes are in a constant state of development. The advice I (or others) give as this is written is different from what I would have given 10 years ago and may well be different from what I will advise 10 years in the future. Because of this state of flux, and because much advice about fit indexes is based on the experience of the user, my advice may also be different from that of others. See the section at the end of this chapter for additional thoughts concerning fit indexes.

Focus again on Figure 14.11. The RMSEA for our Homework model was .009, with a 90% confidence interval of .000 to .064 (Lo 90 to Hi 90 in the figure). The CFI and TLI were 1.0 and .999, respectively. Although not shown in the figure, the SRMR for this model was .0085, suggesting an average difference between the actual and predicted correlations of only .0085. All indexes suggest a good fit of the model to the data; the model could indeed have generated the data.

Comparing Competing Models

Another major advantage of SEM programs is that we can use them to compare competing theoretical models (and the hypotheses embedded in these competing models) via the fit statistics. An example will illustrate.

Suppose in your reading of the literature on the effects of homework you came across evidence that homework and school learning are unrelated. Perhaps the evidence is in the form of research that suggests that homework has no real effects on achievement or grades. Or perhaps the evidence is in the form of informal theory that suggests that homework really should not affect learning, or vice versa. Whichever is the case, we could test these hypotheses by comparing models embodying them with our initial model (Figure 14.8). One such model will delete the paths from Previous Achievement to Homework and from Homework to Grades. This model asserts that students' previous achievement has no effect on the time they spend working on homework, and such time spent on homework also has no effect on students' grades. Stated differently, this model embodies the hypothesis that homework is unrelated to academic performance, either as an effect (the path from Previous Achievement to Homework) or as an influence (the path from Homework to Grades).

The standardized results of this model are shown in Figure 14.12, which also shows some of the relevant fit indexes. We will focus primarily on the RMSEA (.128), which suggests a poor fit of this model to the data. This assessment is supported by the TLI of .797 and the statistically significant χ^2 of 69.61 with 4 degrees of freedom. The CFI (.919) and the SRMR (.071), in contrast, suggest a so-so model. The model is, among other things, a good illustration that the various fit statistics often present different pictures and lead to different conclusions if used in isolation. Nevertheless, given the mixed fit, and with a primary focus on RMSEA, we conclude that this model does not fit the data well, and we will likely reject the model as a good explanation of the relations between homework and learning.

We can address our primary questions more directly, however, by comparing the results of this model with the results of our initial model. That model fit well, whereas this model did not; but are the *differences* between the fits of the two models meaningful or statistically significant? We can use the fit indexes to make these comparisons, as well. Interestingly, although χ^2 has problems as a measure of fit of a single model (what I will henceforth call a "stand-alone" measure of fit), it often works well for comparing competing models. Furthermore, if the models are nested (meaning that one can be derived from the other by deleting paths), this comparison can be statistical rather than qualitative.

When two models are nested, the more parsimonious model (the model with fewer free, or estimated, parameters) will have a higher df (recall that df is a measure of parsimony) and a larger χ^2 . The χ^2 and df for the less parsimonious model can be subtracted from those of

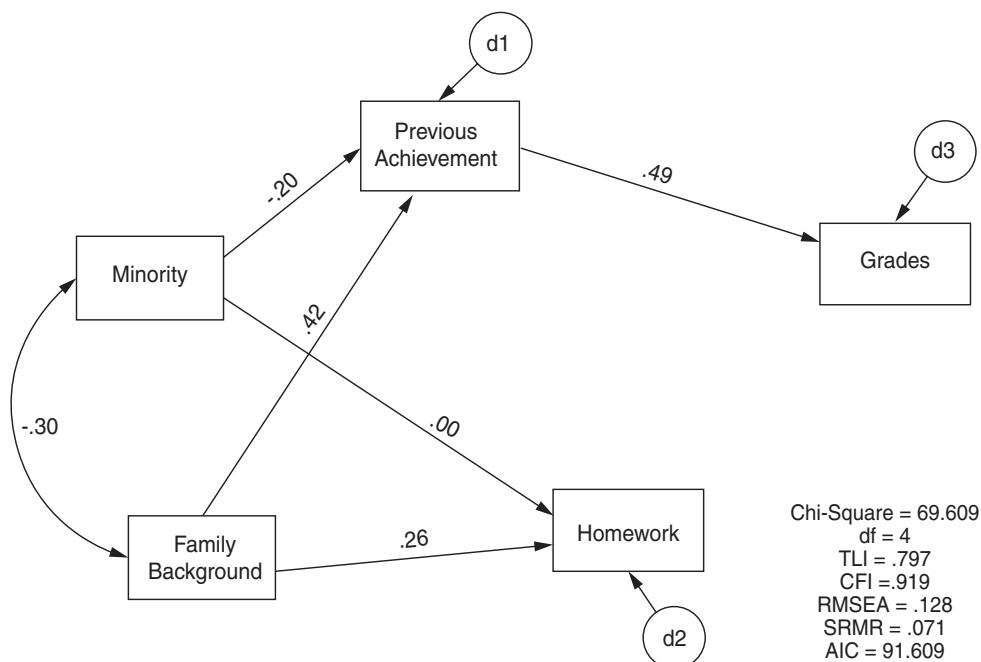


Figure 14.12 Does Homework indeed affect Grades? Compare the fit of this model with the earlier Homework model.

the more parsimonious. The resulting change in χ^2 ($\Delta\chi^2$) is also a χ^2 distribution and may be compared to the change in df for the two models. Again, models are nested when one can be derived from the other by deleting paths or correlations. This second model—the one with one or more paths deleted and the higher df —will be a subset of the first and nested within the first model.

The model shown in Figure 14.12 (no-homework-effects model) is a more parsimonious, more constrained version of the model shown in Figure 14.7 (the initial model); two paths that were estimated in the initial model were constrained to zero in the no-homework-effects model. This model is nested within the initial model. The no-homework-effects model had a χ^2 of 69.609, with 4 df . We subtract the corresponding values for the initial model ($\chi^2 = 2.166, df = 2$) from those for the no-homework-effects model to obtain a $\Delta\chi^2$ of 67.443, with a Δdf of 2. If you look up these values in a probability calculator or spreadsheet,⁴ you will find an associated probability of $< .001$; the additional constraints on the no-homework-effects model resulted in a statistically significant increase in $\Delta\chi^2$.

This finding, that the $\Delta\chi^2$ is statistically significant, means that not only does the no-homework-effects model fit worse than the initial model, but it fits statistically significantly worse. Although the no-homework-effects model is more parsimonious than the initial model, the parsimony comes at too great a cost in terms of model fit, and we reject these constraints on the model and stick with the initial model. What this means, in turn, is that we can reject the hypothesis that time spent on homework is unrelated to academic performance.

The process of comparing competing models can be used to test competing models and hypotheses, but it can also bolster, or undermine, our faith in our preferred models. “The fact that one model fits the data reasonably well does not mean that there could not be alternative models that fit better. At best, a given model represents one tentative explanation of the

Table 14.4 Comparison of Fit Indexes for Alternative Models of the Effects of Homework on High School Students' Grades

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	RMSEA (90% CI)	SRMR	CFI	AIC
Initial	2.166	2				.009 (.000–.064)	.009	1.00	28.166
No Homework Effects	69.609	4	67.443	2	<.001	.128 (.103–.155)	.071	.919	91.609
Background Effect	1.329	1	.837	1	.360	.018 (.000–.089)	.008	1.00	29.329

data. The confidence in accepting such an explanation depends, in part, on whether other, rival explanations have been tested and found wanting" (Loehlin & Beaujean, 2017, p. 63).

We can also use $\Delta\chi^2$ to test the assumptions we made when we developed our initial model. Recall that we assumed that Minority and Family Background had no direct effect on students' Grades, but that their effects were indirect through Previous Achievement and Homework. We could test whether these assumptions are, in fact, supported by freeing these parameters and studying the change in fit of the model. Table 14.4 shows fit statistics for the two models already discussed, plus a model labeled Background Effect, in which the path from Family Background to Grades was freed, or estimated. As you can see, this model is less parsimonious than the initial model. The $\Delta\chi^2$ for this model was .837 with 1 df ; the $\Delta\chi^2$ is not statistically significant. In this case, the two models had nearly equivalent fit. The more relaxed (background effect) model did not fit statistically significantly better; the more parsimonious (initial) model did not fit statistically significantly worse. In other words, the models had equivalent fit. In this case, we favor the more parsimonious of the two models, the initial model. Therefore, our initial assumption that Family Background would affect Grades only through other variables was supported. (Earlier in the text I suggested that you memorize the factoid that with a reasonable sample size, a t of approximately 2 is statistically significant. It would also be worth remembering that with 1 df , a $\Delta\chi^2$ of around 3.9 is statistically significant.)

Note that we could also have evaluated the statistical significance of the path from Family Background to Grades by focusing on the CR (critical ratio, or z) in the Amos printout. The z was .915, which is not statistically significant, thus also supporting our initial assumption of the lack of direct effect of Family Background on Grades. When single parameters are tested, $\Delta\chi^2$ and z will usually, but not always, give the same answer. $\Delta\chi^2$ can be used to test the statistical significance of multiple changes to a model, whereas t focuses on only one parameter at a time.

We could have freed both paths that were constrained to zero in the initial model (Family Background to Grades and Minority to Grades). In this case, the new model will be just-identified, with χ^2 and the df both equal to zero. Thus, the $\Delta\chi^2$ comparing this model with the initial model equals the value for the χ^2 for the initial model (2.166, $df = 2$), which was not statistically significant ($p = .338$). Perhaps this comparison makes it obvious that, strictly speaking, what we are testing with overidentified models is the overidentifying restrictions (constraints) on the model, *not* the model as a whole.

We can also use fit statistics to clean up our models. Note that the path from Minority to Homework was not statistically significant in the initial homework model. One alternative

model worth investigating is one in which this path is deleted. With this change, $\Delta\chi^2$ is statistically not significant; this more parsimonious model thus fits as well as does the initial model. Although it is perfectly reasonable to use $\Delta\chi^2$ and other fit statistics to clean up models, keep in mind that this process is fundamentally different from the other model comparisons we have made. Our previous model comparisons were designed to test hypotheses drawn from theory and previous research. Model modifications to remove statistically nonsignificant paths are not theoretical; instead, they are based on the data themselves. They should not be accorded the same weight as theoretically derived model modifications until they are tested against new data. If you do a lot of data-based model revisions, you should recognize that you are conducting exploratory, rather than theory testing, research.

To reiterate, our rule of thumb is that if $\Delta\chi^2$ is statistically significant it means that the more parsimonious model has a statistically significantly worse fit than does the less parsimonious model. If you use this methodology, you would then reject the more parsimonious model in favor of the less parsimonious one. If, on the other hand, the $\Delta\chi^2$ is not statistically significant, then this means that the two models fit equally well (within a reasonable margin of error). Because we value parsimony, in this case you would reject the less parsimonious model in favor of the more parsimonious one.

Table 14.4 also includes one more fit index that can be used to compare competing models. The Akaike Information Criterion (AIC) is a useful cross-validation index in that it tends to select models that would be selected if results were cross-validated to a new sample (Loehlin & Beaujean, 2017). Another useful feature of AIC is that it can be used to compare competing models that are not nested. Smaller values of AIC are better, and thus if we use the AIC to compare the models in Table 14.4, we will continue to favor the initial model over its competitors. Another, related measure is the Bayes Information Criterion (BIC in the Amos output); the BIC includes a stronger adjustment for parsimony than does the AIC. Another, related index that I often use is the sample size adjusted BIC, the aBIC. Its parsimony adjustment is between that of the AIC and the BIC. The aBIC is not currently computed in Amos but is relatively easy to calculate using other fit information provided. See, for example, David Kenny's web site (<http://davidkenny.net/cm/fit.htm>). The Amos manual shows how to calculate the fit indices used in that program. aBIC is produced in Mplus.

The table shows the values for the RMSEA, along with its 90% confidence interval. These values can also be used to compare competing models either informally, by choosing the model with the lowest RMSEA, or more formally, by comparing the point value for one model with the 90% CI for another model. Using either approach, we will favor the initial model as being better fitting than the no-homework-effects model and more parsimonious (but equivalent fitting) compared to the background effect model. Some researchers also use the CFI to compare competing models in tests of invariance (e.g., Cheung & Rensvold, 2002; see chapter 20).

I currently use $\Delta\chi^2$ as my primary index for comparing competing models if these models are nested and given a reasonable sample size (say 150 to 1000). For nonnested models, the AIC and aBIC have worked well in my experience.

MORE COMPLEX MODELS

Equivalent and Nonequivalent Models

Equivalent Models

We saw in Chapter 13 that with just-identified models path directions could be reversed, leading to very different conclusions, without any warning that one model was correct and the other incorrect. In other words, these models with reversed paths (e.g., Figures 13.3

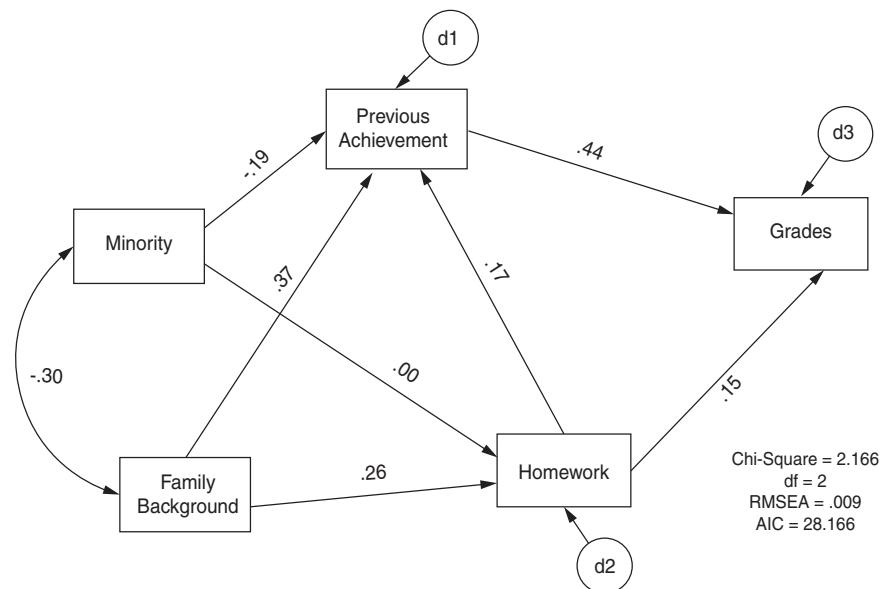


Figure 14.13 An equivalent model. Note that the Previous Achievement to Homework path is reversed, but the fit indexes are identical to those of the initial Homework model.

and 13.7) are equivalent; we cannot differentiate them by their fit. This makes sense, because just-identified models fit perfectly, and thus we cannot differentiate them by their fit.

We have seen in this chapter that one advantage of overidentified models analyzed through SEM programs is that they provide measures of fit of the model to the data. We can use these fit indexes to compare models, to reject those that fit less well and tentatively accept those with better fit. What may not be obvious is that it is also possible, in fact likely, to have equivalent overidentified models. Equivalent models are those that produce the same fit statistics as the original model and thus cannot be differentiated from that model based on fit. It is often possible to reverse a path or to replace a path with a correlation without any change in the fit of the model. For example, Figure 14.13 shows the results of an analysis in which the path from Homework to Previous Achievement was reversed (compared to Figure 14.8). As can be seen in Figure 14.13, the χ^2 , df , and RMSEA are all the same as in the initial analyses of this model (Figure 14.11; and although not shown, the rest of the fit indexes are also identical). The two models, with the path between Homework and Previous Achievement drawn in opposite directions, are statistically equivalent and cannot be differentiated. There are, in fact, numerous equivalent models to most target models, and you should consider them as you focus on a particular model.

Stelzl (1986) and Lee and Hershberger (1990) provided rules for generating equivalent models. The main gist of these rules is summarized briefly here. For this presentation, I have assumed that the beginning models are recursive.

1. For a just-identified model, a path from a to b (symbolized as \rightarrow) may be replaced by a path from b to a (\rightarrow) or by a correlation between a and b (if a and b are exogenous). Endogenous variables may not have simple correlations, but their disturbances may be correlated.⁵ Thus, a path from endogenous variable a to endogenous variable b may be replaced by a correlation between the disturbances of a and b (I will symbolize both types of correlations by $\curvearrowleft \curvearrowright$ for this discussion). All these possibilities are equivalent,

- meaning you can also replace \leftrightarrow with \leftarrow . This is simply another way of stating that all just-identified models are statistically equivalent because they all fit the data perfectly.
2. More importantly, for overidentified models, portions of these models may be just identified. For the just-identified portions of models, these same rules apply. That is, you can replace \rightarrow by \leftarrow or by \leftrightarrow (or vice versa), and the model will be equivalent. So, for example, in Figure 14.13 the model is just-identified through the variable Homework. This is why we can reverse the Homework–Previous Achievement path and still have an equivalent model.
 3. For portions of the model that are overidentified, if a and b have the same causes, \rightarrow (a path from a to b) may be replaced by \leftrightarrow or by \leftarrow . Thus, for the model in Figure 14.7, reversing the path from Homework to Grades will not result in an equivalent model, because the two variables do *not* have the same causes.
 4. For portions of the model that are overidentified, when a and b do *not* have the same causes, the substitutions are slightly more complex. A path from a to b may be replaced by \leftrightarrow if the causes of b *include* all the causes of a . You could not replace the path from Homework to Grades with a correlated disturbance between d_2 and d_3 because the causes of Grades do not include all the causes of Homework. Minority and Family Background are influences on Homework but not Grades. In addition, correlated disturbances can be replaced by a path from a to b if b includes all causes of a .

Figure 14.14 shows several equivalent models to our original Homework model (from Figure 14.7, also shown as model A in Figure 14.14). Make sure you understand why each is equivalent to the original model. It is worth noting that these rules can be applied repeatedly, which is how the final model (model F) is derived. The derivation of each model is explained in note 6.⁶

It should be obvious from a study of Figure 14.14 that the presence of equivalent models may threaten the causal conclusions from our research. If all these models are statistically

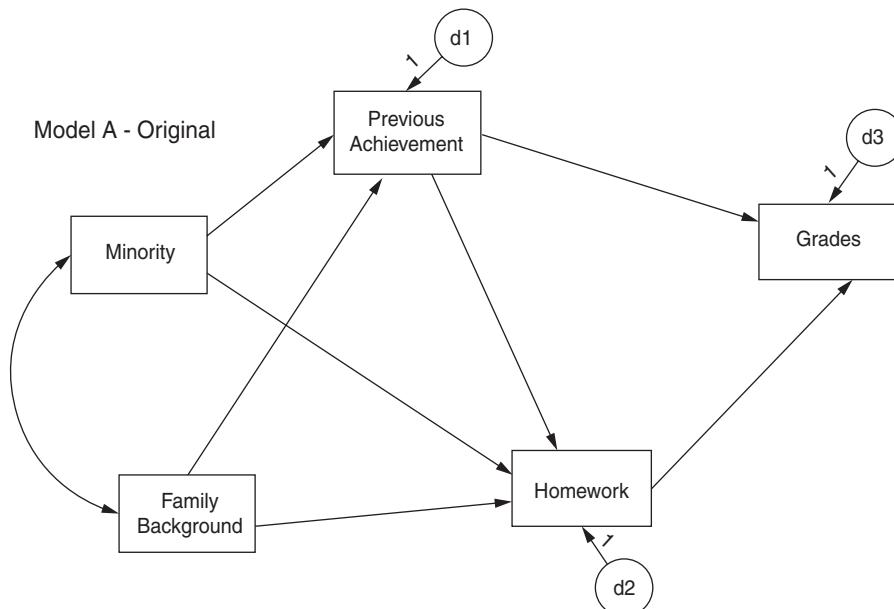


Figure 14.14 Equivalent models. All the models shown below are equivalent to Model A and cannot be differentiated from it based on fit.

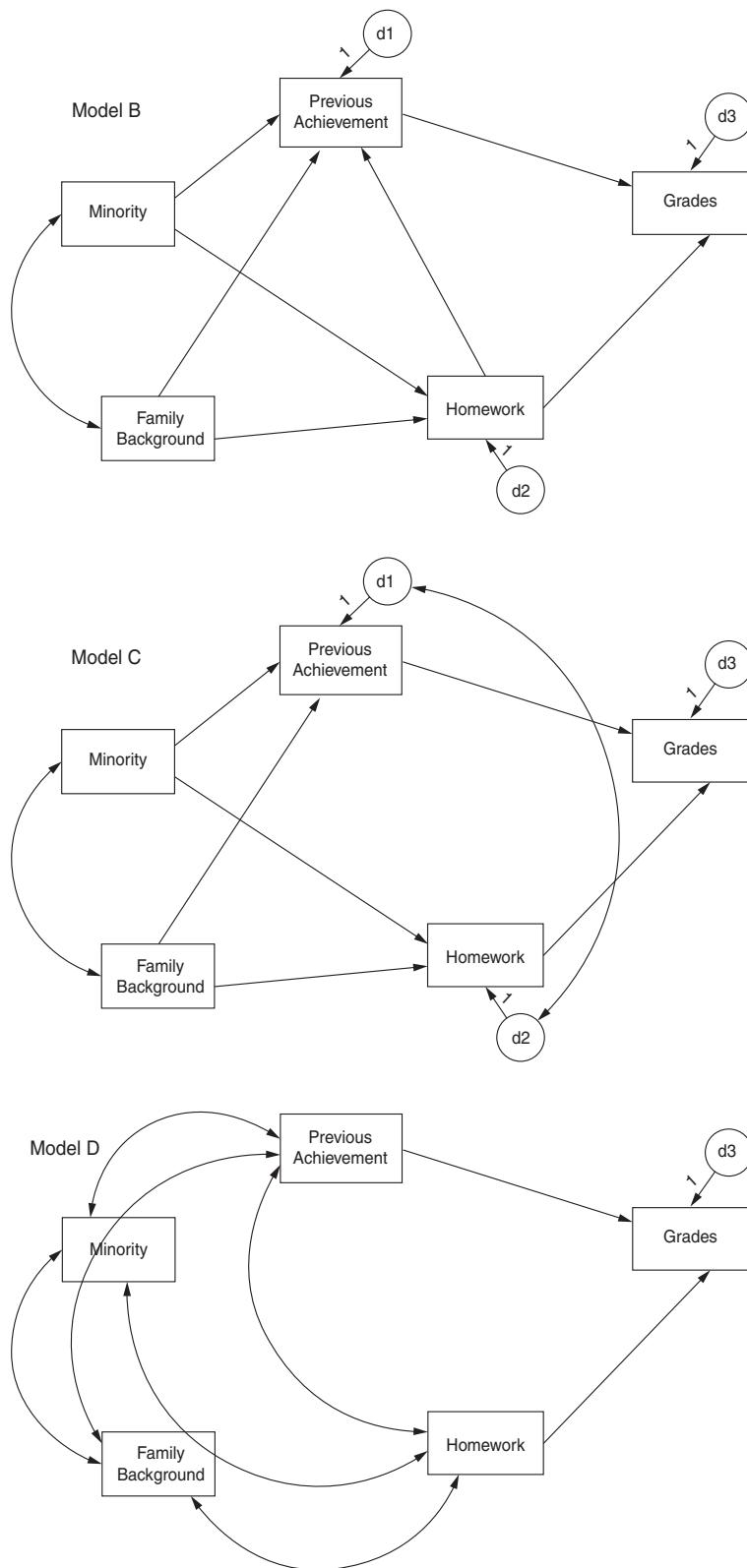
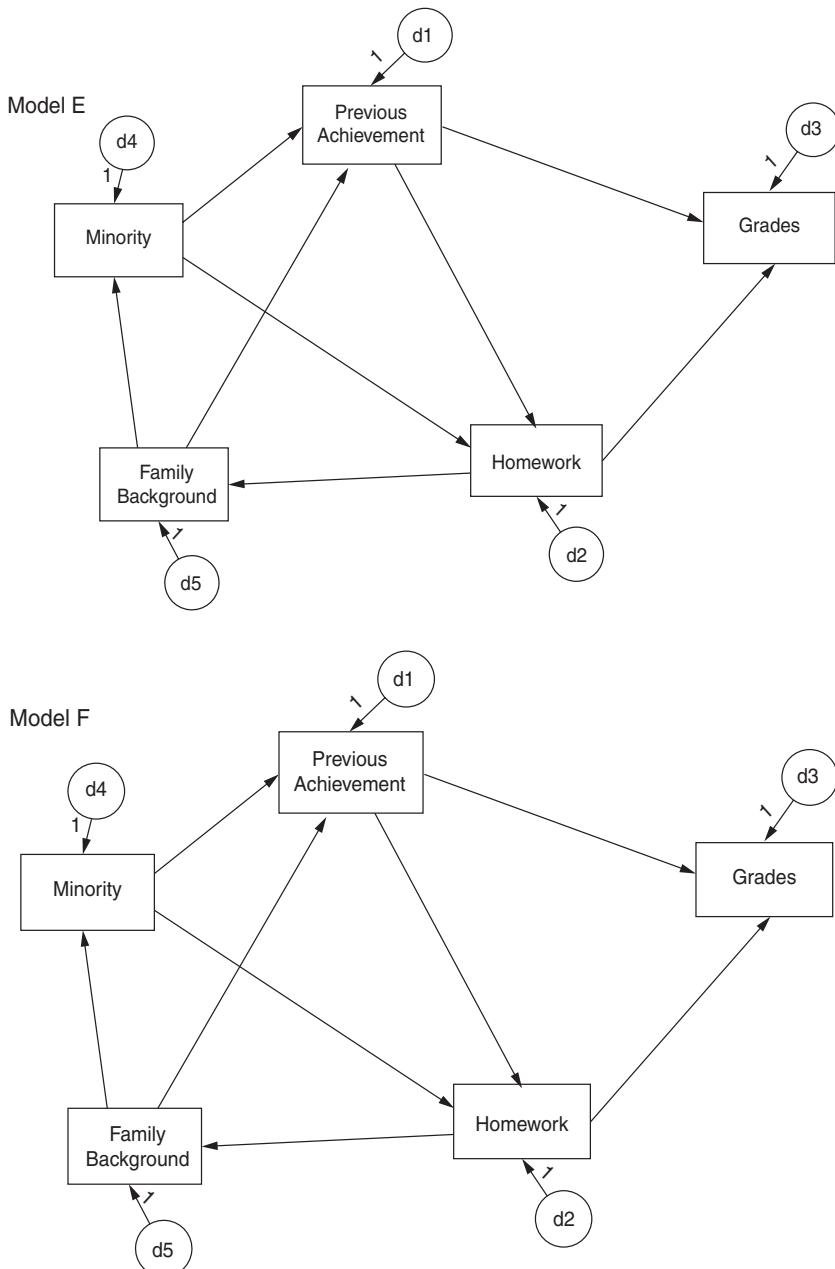


Figure 14.14 (Continued)

**Figure 14.14 (Continued)**

equivalent to our preferred model, how can we assert, for example, that Previous Achievement affects Homework, rather than Homework affecting Previous Achievement? I encourage you to generate and consider alternatives to your model of choice. You may discover alternatives that make as much sense as your original model, or you may begin to feel more comfortable with your initial model. It is certainly better to consider equivalent models and either revise your models or defend your reasoning prior to publication rather than after! But, in reality, the answer to the threat of equivalent models is the same as the method of devising strong models to begin with: consider logical and actual time precedence, build in relevant theory and research, and carefully consider the variables involved.

What should we do, however, when equivalent models remain plausible even after such considerations? As we will see, one possible solution is to devise *nonequivalent* models that evaluate the different possibilities; another possibility is the use of longitudinal data.

The Lee and Hershberger rules apply to portions of nonrecursive models as well, but the rules presented here will cover most models of interest in this text. See Lee and Hershberger (1990) for more information; the rules are also summarized and well illustrated by Hershberger (2006) and by Kline (2016). MacCallum, Wegener, Uchino, and Fabrigar (1993) illustrated problems that arise from not considering equivalent models.

Directionality Revisited

If some overidentified models are equivalent, it follows that some overidentified models are not equivalent and that we can use the same rules to generate nonequivalent models. These, in turn, may help us deal with one problem we encountered with simple just-identified models: uncertainty concerning causal direction. As you will see, although the problem of equivalent models is a danger to SEM interpretation, an understanding of the rules of equivalent models can lead to the development and testing of nonequivalent models, which can be a blessing.

Figure 14.15 shows one more version of the homework model, one in which the path from Homework to Grades is reversed, replaced by a path from Grades to Homework. This direction does not make sense based on time precedence (Homework includes information from 8th and 10th grades, whereas Grades are from 10th grade). Still, as demonstrated in Chapter 13, if we estimate a just-identified version of this model, there will be nothing in our analysis to tell us that it is incorrect. The current version is overidentified. More importantly, this model is not equivalent to the original model. Grades and Homework do not have the same causes (rule 3), and thus the reversal of the path does not result in an equivalent model. If the models are not equivalent, does that mean that the fit indexes may help spot the error in our model? In a word, yes.

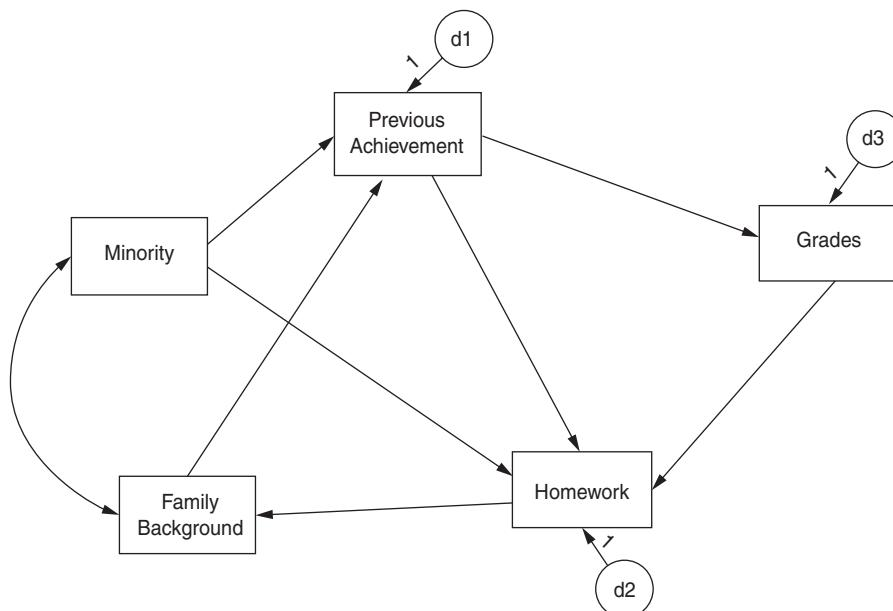


Figure 14.15 Reversing the Homework to Grades path results in a nonequivalent homework model.

Figure 14.16 shows the solved “wrong direction” model with a few of the relevant fit indexes. Note that if we look at the RMSEA (or other stand-alone fit indexes), this model will be deemed acceptable. Of more interest, however, is to compare this model with the initial “correct” homework model. We can’t use $\Delta\chi^2$ because the two models are not nested; you cannot arrive at one by deleting paths from the other. Indeed, the models are equally parsimonious (they have the same df). We can still use the AIC to compare nonnested models, however. As you can see, if you compare the AIC from Figure 14.16 with the fit indexes for the original model (shown in Figure 14.11), the AIC for the original model is smaller. The rule of thumb for AIC is to favor the model with the lower value; we would thus favor the original model over the model with the Homework–Grades path drawn in the wrong direction. The judicious use of nonequivalent models may indeed help us answer nagging questions of directionality!

You may wonder why this should work. Recall the genesis of the fit indexes: a comparison of the actual correlation–covariance matrix with the matrix implied by the model. Quite simply, Figure 14.16 implies a slightly different covariance matrix than does the model shown in Figure 14.8, and the matrix implied by the model shown in Figure 14.8 comes closer to the actual matrix.

Practically, the easiest way to develop such nonequivalent models is to include variables that uniquely cause the variables in question. That is, include variables in the model that are influences of the presumed cause but not the presumed effect and variables that are influences on the presumed effect but not the presumed cause. In other words, include some relevant *noncommon* causes in the model. Thus, although we saw in Chapter 13 that non-common causes are not *required* for the model to be valid, we now see they may help in dealing with other problems. Likewise, intervening (mediating) variables can help in the development of nonequivalent models and thus may be valuable for this purpose as well.

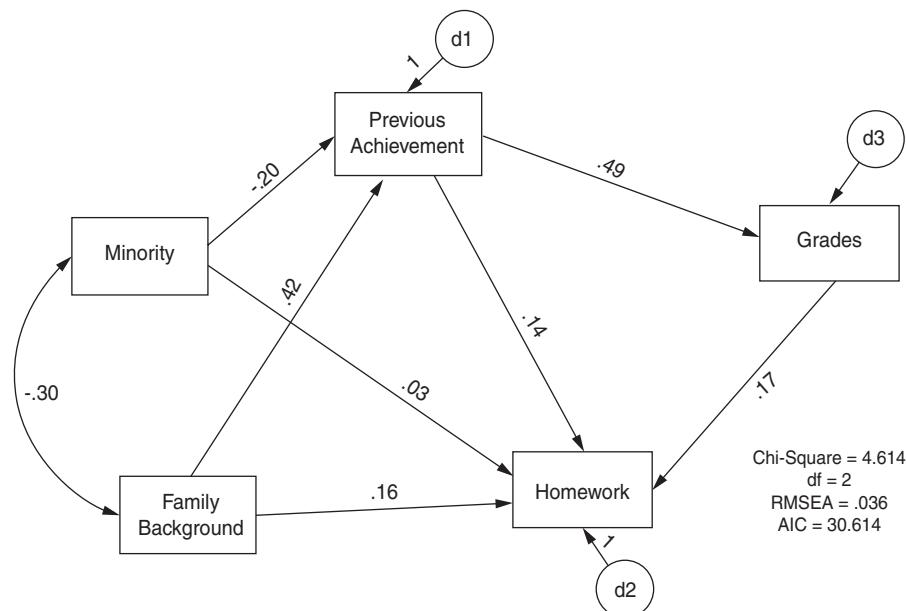


Figure 14.16 The nonequivalent homework model demonstrated a worse fit to the data.

Nonrecursive Models

Another advantage of SEM programs is that they can be used to analyze nonrecursive models, or models with feedback loops. Suppose you were interested in the influences on partners' levels of trust in marriage and other close male-female relationships. It makes sense that my level of trust in my wife may be affected, in part, by my own personal and psychological characteristics. My trust may also be affected by my wife's level of trust in me, however, and vice versa. If I trust my wife more, she will likely trust me more, and so on. Trust likely has reciprocal effects. Your theoretical model might look something like that shown in Figure 14.17. The model posits that one's trust in his or her partner is affected by one's own characteristics (self-esteem and perception of the partner's desire for control), as well as by the partner's own level of trust. This model is a smaller version of one posited and tested by John Butler (2001).

Recall that the tracing rule does not work with nonrecursive models but that we can develop formulas for the paths using the first law of path analysis. If you develop equations for the model shown in Figure 14.17, you find that, unlike recursive models, the formulas no longer are equivalent to those for regression coefficients from multiple regression. This is simply a convoluted way of saying that with nonrecursive models you cannot use ordinary multiple regression to estimate the paths.

It is possible, however, to use SEM programs to estimate models such as those shown in Figure 14.17. Some results of this analysis are shown in Figure 14.18; they suggest that each partner's trust is indeed affected by the other's trust. Self-Esteem had a positive effect on Trust, and Perception of Control had a negative effect, although the relative magnitudes of these effects were different for men and women. You will have a chance to return to this model in the exercises. (The data that produced these results are simulated because the original article did not include the correlation or covariance matrix. These simulated findings are consistent with those of the original article, however.)

I have presented this model as an example of the use of nonrecursive models to answer questions in which we expect there to be reciprocal effects. These are common in analyses of data from couples or other pairs of people. One of the best-known nonrecursive models, extensively analyzed and used as an example in many SEM manuals, was devised by

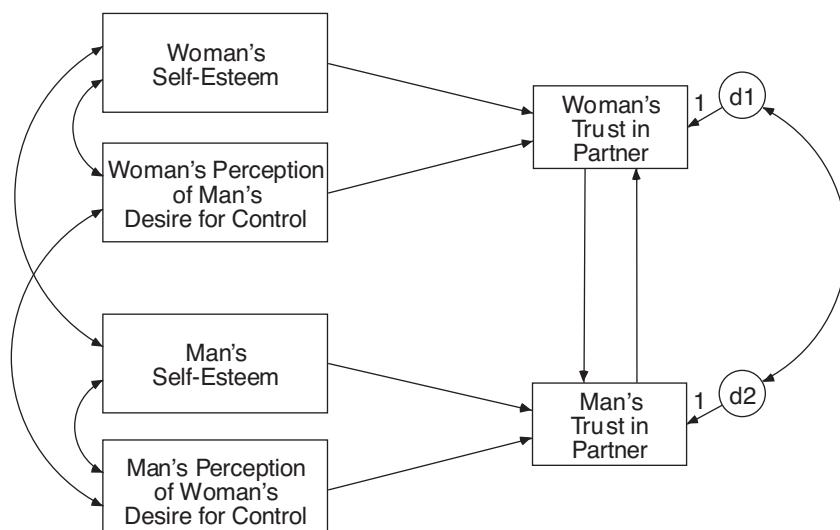


Figure 14.17 Nonrecursive model to test the reciprocal effects of partners' trust in each other.

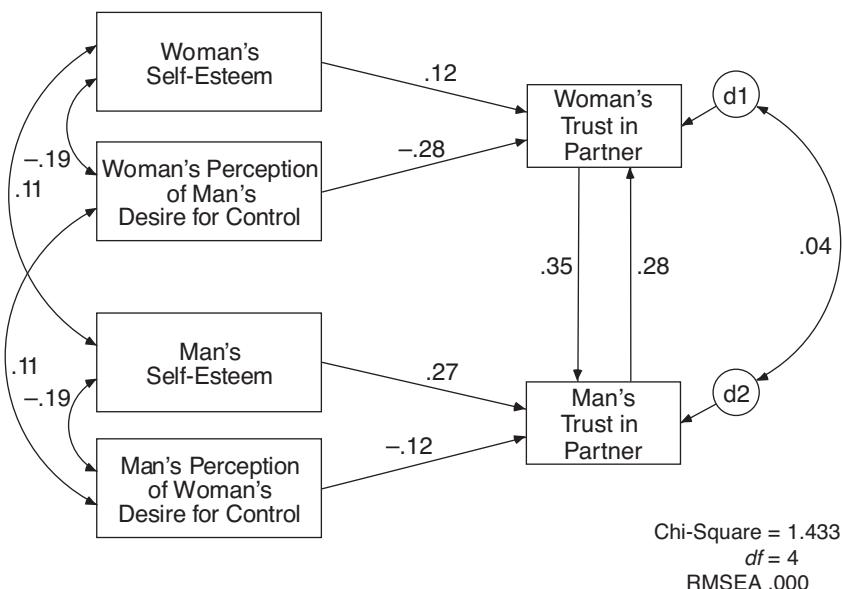


Figure 14.18 Standardized solution, partner trust model. The data are simulated, but based on research reported by Butler (2001).

sociologist Otis Dudley Duncan and colleagues to estimate the effects of friends on each other's occupational and educational aspirations (Duncan, Haller, & Portes, 1971). As you might expect, nonrecursive models are also used to settle questions of causal sequence (e.g., Reibstein, Lovelock, & Dobson, 1980).

Nonrecursive models are considerably more complex than this simple overview, however, and are beyond the scope of this book. If you are interested in pursuing nonrecursive models, you will need to study such models in considerably more depth. Kline (2016) and Loehlin and Beaujean (2017) provide a more detailed introduction, Rigdon (1995) presents a detailed discussion of identification issues for nonrecursive models, and Hayduk (1996) presents interesting issues related to nonrecursive models.

Longitudinal Models

Another method of answering questions about the reciprocal effects of variables on one another is through longitudinal models. Indeed, if you focus on our homework models, you will see that they take advantage of this technique. These models focus on the effects of homework on learning in later grades (subsequent GPA), while controlling for achievement in an earlier grade (Previous Achievement in 8th grade).

Do job stress and emotional exhaustion (or burnout) have reciprocal effects? Figure 14.19 shows a longitudinal model designed to answer this question for physicians surveyed in the United Kingdom (McManus, Winder, & Gordon, 2002). The physicians were surveyed in 1997 and again in 2000; the variables in the model should be self-explanatory. The data (stress burnout longitudinal.xls) and this model are on the Web site (stress burnout longitudinal 5.amw).

The model is barely overidentified (with 1 *df*); there is no path from Personal Accomplishment at time 1 to Stress at time 2. The results suggest that Stress and Emotional Exhaustion indeed have reciprocal effects. Stress increases Exhaustion, which, in turn, increases

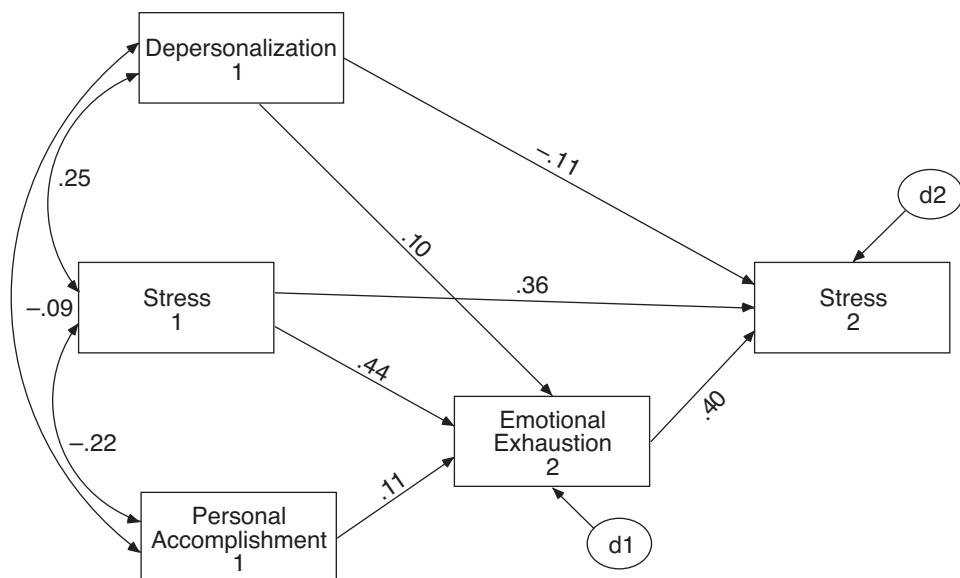


Figure 14.19 Reciprocal effects of Stress and Emotional Exhaustion, estimated via longitudinal data. The model is based on research with physicians (McManus et al., 2002).

subsequent Stress. It is worthwhile to compare this model to one in which it is assumed that Stress affects future Stress only via the indirect effect through Exhaustion (full mediation).

Longitudinal models can also help bolster the reasoning behind the paths we draw, even in the presence of equivalent models. If Emotional Exhaustion is measured in 1997 and Stress in 2000, it is easier to argue that the proper direction is from Exhaustion to Stress than if they are measured concurrently. Still, I don't want to oversell the ability of nonrecursive and longitudinal models to answer questions about the direction of influence; the results are not always as clear as we would like them to be. I have provided fairly clean and clear-cut examples here to illustrate the possibilities.

Figure 14.20 shows a special type of longitudinal model known as a panel model. A panel model has the same set of two or more variables measured two or more times. It is often used to test questions of reciprocal causation or settle issues of causal predominance. The model shown could be tested with the NELS full data (including the 12th-grade data, not included in our NELS subsample). Note that the Achievement tests (the same or similar tests) and the locus of control measure are administered three times; the model, as shown, has 15 df. If the results of the analysis showed a substantial effect from Achievement at every time period to Locus of Control at the next but not the reverse (Locus not affecting Achievement) then we would feel more comfortable in specifying a path from Achievement to Locus of Control in subsequent cross-sectional or longitudinal research. Note the correlated disturbances for the two variables of interest in 8th grade. In this case, the correlated disturbance may serve two purposes: it can take into account that we have not specified any effects between Achievement and Locus of Control at Grade 8 (perhaps additional correlated disturbances are needed at the other time points?), and that there may be other common causes of these variables that we have not considered. The time lag between measures in panel and other longitudinal model eases our concerns about specifying a causal direction, but keep in mind that the lag needs to be long enough for the causal process to have worked. For more information about longitudinal models, in general, and panel models, in particular, see Little, 2013. A latent variable version of this model is analyzed in Chapter 18.

Locus of control and achievement
Model Specification

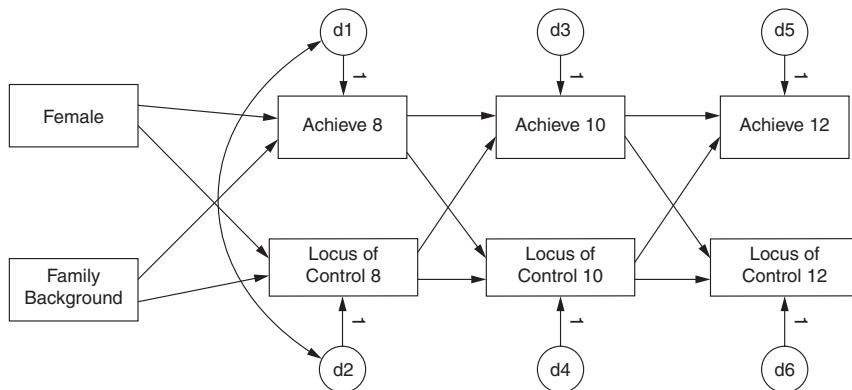


Figure 14.20 Potential longitudinal panel model designed to determine the extent of the effect of self-concept on achievement, and vice versa.

ADVICE: MR VERSUS SEM PROGRAMS

We have seen that with just-identified models SEM programs provide the same information for a path analysis as we get with multiple regression programs. With overidentified models, however, there are advantages in using SEM programs. If you have a choice, which should you use? Here's my advice:

1. If you plan to analyze a single, just-identified recursive model, either MR or a dedicated SEM program will work just fine.
2. If you plan to analyze an overidentified model or compare several competing models, use a SEM program. If you plan to analyze a nonrecursive model, use a SEM program.
3. If you are using a MR program to conduct a path analysis, there is no real benefit in specifying overidentified models. Instead, what I suggest is a more qualitative evaluation of fit. By this I mean that prior to analysis you should try to predict, based on previous research and theory, which paths will be close to zero, which should be large, which should be positive, which should be negative, and so on. I'm not suggesting that you necessarily need to make these as formal predictions, but you should spend some time thinking about what you expect each path to look like. After conducting the analysis, see how your predictions fared. If the paths you thought should be close to zero were, in fact, close to zero, and so on, you can have much more faith that your model may be a reasonable approximation of the way the phenomenon you are studying actually works. If, on the other hand, many of your predictions were wrong, you should be more cautious in your interpretation and should rethink your model and double-check your analyses.
4. If you are using a SEM program to conduct a path analysis, it is worthwhile to try to specify overidentified models rather than just-identified models. Again, spend some time comparing your model to what you know based on theory and previous research. Are there paths that you can set to zero based on such information? If so, delete them from your model (you can always test these no-effect hypotheses in subsequent

models). Again, it is preferable to specify these no-effect hypotheses prior to analyzing the data, rather than after running a just-identified model and noting which paths were statistically nonsignificant. If you are using a SEM program, you should also consider the substantive hypotheses you can test by comparing competing models.

ADVICE: MEASURES OF FIT

If you are using a SEM program to conduct path analysis (or CFA or latent variable SEM) you should strive for overidentified models and use the fit information to evaluate the models and to compare competing models. It is with some trepidation that I write this section attempting to consolidate and expand my earlier advice on fit indices. Quite simply, given the number of model characteristics (e.g., sample size, number of variables, degrees of freedom, model misspecification, and unique variances) my advice will often be wrong. But if you are a beginner, you need to start somewhere. Please also see the caveats at the end of this section.

Evaluating a Single Model

As noted earlier in the chapter, I have found RMSEA, SRMR, and CFI or TLI useful for evaluating the fit of a single model, or what I've called useful "stand-alone" fit indexes. For CFI and TLI I will usually use one or the other. Common criteria for these fit indices are shown in Table 14.5. As noted earlier, these criteria have been generally supported in simulation studies (e.g., Hu & Bentler, 1998, 1999). These authors (Hu & Bentler, 1999) recommended using them in combination, such as SRMR and CFI. But things are not that simple. More recent research, however, has shown potential problems with cut-off criteria for good versus poor fitting models (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Fan & Sivo, 2007; Marsh, Hau, & Wen, 2004). Many things affect fit indices, so adherence to rigid cutoff criteria simply will not work. For example, concerning RMSEA, "The authors' analyses suggest that to achieve a certain level of power or Type I error rate, the choice of cutoff values depends on model specification, degrees of freedom, and sample size" (Chen et al., p. 462). I continue to use the criteria listed in the Table but not as a fixed good model/bad model criteria. If all the fit indices look good, I'm tentatively OK with a model. If some are good and some are not so good, I try to understand why and investigate how the model could be improved. Loehlin and Beaujean recalled the adage that a man with two watches is never sure what time it is (Loehlin & Beaujean, 2017, chap. 2). With the multitude of fit statistics available we have many more than two watches. My advice is that if those watches are fairly close to one another, you will have a pretty good idea of the correct time. If they tell you vastly different times, you'd better investigate further.

I have mixed feelings about the use of χ^2 as a primary measure of fit for a single model; you should realize that other writers are more supportive. Kline, for example, suggests always reporting χ^2 and its associated df and p , and for models with a statistically significant χ^2 , carefully examining the residual correlation matrix for the sources of misfit. I am less enamored with χ^2 , but that may be because most of my research involves large samples (thousands of cases) and, given that χ^2 is so affected by sample size, my χ^2 's are usually statistically significant. But this is not bad advice, especially if you use sample sizes in the 75 to 200 or maybe even 400 range (<http://davidakenny.net/cm/fit.htm>, retrieved March 22, 2018). With larger samples, I think χ^2 is less useful as a stand-alone measure of fit. And the more general point here is even more useful: when fit, as measured by your preferred indexes, is less than stellar, then you should investigate further. The residual correlations and the standardized residuals (covariances) are an excellent resource for doing so. The modification indices (and

Table 14.5

Fit Index	May be useful for & other notes	Common criteria
χ^2	Useful as stand-alone measure with $N = 75$ to 400. Tested for statistical significance with df	non-significance supports the model
$\Delta\chi^2$	Comparing competing, nested models, $N \leq 1000$	Non-significance supports the model with larger df ; significance supports the model with smaller df
RMSEA	Stand-alone measure of fit. Can calculate confidence intervals around RMSEA, and test whether an obtained RSMEA is statistically significantly different from some value (e.g., .05)	$\leq .05$ = good fit (close fit) $\leq .08$ = adequate fit $\geq .10$ = poor fit
SRMR	Stand-alone measure of fit. Intuitively appealing	$\leq .08$ = good fit, although $\leq .06$ may be a better criterion
CFI	Stand-alone measure of fit. Some research suggests ΔCFI may be useful in invariance testing (see Chapter 20)	$\geq .95$ = good fit $\geq .90$ = adequate fit
TLI	Stand-alone measure of fit	$\geq .95$ = good fit $\geq .90$ = adequate fit
AIC	Comparing competing models—nested or non-nested	Smaller is better
BIC	Same as AIC, but larger reward for parsimony	Smaller is better
aBIC	Same as AIC, in between AIC and BIC in reward for parsimony	Smaller is better

associated expected parameter change) are also useful (Heene, Hilbert, Freudenthaler, & Bühner, 2012); these will be discussed in subsequent chapters.

This difference highlights several important points. First, different writers will emphasize different measures of fit and will give different advice. Second, knowledge about the performance of various fit indexes will increase over time, and common wisdom concerning fit indexes will change over time. If you are to be a responsible user of SEM for research, you need to stay attuned to new developments. I've already noted Kenny's web pages as a good source of current advice; presumably he will continue to update his advice. You should also pay attention to the conventions and norms in your own area of research, because these will differ from one area to another. Third, you should always keep in mind that even when a model fits the data well, that does not mean that the model is correct, and that you have found "truth." There may be alternative models with equivalent or better fit. And even if your model beats out all alternatives, it's just a model; it does a good job in explaining the observed relations among the variables you have looked at, and those are just a small slice of the infinite number of variables you could have considered. Fourth, and very importantly, what you should NOT do is cherry-pick your fit index to support the model that you prefer. Although you can change your preferences for fit indices over

time, that change should be based on knowledge and experience, not the desire to support a particular model.

Comparing Competing Models

This lack of concrete, universally accepted rules of thumb concerning what constitutes a good model, and the fact that good models are not “correct” models, highlights a fifth important point: although stand-alone measures of fit are very useful, it is even better when we can compare the fit of alternative, competing models. As already noted, I’ve found $\Delta\chi^2$ useful for this purpose, when those models are nested, and given reasonable sample sizes (say up to 750 or 1,000 or so). Also useful are the AIC and other information criteria indexes, and these have the advantage of being usable and useful when models are non-nested. The various information criteria indices (AIC, BIC, aBIC) give different rewards for parsimony (Mulaik, 2009 has shown that these parsimony “rewards” depend on sample size, and disappear with large samples). At least in my recent research I have found the aBIC to provide a happy medium between too strict versus too forgiving (although the AIC seemed to work better in a confirmatory factor analysis simulation study our research group did recently: Keith, Caemmerer, & Reynolds, 2016). All of these indexes (AIC, BIC, aBIC) are only useful for comparing competing models (they are not used or useful as stand-alone indexes), and smaller is better.

Finally, please recognize that my term “stand-alone” fit index is not common. I think it makes sense to talk about stand-alone measures of fit versus measures useful for comparing competing models, but this is not common usage. More commonly, writers will refer to measures such as CFI and TLI as incremental or relative fit indexes (because they compare the target model with a null model), and measures such as RMSEA, SRMR, sometimes and χ^2 as absolute fit indexes. Additional categorizations vary from writer to writer. Kenny adds the term “comparative fit” indexes for indexes such as AIC that are only useful for comparing competing models (<http://davidakenny.net/cm/fit.htm>, retrieved March 22, 2018), others refer to AIC and related measures as information-theoretic measures (e.g., Arbuckle, 2017), and so on.

Table 14.5 shows the fit indices we have discussed so far, and their usefulness (in my opinion) for evaluating a single model or competing models. Some other indices are included as well.

SUMMARY

We covered a great deal of ground in this chapter; a review is needed. In this chapter we made the transition from estimating path models using multiple regression analysis to estimating these models with programs specifically designed for structural equation modeling (SEM). Several such programs are available, each with its own advantages. Some programs have student, or demonstration, versions available, downloadable from the Web; these student versions work the same as do the full-featured programs, but generally limit the number of variables that can be analyzed. There are SEM modules available for R, the free statistical programming language. I have used the Analysis of Moment Structures (Amos) program to illustrate SEM programs. The illustrations and explanations should translate easily to other SEM programs, and the web site illustrates input and output from several SEM programs (www.tzkeith.com).

All our previous discussions of path analysis translate directly to path analysis via SEM programs. To illustrate, we re-estimated the parent involvement path model from Chapter 13 using Amos. One advantage of Amos is that a drawing of a path model is used as the specification of the model, and the drawing, along with the data, is sufficient for conducting the

analysis. The input drawing for reanalysis of the parent involvement example was similar to the conventions we have used previously for developing path models. The one difference was that, by convention, we set the paths from the disturbances to the endogenous variables to 1, which allowed us to estimate the variance of the disturbance. (In multiple regression the variance of the disturbance was assumed to be 1, but the path was estimated.) We will follow this convention with other unmeasured-latent variables as well: setting the scale of the unmeasured variable by setting the path from it to one measured variable to 1; this convention merely says the scale of the unmeasured variable is the same as that of the measured variable.

Output from the SEM program (in this case Amos) included standardized and unstandardized path models, as well as detailed output. The more detailed output included standard errors of unstandardized coefficients and their associated z (or t) statistics, as well as tables of direct, indirect, and total effects.

Our next example was an overidentified model designed to determine the extent of the influence of Homework time on high school Grades. The model did not include all the paths that could have been drawn, a specification that is the same as drawing the paths but constraining them to a value of zero. The solved model suggested that Homework had a moderate effect on Grades, and Previous Achievement and Family Background had moderate to strong effects on time spent on Homework.

In earlier chapters I noted that overidentified models can be used to provide feedback about the adequacy of the model. A chief advantage of SEM programs is that they naturally provide such feedback. We can solve for paths using covariances, but we can also do the reverse: solve for the covariances using the solved path model. When models are overidentified, these two matrices (the actual and the implied covariance matrices) will differ to some degree. Fit statistics or indexes describe this degree of similarity or dissimilarity and provide feedback as to the adequacy of the model in explaining the data.

The degrees of freedom for a model describe the extent to which it is overidentified, or the parsimony of the model. The Homework model had 2 degrees of freedom; there were two paths we could have drawn but did not. The more we constrain values in the model to zero (or some other value), the more parsimonious the model and the larger its degrees of freedom.

Numerous fit indexes are provided by SEM programs. We focused on the root mean square error of approximation (RMSEA) as a primary index of fit for a single model; RMSEAs of .05 or less suggest a good fit, with values of .08 or less suggesting an adequate fit (cf. Browne & Cudeck, 1993). I also discussed using the comparative fit index (CFI), and the Tucker-Lewis index (TLI) as methods of assessing the fit of a single model. For these indexes, values above .95 suggest a good fit, and values above .90 suggest an adequate fit. The standardized root mean square residual (SRMR) is an intuitively appealing index of fit, and represents the average difference between the actual correlations among measured variables and those predicted by the model; SRMR values below .08 (or perhaps .06) represent a good fit. χ^2 , along with the df and its associated probability, may be used to assess the fit of a model, with statistically significant values suggesting a lack of fit and statistically not significant values suggesting a good fit of the model to the data. Although common, χ^2 has problems as a measure of the fit of a single model.

A major advantage of SEM programs and measures of fit is that they may be used to compare competing theoretical models. We compared the fit of the initial Homework model to several competing models; these comparisons tested basic hypotheses embodied in these models. Although I downplayed the use of χ^2 as the measure of fit of a single model, I argued that if models are nested (one is a more constrained version of the other) χ^2 can be a useful method of comparing the two models. The more parsimonious model (the one with the larger df) will also have a larger χ^2 . If the change in χ^2 is statistically significant

compared to the change in df , our rule of thumb is to prefer the less parsimonious model; but if the $\Delta\chi^2$ is statistically not significant, our preference is for the more parsimonious model. A $\Delta\chi^2$ of close to 4 is statistically significant with 1 df . Other fit indexes for comparing competing models are the Akaike Information Criterion (AIC) and the sample size adjusted Bayes Information Criterion (aBIC), in which smaller values are better.

Although overidentified models allow us to compare competing models, representing competing hypotheses about the effects of variables on each other, there may be several or many models that are equivalent to our preferred model. These equivalent models may also represent competing hypotheses about effects but are statistically indistinguishable from our preferred model. I briefly explained and illustrated the rules for generating equivalent models, and noted that you should consider such equivalent models as you develop your own models. You can guard against the threat represented by equivalent models in the same way you build valid models in the first place, through careful consideration of theory, previous research, time precedence, and so on.

The flip side of equivalent models is that there are other overidentified but nonnested models that are not equivalent with the model under consideration. Such models can be very useful for testing and rejecting threats to path models. Knowing the rules for generating equivalent models also allows us to develop nonequivalent models. We illustrated this value by testing a nonequivalent version of the Homework model with the path from Homework to Grades reversed.

Other advantages of SEM programs are that they can be used to analyze nonrecursive models and can provide for more powerful analysis of longitudinal models. Longitudinal data may also be useful for overcoming some challenges posed by equivalent models by clarifying causal direction. I briefly illustrated such models but did not delve into them in detail.

We now have two methods for analyzing path models: multiple regression analysis via a generic statistical analysis program or SEM programs. If you are using MR to conduct path analysis, there is no real benefit for developing overidentified models. If you are using a SEM program, however, it is worth developing overidentified models when possible, because of the fit information the programs provide. Similarly, if you are interested in overidentified models, comparing competing models, or in more complex forms of path models, I encourage you to use a SEM program to estimate these models.

EXERCISES

1. Reproduce the Homework models used in this chapter. Make sure your results match mine (note there may be minor differences in estimates if you are using programs other than Amos). Are there additional models that you might test?
2. Try estimating a similar homework model using the NELS data.
3. In the section introducing overidentifying models, I stated that “not drawing a path is the same as drawing a path and fixing or constraining that path to a value of zero.” Demonstrate the truth of this statement. Using the homework model, constrain, for example, the path from Previous Achievement to Grades to zero and check the fit of the model. Now delete that same path. Is the fit the same? Are the parameter estimates the same for the two models?
4. Focus on the equivalent models in Figure 14.14. Note the difference between these and the initial model (model A). Which rule or rules were used to produce each equivalent model? Check your answers against those in note 5. Try estimating one or two of these models to demonstrate that they are indeed equivalent.
5. Henry, Tolan, and Gorman-Smith (2001) investigated the effect of one’s peers on boys’ later violence and delinquency. Figure 14.21 shows one model drawn from their study,

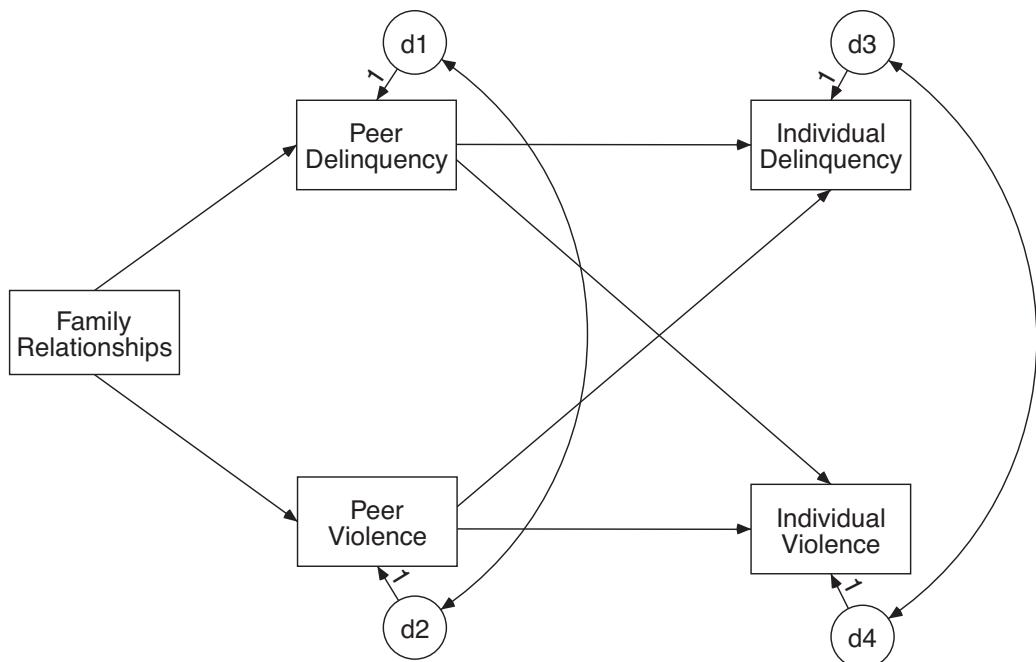


Figure 14.21 One model from Henry et al. (2001).

their “fully mediated” model. Family Relationships is a composite of measures of family cohesion, beliefs about family, and family structure, with high scores representing a better functioning family; the violence and delinquency variables are measures of the frequency of violent and nonviolent delinquent offenses for peers and individuals. The model is longitudinal, with Family Relationships measured at age 12, Peer variables at age 14, and Individual variables at age 17. The model is also contained in the file “henry et al.amw” on the Web site.

Data consistent with those reported in the original article are in the SPSS file “Henry et al.sav” or the Excel file “Henry et al.xls.” Analyze and interpret this model. Which variable had a more important effect on boys’ delinquency: peers who are delinquent or peers who are violent? Which variable was more important for boys’ violence? What were the indirect effects of Family Relationships on Individual’s Violence and Delinquency? Test an alternative model to determine whether Family Relationships directly affect the outcome variables. (The Henry et al., 2001, article reported correlations among variables. The data used in this example were simulated data designed to mimic these correlations. The Family Relationships variable used here was a combination of three variables from the original article.)

6. Estimate the nonrecursive trust model from Figure 14.17. The model (trust nonrecursive model 1.amw) and the data (trust norec sim data.xls) are included on the accompanying Web site. Second, assume that the Man’s Trust affects his partner, but not the reverse: delete the path from Woman’s Trust to Man’s Trust, along with the correlated disturbance. Are these models nested? Why? Compare the fit of the two models. What conclusions do you reach from these model comparisons?
7. Exercise 6 in Chapter 4 was “designed to explore further the nature of common causes, and what happens when non-common causes are included in a multiple regression.”

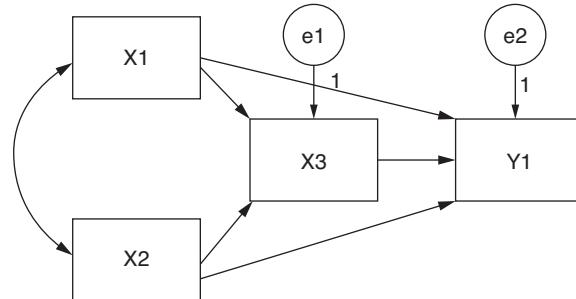


Figure 14.22 Understanding common versus non-common causes, and their effects on path estimates.

We will begin our analysis of these data here, and will return to them in Part 2 when we have the tools to explore them more completely.” This example was also used in Chapter 9 to illustrate the effects of common versus non-common causes.

You now have the tools to explore these data and this topic more completely. To review, there were three data files for this exercise, all including variables labeled X1 X2 X3 and Y1. For the data in the first file (common cause 1.sav), X2 influenced both X3 and Y1 (it was a common cause). In the second file (common cause 2.sav) variable X2 had no effect on Y1. In the third file (common cause 3.sav), variable X2 had no effect on variable X3.

Analyze these data using an SEM program. For all three data sets, the model you should estimate is illustrated in Figure 14.22. Compute and examine the correlations among the variables in all the data sets. All correlations are statistically significant, correct?

Now analyze the model shown for all three data sets. Notice that for data set 2 the effect of X2 on Y1 is essentially zero. For data set 3, what is the effect of X2 on X3? Is X2 a common cause of X3 and Y1 in either model? Now notice the effect of X3 on Y1 in each model. What should happen to this path when the variable X2 is removed from the model?

Analyze each data set without variable X2 in the model. What happens to the magnitude of the path from X3 to Y1?

Consider what your findings mean concerning the nature of controlling for common versus non-common causes.

Notes

- 1 We could also analyze the NELS raw data, but would then need to consider methods of dealing with missing data in more depth than I want to right now. We will return to this issue in the chapter on latent means in SEM, Chapter 19.
- 2 How could you do so? It is fairly easy to do so using the tracing rules. For example, to calculate the correlation between Minority and Grades implied by the model, here are the possible tracings between Minority and Grades (where → represents a path and ↘ represents a correlation):
 1. Minority → PreAch → Grades + Minority → PreAch → Homework → Grades + Minority → Homework → Grades = $-.20 \times .44 + -.20 \times .22 \times .15 + -.04 \times .15 = -.089$, and
 2. Minority ↘ FamBack × (FamBack → PrevAch → Grades + FamBack → PrevAch → Homework → Grades + FamBack → Homework → Grades) = $-.30 \times (.42 \times .44 + .42 \times .22 \times .15 + .17 \times .15) = -.067$.

When added together, these equal $-.156$, the implied correlation between Minority and Grades. Another way to think about this is that the tracings listed under 1 are the total effects of Minority

- on Grades, and those listed under 2 are the total effects of Family Background on Grades, times the correlation of Minority and Grades.
- 3 I know properly it should be chi-squared, but, by convention, it's chi-square.
 - 4 For example, type the χ^2 and *df* into two cells in Excel. Click on another cell, then Insert, Function. Click on CHIDIST and follow the directions to obtain the probability associated with χ^2 with the indicated *df*.
 - 5 What do correlated disturbances mean? Focus on model C in Figure 14.14, which shows a correlated disturbance between d1 and d2. The disturbances represent all other influences on the corresponding variables other than those shown in the model. The correlation between d1 and d2 in this model suggests that the other influences (other than Minority and Family Background) on Previous Achievement and Homework may be correlated. What this means, in turn, is that there may be other common causes of Previous Achievement and Homework not included in the model. Correlated disturbances can also be used to denote an agnostic causal relation; that is, we think that Previous Achievement and Homework are causally related but don't know the direction. As shown in Appendix C, one helpful way of thinking about partial correlations is that they represent the correlation between disturbances.
 - 6 Model B and Models C and D resulted from the application of rule 2. Model E, with the paths between Homework and Minority and Homework and Family Background reversed, also resulted from the application of this rule. Model F builds on Model E. Note that with model E Homework and Grades now have the same causes. We can therefore apply rule 3 to Model E and reverse the path from Grades to Homework. It may not be obvious, but Models E and F are nonrecursive models. Note that in Model F, for example, Homework affects Background, which affects Previous Achievement, which affects Homework, and so on.

15

Error

The Scourge of Research

Effects of Unreliability	335
<i>The Importance of Reliability</i>	335
<i>Effects of Unreliability on Path Results</i>	336
Effects of Invalidity	339
<i>The Meaning and Importance of Validity</i>	339
<i>Accounting for Invalidity</i>	340
Latent Variable SEM and Errors of Measurement	343
<i>The Latent SEM Model</i>	344
Summary	346
Exercises	347
<i>Notes</i>	347

Recall the assumptions required to interpret regression coefficients (paths) as estimates of effects of one variable on another:

1. There is no reverse causation; that is, the model is recursive.
2. The exogenous variables are perfectly measured, that is, they are completely reliable and valid.
3. A state of equilibrium has been reached. This assumption means that the causal process has had a chance to work.
4. No common cause of the presumed cause and the presumed effect has been neglected; the model includes all such common causes (Kenny, 1979, p. 51).

We have dealt with several of these assumptions, such as the effect of neglecting a common cause, and I promised we would return to assumption 2: the assumption of perfect or near perfect measurement of the exogenous variables. Obviously, this assumption is violated routinely—perfect measurement is rare to impossible—but what effect does this violation have on our research? In addition, inaccurate measurement of the endogenous variables also affects estimates in path models.

It is worth noting that issues of reliability and validity of measurements affect *all* research, not just that based on path analysis and multiple regression. Many of us think of measurement as separate from statistics, but they are inexorably intertwined. In a laboratory experiment our experimental conditions (the exogenous variable) may be clear-cut and thus perfectly

measured (e.g., treatment versus control), but the dependent (endogenous variable) (e.g., a measure of self-esteem) may be considerably less reliable. This lack of reliability may result in an underestimation of the effect of the experimental treatment, with even a truly meaningful finding showing up as statistically nonsignificant. In applied research, there may be variations in the treatments by those responsible for providing the experimental treatment. Teachers in an experiment designed to compare the effects of two methods for teaching reading may use other methods outside the experimental procedure. This variation is, in fact, unreliability and invalidity in the independent (exogenous) variable, which will also cloud the results of the research. In fact, the effect of measurement on decision making affects *every* aspect of life. Your physician may prescribe or not prescribe medication for high blood pressure depending on her measurement of your blood pressure; if her measurements are unreliable, however, you may receive unnecessary treatment or not receive needed treatment. You may have costly repairs completed on your car based on unreliable measurement, and so on. Measurement accuracy affects all research and all decisions made from these measurements. Why, you may wonder, does it?

EFFECTS OF UNRELIABILITY

The Importance of Reliability

In classic measurement theory, we might administer a test, or survey, or other measurement to a group of people. There will be variation in their scores; some people will score high, some low. We also know that there will be error in their scores; all measurement involves error. This aspect of scores is represented in Figure 15.1. V represents the total variation in a set of scores on some measurement. This total variance can be divided into variation due to error (V_e) and true score variation (V_t): $V = V_t + V_e$. Using this definition, reliability is the proportion of the true score variance to the total variance: $\frac{V_t}{V}$. This makes sense: the greater the error in a set of scores, the less a person's score on that measure is a result of true variation and the less reliable the measurement.

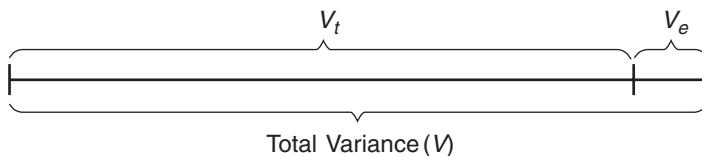


Figure 15.1 Variance definition of reliability. Reliability is the proportion of true score variance to total variance ($\frac{V_t}{V}$).

Figure 15.2 illustrates the effects of unreliability in path analytic format. In this graphic, a person's score on any measurement is affected both by the person's *true score* and by *errors* of measurement. In this graphic, error is equivalent to V_e and the true score to V_t . Note that the actual, measured score is the only measured variable in this model; both the true score and the error are unmeasured and unknown.

The reliability of a test, scale, survey, or other measure places an upper limit on the correlation that the measurement can have with any other measurement. As a general rule, a second variable will correlate with the measured score through correlation with the true score. That is, other variables will generally correlate with the V_t portion of the variable illustrated in Figure 15.1, not the V_e portion. This, then, is the reason that measurement quality affects statistics and research: a less reliable measurement limits the correlations a variable can have with any other variable. Since correlations are the statistic underlying multiple regression, path

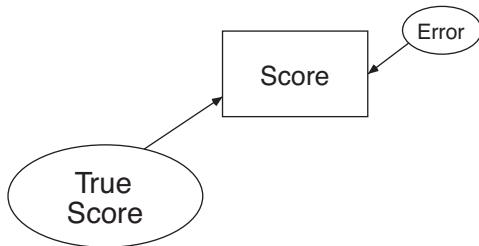


Figure 15.2 Path analytic definition of reliability; a person's score on a test or measurement is affected by their true, but unknown, score and by error.

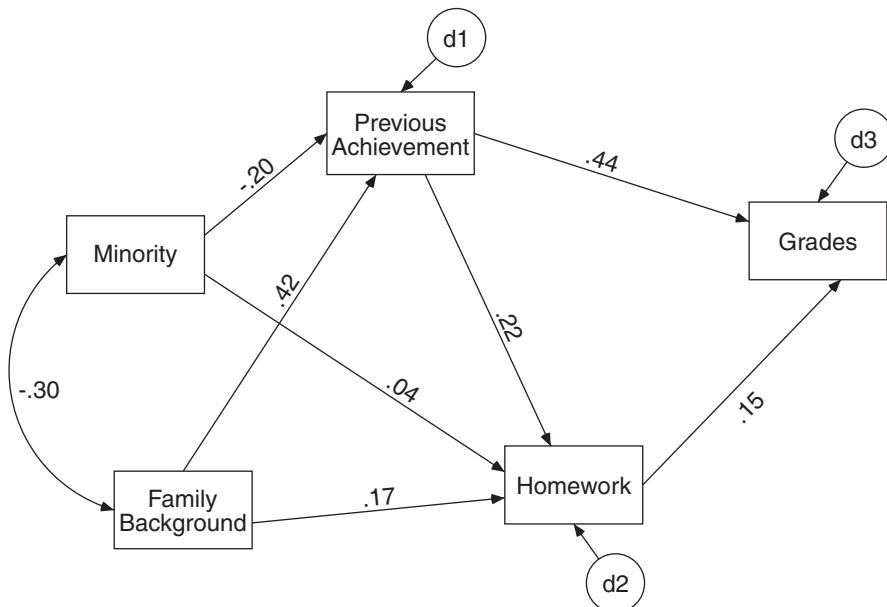


Figure 15.3 Homework model from Chapter 14 revisited.

analysis, ANOVA, and other derivatives of the general linear model, unreliable measurement causes us to underestimate the effects of one variable on another in *all* these methodologies.

Effects of Unreliability on Path Results

What effect does measurement error have on path analytic results? Figure 15.3 shows the results for the homework model from Chapter 14. In this model, whether we realize it or not, we are assuming that all the variables in the model are measured without error, with perfect reliability. As researchers, we may recognize that the variables in the model are measured with different degrees of error, but the model assumes they are all error free.

Focus on the variable of homework. Homework is based on student self-report of the average amount of time students spend on homework in several academic areas. Undoubtedly, error is inherent in this variable, not only because of the self-report nature of the questions, but also because, perhaps more importantly, students were asked to approximate their average amount of time per week. I would not be surprised to discover that this variable had a reliability of only about .70, with a corresponding error of 30%. If we build such estimates into the path model, what will be the effect on the estimates of paths?

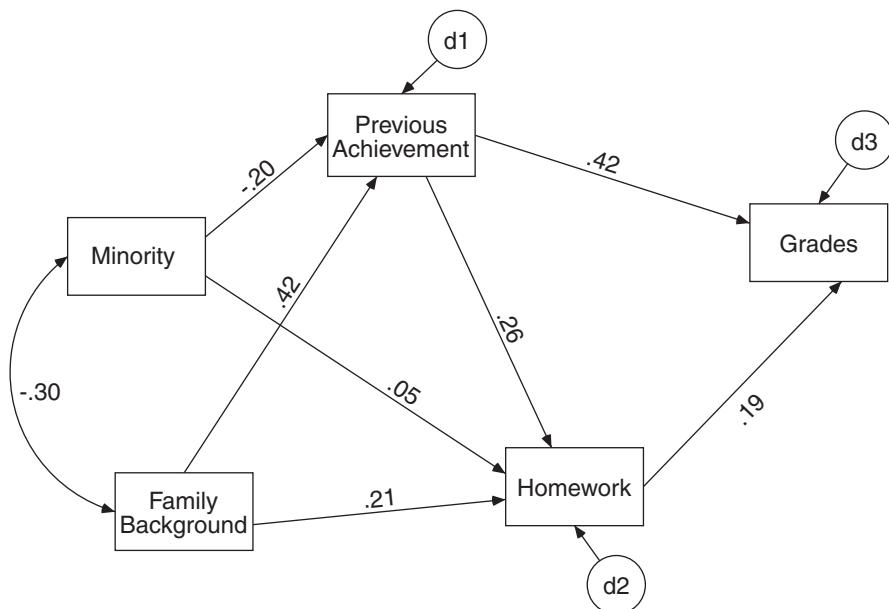


Figure 15.4 Effects of error. This model recognizes and accounts for the unreliability in the Homework variable; with this recognition, the apparent effect of Homework on Grades increases.

Figure 15.4 shows a model that recognizes this unreliability ($\text{reliability} = .70$, $\text{error} = .30$) in the Homework variable. Note the increase in the apparent effect of Homework on Grades, from .15 in Figure 15.3 to .19 in Figure 15.4. What this means is that when we assumed that the error-laden Homework variable was perfectly reliable, as in Figure 15.3, we *underestimated* the true effect of Homework on Grades. In contrast, when we recognize the error inherent in this variable, we obtain a more realistic and larger estimate of the effect. This is also the most common effect of error in models: unreliability artificially reduces our estimates of the effects of one variable on another.

Note also that many of the other paths in the model are different from those in Figure 15.3. Indeed, all paths to Homework increased in magnitude, and the path from Achievement to Grades decreased slightly. Recognition of the error that exists in the Homework variable resulted in changes in many of the paths in the model.

But Homework is not the only less-than-perfectly-reliable variable in the model. What about Grades? Grades were also based on student self-report, plus there are well-known problems with Grades as measures of student learning, including variations in grading standards from teacher to teacher, the unreliability of teacher-made tests and other components of grades, and the likely clouding of other variables (e.g., students' apparent interest and motivation) in teachers' grading practices. Given these deficiencies of Grades, it is probably reasonable to estimate their reliability at a maximum of .80 (and 20% of the variation in scores due to error).

Figure 15.5 shows the results of recognition of this level of error for the Grades variable (assuming perfect reliability for the other variables in the model). In this model, compared to Figure 15.3, the magnitude of the paths to Grades from both Previous Achievement (from .44 to .50) and Homework (from .15 to .17) increased.

Although it is not obvious from these figures, the effects of unreliability are different depending on whether the variable in question is exogenous or endogenous. Briefly, error

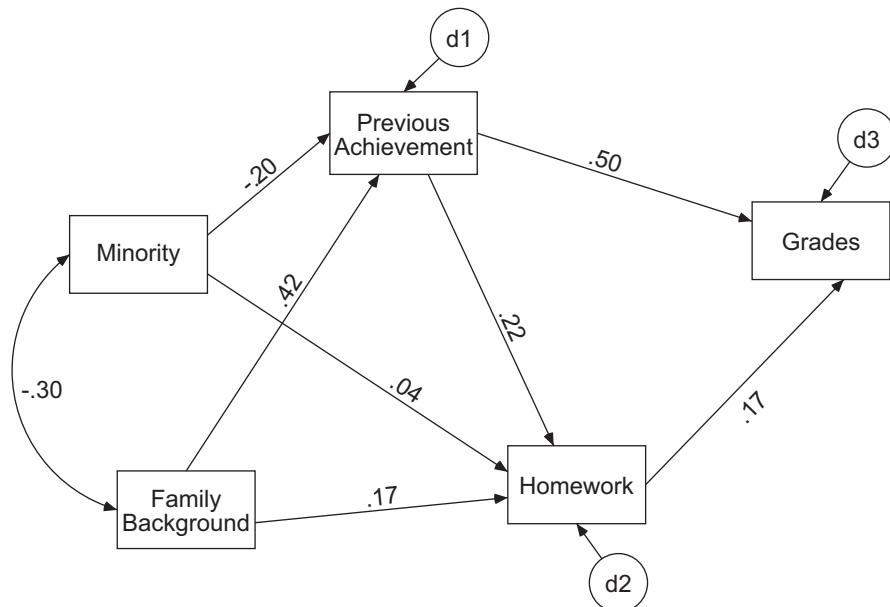


Figure 15.5 Effects of error. This model shows the result of recognizing the error inherent in the Grades variable.

in an exogenous variable affects both the standardized and unstandardized paths, as well as their statistical significance. Paths from other exogenous variables (in addition to the error laden one) may be affected. In contrast, error in an endogenous variable affects only *standardized* estimates of effects, leaving unstandardized effects unchanged. The unstandardized paths for the model shown in Figure 15.5 would be the same as those for the model shown in 14.3, despite the differences in the standardized paths. This difference is why error in exogenous variables is more consequential than error in endogenous variables. When a variable is in the middle of a model—exogenous in relation to some variables, endogenous for others—the results of error are more complex, as in the example recognizing error in Homework (Figure 15.4). The bottom line is that measurement error affects estimates of effects, but is more serious for exogenous variables [for more information, see Bollen, 1989 (chap. 5); Rigdon, 1994; or Wolfe, 1979].

These examples have corrected for unreliability in a single variable. What would happen if we were to recognize the unreliability in *all* the variables in the model? If you think about it, all the variables in the model are unreliable to one degree or another. Even Ethnic Minority background, probably the most reliable variable, likely has some error. Students may not read the survey question accurately, students who could legitimately claim to belong to more than one ethnic group are allowed only one answer, some students simply knowingly mark the wrong response, and there may be errors in coding of students' responses. For whatever the reason, even this variable likely includes some error.¹

The model shown in Figure 15.6 attempts to recognize the error inherent in every variable in the model. For this example, I assumed that error was responsible for 30% of the variability for Homework, 20% for Grades, 5% for Minority, 20% for Family Background, and 10% for Previous Achievement. These are plausible estimates. Note that every parameter estimate in the model changed from those shown in Figure 15.3. Most estimates increased in magnitude, but one, the path from Minority to Previous Achievement, decreased (from $-.20$

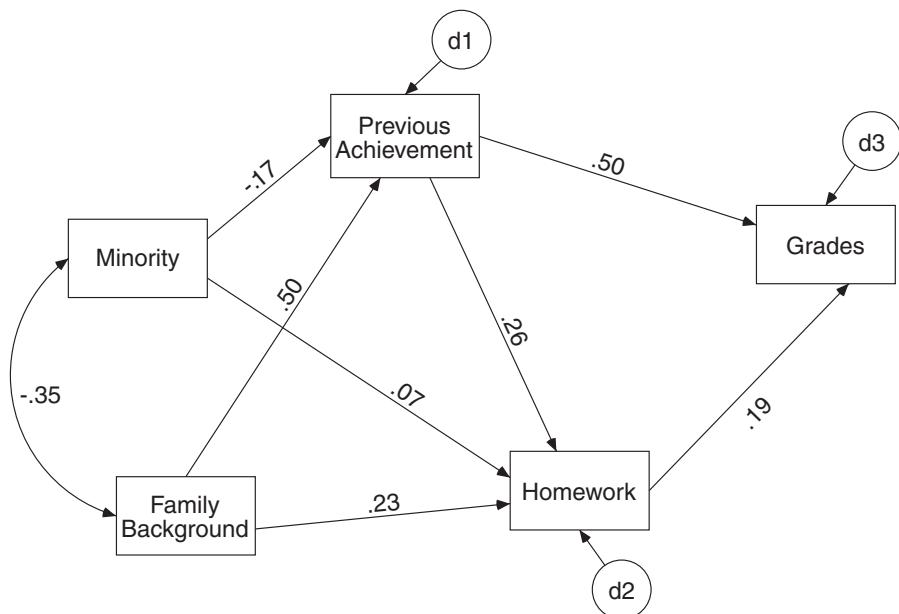


Figure 15.6 Effects of error. This model recognizes the error inherent in all the variables in the model. Compare the coefficients here with those shown in Figure 15.3.

in Figure 15.3 to $-.17$ in Figure 15.6). Recognition of the error inherent in the variables in our models will often, although certainly not always, result in larger estimates of the effects of one variable on another. With such complex patterns of errors, estimates may increase, decrease, or stay the same.

These examples illustrate the effects of measurement error on estimates of the influence of one variable on another in path analysis (as well at MR, ANOVA, etc.). What can researchers do to avoid misestimating such effects? We can strive for better measures, but no measures are error free. We could also correct the correlations for all the variables in the model using estimates of each variable's reliability and the common formula for correcting for attenuation, $r_{T_1 T_2} = r_{12} / \sqrt{r_{11} \times r_{22}}$, where $r_{T_1 T_2}$ is the corrected, or "true" correlation, r_{12} is the original correlation, and r_{11} and r_{22} are the reliabilities of the two variables. This solution is not very satisfying for several reasons. First, it divorces the correction from model testing; indeed, the process smacks of statistical voodoo. Second, when there are multiple estimates of reliability, such as with several studies providing estimates, it is unclear which estimate should be used. Conversely, no estimates of reliability may be available for a given measure. Finally, although this method might deal with unreliability of measures, it ignores problems of invalidity.

EFFECTS OF INVALIDITY

The Meaning and Importance of Validity

What effect does invalidity have on estimates of effects? In classic measurement theory, validity may be considered as a subset of reliability. An example will illustrate how these measurement concepts are related. Suppose that you are interested in the effects of reading comprehension on subsequent delinquent behavior. One task is to measure reading comprehension. You will find that different tests of reading comprehension use different methods of measurement. Test 1, for example, asks research participants to read a passage on one page

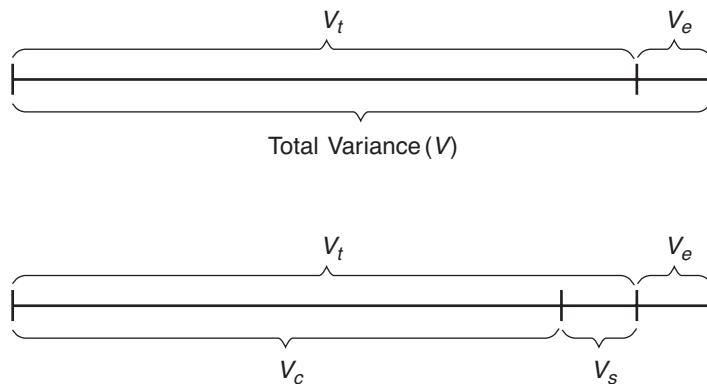


Figure 15.7 True score variance may be further subdivided into common variance (V_c) and specific or unique variance (V_s). Validity is related to common variance.

and then point to one picture (out of four choices) on the next page that best illustrates what they read in the passage. Test 2, in contrast, asks participants to read a passage (e.g., “stand up, walk around the table, then sit down”) and then do what the passage requested. Test 3 uses a “cloze” procedure; the participant reads a passage with one or several words missing and then supplies the missing words based on the meaning of the text.

It is clear that each of these tests measures reading comprehension to some degree. But each test also measures something else, something other than reading comprehension. Test 1 also measures the ability to translate something read into a picture; Test 2 measures the ability to act out something read; Test 3 measures the ability to pick from one’s knowledge store the word or words that will make the most sense when inserted in a passage. Each test may measure these unique skills reliably, but these skills are not the same as reading comprehension.

We are also not interested in the variation in scores due to these unique skills. We are interested in the effects of *reading comprehension* on delinquent behavior, not the effects of the ability to translate text into mental pictures (Test 1) or the unique skills measured by other tests on delinquent behavior. This variation due to these unique skills will not be removed through correction for attenuation, however, because these skills are measured reliably and are not due to error.

As shown in Figure 15.7, it is possible to extend the earlier variance definition of reliability. The true score variation (reliability) can be divided further. Using the reading comprehension example, one component of the true score variation for each test is the variance that these three tests have in common, the common variance, or V_c . What do the three Reading Comprehension Tests measure in common: reading comprehension! Each test also measures something unique or specific, however, and this component of the true score reliability is symbolized as V_s , for specific variance. The common variance, V_c , is an estimate of the validity of each test and thus demonstrates that validity is a subset of reliability. The V_s , the unique or specific variance of each test, is sometimes called the *specificity*, or the unique variance. For our present purposes, it represents invalidity and needs to be taken into account in our research on the effects of reading comprehension on delinquent behavior.

Accounting for Invalidity

How can we take this invalidity into account? Another way of conceptualizing the problem is as a path model, as shown in Figure 15.8. The diagram illustrates the influences on individuals’ scores on the three Reading Comprehension Tests. Each person’s score on each test is first

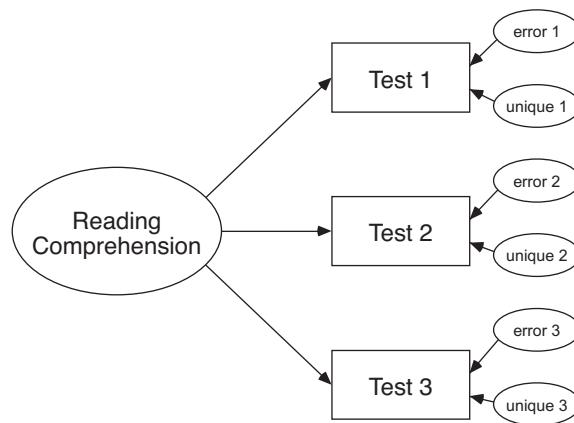


Figure 15.8 Using path models to understand validity. Individuals' scores on three tests of Reading Comprehension are affected by their true level of Reading Comprehension and by error and the unique aspects of each test.

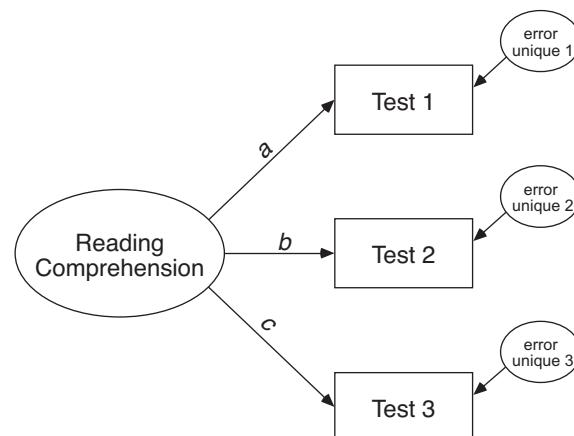


Figure 15.9 Reading Comprehension measurement model; we can generate equations to solve for the paths from Reading Comprehension to the three Tests.

affected by his or her level of reading comprehension. Reading Comprehension—the true level of reading comprehension—is an unmeasured or latent variable and is thus enclosed in an oval. Each person's scores on each test are also affected by error (unreliability) and by that person's level of the unique skills measured by each test (one's ability to translate text into pictures, and so on). These are also unmeasured variables. Our primary interest, of course, is in the Reading Comprehension latent variable.

Figure 15.8 is just another path model, and we can solve it in much the same way we solved the path models in Chapter 12. Figure 15.9 shows a slight revision of the model, with the error and unique variances combined for each variable and the paths labelled to help develop equations. Figure 15.10 shows the correlations among the three tests. As in Chapter 12, we can use the tracing rule to develop equations:

$$\begin{aligned}
 r_{12} &= ab, \\
 r_{13} &= ac, \text{ and} \\
 r_{23} &= bc.
 \end{aligned}$$

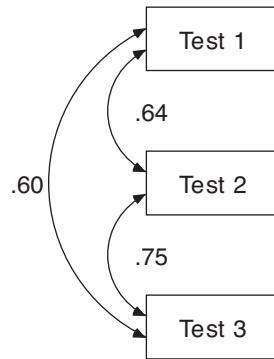


Figure 15.10 Correlations among the three Tests used to solve for the paths.

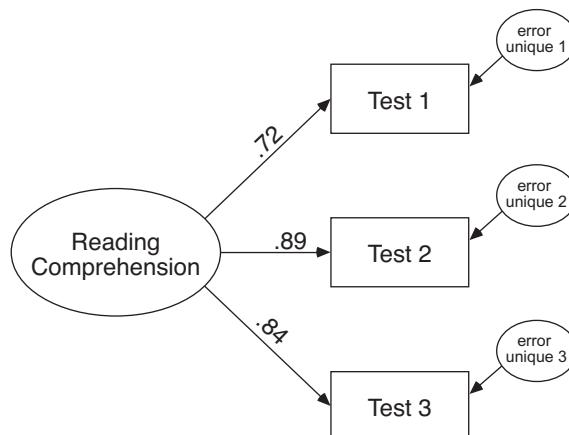


Figure 15.11 Solved Reading Comprehension measurement model.

If we combine the first two equations, we get $r_{12}r_{13}=abac$, which can be simplified as $a^2bc = r_{12}r_{13}$, or $a^2 = r_{12}r_{13}/bc$. Because $bc = r_{23}$ from the third equation, $a^2 = r_{12}r_{13}/r_{23}$ and $a = \sqrt{r_{12}r_{13}/r_{23}}$. We can also solve for b and c : $b = \sqrt{r_{12}r_{23}/r_{13}}$ and $c = \sqrt{r_{13}r_{23}/r_{12}}$. If you substitute the correlations in these equations, $a = .716$, $b = .894$, and $c = .839$. Figure 15.11 shows the model with the path estimates inserted.

Interestingly, what we have done by solving for the paths in Figure 15.11 is a simple (confirmatory) factor analysis. Figure 15.12 shows output from a factor analysis of these three

Factor Matrix^a

	Factor
	1
TEST_1	.716
TEST_2	.893
TEST_3	.839

Extraction Method: Principal Axis Factoring.

a. 1 factors extracted. 11 iterations required.

Figure 15.12 Reading Comprehension measurement model solved via factor analysis. Our measurement model is a (confirmatory) factor analysis.

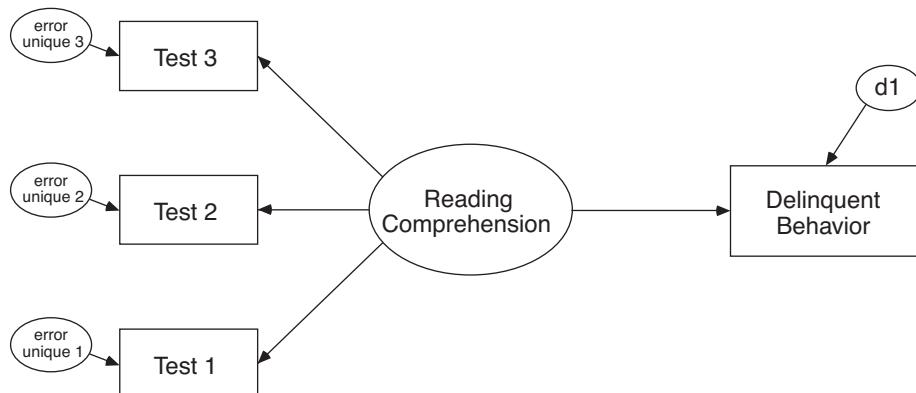


Figure 15.13 We could use the Reading Comprehension factor, or latent variable, in a structural equation to more accurately determine the effect of Reading Comprehension on Delinquent Behavior.

items in SPSS; the factor loadings from the output are the same as the paths from the Reading Comprehension latent variable to the three reading Tests.² The example nicely illustrates the thinking underlying factor analysis: there is a latent, or unmeasured, variable, or factor, that affects individuals' scores on these three Tests and does so to different degrees. The example also illustrates the equivalence of several terms. What we have been referring to as latent or unmeasured variables are equivalent to the *factors* from factor analysis. These latent variables or factors are also much closer to the *constructs* we are interested in than are our normal, error-laden measurements.

Our primary interest, of course, was the influence of Reading Comprehension on Delinquent Behavior. Because we can solve the model to estimate the Reading Comprehension latent variable, we could also use the latent variable in an analysis of the effects of Reading Comprehension on Delinquent Behavior, as in Figure 15.13 (once we were able to measure Delinquent Behavior).

LATENT VARIABLE SEM AND ERRORS OF MEASUREMENT

To return to our more general problem, perhaps this means that the solution to the problem of less-than-perfect measurement is not to correct all the correlations for attenuation but to obtain multiple measures of each construct in our path model, separately factor analyze these items, and then use the factor scores in our path analyses, rather than the original items or tests. This process will rid our measures of both invalidity and unreliability (because ridding the measure of invalidity will rid it of unreliability) and will allow us to get closer to the constructs we are interested in. Although this solution makes sense conceptually, it too has drawbacks. The multistep process separates the different factor analyses (the measurement model) from the testing of the path model (the structural model). It would be preferable to be able to conduct *all* analyses simultaneously.

This is what latent variable SEM does: it performs confirmatory factor analysis and a path analysis of the resulting factors at the same time. In the process, latent variable SEM removes the effects of unreliability and invalidity from the estimation of the effect of one variable on another. By doing so, the method gets closer to constructs we are really interested in. Thus, instead of doing research on the effects of a measure of Reading Comprehension on a measure of Delinquent Behavior, we can come closer to studying the effect of *true* Reading Comprehension on *true* Delinquent Behavior. As another example, if we are interested in the effects of income on job satisfaction, we are not interested in the effects of reported income

(the number someone reports on a survey) on perceptions of job satisfaction. Instead, we are interested in the effects of *true* income on *true* job satisfaction. In other words, we want to strip away the fog of invalidity and measurement error and get at the true constructs of interest. Likewise, if we are studying the effect of social skills on peer acceptance, we are not really interested in the effects of someone's perceptions of peoples' social skills on their perceptions of acceptance; we are interested in the effects of *real* social skills on *real* acceptance. Latent variable SEM helps us get closer to this level of analysis.

The Latent SEM Model

Figure 15.14 illustrates a generic latent variable structural equation model. To refresh our jargon, latent variables are the same as unmeasured variables or factors. Latent variables are inferred from the measured variables, and they more closely approach the constructs of true interest in the research. Latent variables are enclosed in circles or ovals. Measured variables are also known as observed variables or manifest variables. They are the variables that we actually measure in our research through tests, surveys, observations, interviews, or other methods. Measured variables are enclosed in rectangles. Scores on a reading test, survey items concerning time spent on homework, records of social interactions from playground observations, and a count of errors on a computer task are all examples of measured variables. Actual reading comprehension, time really spent on homework, true social acceptance, and actual mental processing speed are the latent variables we hope to determine through these measured variables. In research we are almost *always* interested in the latent rather than the measured variables, but we often have to settle for the error-laden measured variables as approximations of the latent variables. Not necessarily so with latent variable SEM!

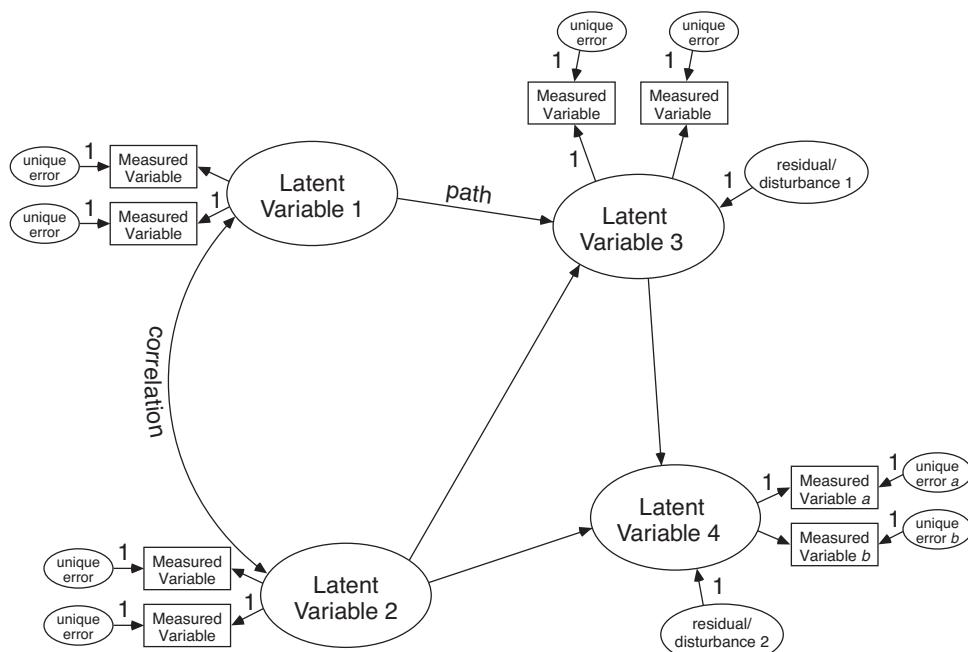


Figure 15.14 Latent variable structural equation model. The model includes a confirmatory factor analysis of the latent and measured variables, as well as a path analysis of the effects of one latent variable on another.

Understanding the Model

The system of paths from the latent to the measured variables is sometimes referred to as the *measurement model*. It is a simultaneous confirmatory factor analysis of all the latent variables in the model. The system of paths and correlations among the latent variables is often referred to as the *structural model*. You can think of it as a path analysis of the latent variables.

You may find it confusing at first glance that both the measured variables and the endogenous latent variables have smaller latent variables pointing to them, but you will soon see that these have previously been defined. Recall that endogenous variables (effects) in a path model have latent variables pointing toward them; these latent variables are generally called either residuals or disturbances. The disturbances represent all *other* influences on the endogenous variables other than those shown in the model. It is the same with *latent* endogenous variables. We need to account for all other influences on the latent variables besides those shown in the model; again we do so with other latent variables known as disturbances or residuals (or errors). The small latent variables pointing to the measured variables represent the unique and error variances that we wish to remove from consideration in the SEM as we focus on the true effects of one (latent) variable on another. These unique and error variances are often simply referred to as error or occasionally by Greek letters (e.g., theta delta, theta epsilon), a convention from LISREL. More generally, both types of variables (errors and disturbances) are sometimes referred to as errors.

In fact, you can think of errors and disturbances in the same way. Latent Variable 2 and Latent Variable 3 are not the only influences on Latent Variable 4; there may be a multitude of other such influences outside the model. Residual/Disturbance 2 represents all the other influences on Latent Variable 4 other than those shown in the model. Likewise, Latent Variable 4 is not the only influence on Measured Variable *a*; unique and error variances also affect this and other Measured Variables. “Unique error *a*” represents these influences. Although I will continue to treat disturbances and errors as different, you can thus think of them as “all other influences” on the measured and latent variables.

Figure 15.15 shows a latent variable SEM version of the homework model used in the last few chapters. Note that each variable in the model, except Minority, was measured via multiple measured variables and thus can be estimated by a latent variable. Ethnic Minority background, still indexed by a single item, is still a measured variable in this model. We will explore this model in more detail in subsequent chapters. What is interesting to note at the present time is that the use of latent variables rather than measured variables increased the estimate of the effect of Homework on Grades from .15 (from the path analysis) to above .20 (in the latent variable SEM; the value is not shown in the figure). Again, the latent variable analysis has the advantage of removing measurement error from consideration in the model and thus getting closer to the level of the constructs we are really interested in (e.g., Homework and Grades). The latent variable estimates in this model should thus be the more accurate ones.

We will explore this example in more depth in subsequent chapters. First, however, we will take an important detour in the next chapter into confirmatory factor analysis, or the measurement model portion of latent variable SEM.

Before leaving this chapter, I reiterate that the problems discussed here—the effects of imperfect measurement in research—apply to all research. Here I have focused on the effects of measurement error in nonexperimental research—path analysis and structural equation modeling—because that is our focus. But measurement error affects all research, experimental and nonexperimental, whether analyzed through ANOVA, correlations, multiple regression, or SEM.

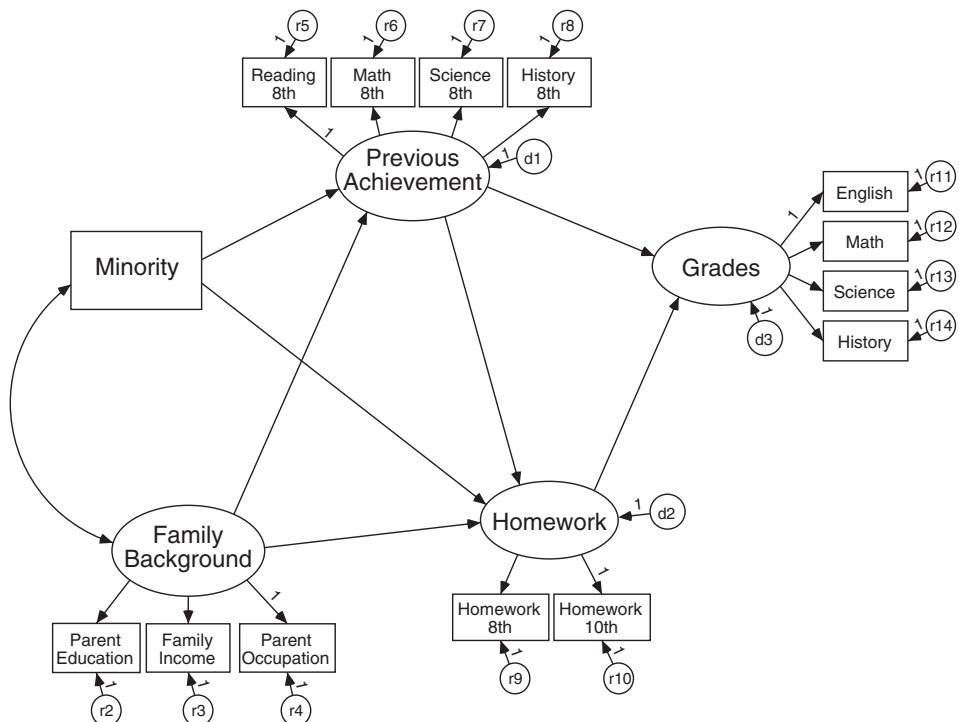


Figure 15.15 A latent variable version of the homework model. All constructs except Ethnicity are indexed by multiple measures. We will examine and test this model in Chapter 18.

SUMMARY

One assumption required to interpret regression (path) coefficients in a causal fashion is that the exogenous variables be measured without error. We rarely satisfy this assumption and thus need to know the effect of this violation on our estimates of the effects of one variable on another. To expand this discussion, I noted that unreliability and invalidity affect *all* types of research, not just path analysis and multiple regression. Problems in measurement in both the independent and dependent variables affect our research results.

Reliability is the converse of error. Error-laden measurements are unreliable, and reliable measurements contain little error. We can consider reliability from the standpoint of variance by thinking of true score variance as the total variance in a set of scores minus the error variance. In path analytic form, we can think of a person's score on a measurement as being affected by two influences: their true score on the measure and errors of measurement. The true score and error influences are latent variables, whereas the actual score the person earns on the measurement is a measured variable. These concepts are important for research purposes, because other variables generally correlate with the true score, but not the error. For this reason, the reliability of a measurement places an upper limit on the correlation a variable can have with any other variable. Unreliable measurements can make large effects look small and statistically significant effects look nonsignificant.

The path models we have been discussing so far assume that the variables in our models are measured with perfect reliability. In a series of models, I demonstrated what would happen when we recognized and quantified the unreliability of these measurements. When unreliability was taken into account in these models, the apparent effects of one variable on

another changed and usually increased. Taking unreliability into account in our research will improve our estimates of the effects of one variable on another.

Reliability is not the only aspect of measurement that needs to be considered, however; there is also validity. I demonstrated that a measurement may be reliable but may focus on some unique skill, rather than the central skill we are interested in. Said differently, a measurement may be reliable but may not be a valid measure of our construct of interest. As it turns out, validity is a subset of reliability. We can get closer to valid measurement, closer to the constructs of interest in our research, by using multiple measures of constructs.

Latent variable structural equation modeling seeks to move closer to the constructs of interest in our research by using such multiple measures. With latent variable SEM, we simultaneously perform a confirmatory factor analysis of the measured variables in our research to get at the latent variables of true interest, along with a path analysis of the effects of these latent variables on each other. In the process, latent variable SEM removes the effects of unreliability and invalidity from consideration of the effects of one variable on another and avoids the problem of imperfect measurement. In the process, latent variable SEM gets closer to the primary questions of interest: the effect of one construct on another.

Although our discussion focused on the effects of imperfect measurement in multiple regression and path analysis, it is worth remembering that measurement affects every type of research, however that research is analyzed. With the addition of latent variables to SEM, we are able to take measurement problems into account and thus control for them.

EXERCISES

1. Pick a research study in your area of interest. Describe the latent variables, the constructs the authors were interested in. What was the construct of interest underlying the independent variable(s)? What was the construct of interest underlying the dependent variable(s)? What measured variables were used to approximate these constructs?
2. How could you convert this research from a measured variable study into a latent variable study? Think of ways to include multiple measures of the researchers' independent and dependent variables. Draw a model incorporating both measured and latent variables.
3. What is the advantage of moving from a measured to a latent variable approach? What might happen to the estimates of effects with this transition?
4. Find an article in your area of interest that uses latent variable structural equation modeling (it may be referred to as structural equation modeling or covariance structures analysis). Read the article. Do the authors discuss reasons for using latent over measured variables? Do they link latent variables with the constructs of reliability and validity? How do they label the disturbances? The error and unique variances of the measured variables?

Notes

1. Some of these examples are actually systematic errors rather than random errors and are thus not considered unreliability. I include them because I want you to consider the errors that can be included in even such a straightforward item.
2. The results are equivalent only because the example is so simple. With more items and multiple factors, the results of a confirmatory analysis will be different from those of an "exploratory" factor analysis (from SPSS), and even the results of an exploratory analysis will differ depending on the method used and the assumptions made. The example is useful for heuristic purposes, however, as a conceptual illustration of what factor analysis is.

16

Confirmatory Factor Analysis I

Factor Analysis: The Measurement Model	348
An Example With the DAS-II	349
<i>Structure of the DAS-II</i>	350
<i>The Initial Model</i>	351
<i>Standardized and Unstandardized Results: The Initial Model</i>	353
<i>Testing a Standardized Model</i>	354
Testing Competing Models	358
<i>Testing Plausible Cross-Loadings</i>	358
<i>A Three-Factor Combined Nonverbal Model</i>	360
Model Fit and Model Modification	363
<i>Modification Indexes</i>	363
<i>Residuals</i>	366
<i>Adding Model Constraints and z Values</i>	368
<i>Cautions</i>	369
Hierarchical Models	369
<i>Higher-Order Model Justification and Setup</i>	369
<i>Higher-Order Model Results</i>	370
<i>Bifactor Model Justification and Setup</i>	373
<i>Bifactor Model Results</i>	374
<i>Comparing the Hierarchical Models</i>	376
Additional Uses of Model Constraints	379
Summary	384
Exercises	385
Notes	387

FACTOR ANALYSIS: THE MEASUREMENT MODEL

This chapter will focus in more detail on the *measurement model* portion of latent variable structural equation modeling, more generally known as *confirmatory factor analysis*. At its most basic level, factor analysis is a reduction technique, a method of reducing many measures into fewer measures. The methodology works by placing scales or items that correlate highly with each other on one factor, while placing items that correlate at a low level with

each other on different factors. Because one primary reason items correlate highly with one another is that they measure the same construct, factor analysis provides insights as to the common constructs measured by a set of scales or items. Because it helps answer questions about the constructs measured by a set of items, factor analysis is a major method of establishing the internal validity of tests, questionnaires, and other measurements. You can also think of factor analysis as a method of establishing convergent and divergent validity: items that measure the same thing form a factor (converge), whereas items that measure different constructs form a separate factor (diverge).

With *exploratory* factor analysis (not covered in this text), one analyzes a set of items or scales that presumably measures a smaller set of abilities, traits, or constructs. Decisions are made concerning the method of factor extraction to use, the method for deciding the number of factors to retain, and the method of factor rotation to use. Given these choices and the data, the results of the analysis will suggest that the items measure a smaller number of factors. For example, factor analysis of 13 scales may suggest that these scales measure four constructs. The output from the analysis will include factor loadings of each scale on the four factors and, if oblique rotation is used, the correlations of the factors with each other. The researcher then decides on names for the factors based on the constructs they presumably reflect, a decision based on the loadings of the variables on the factors, relevant theory, and previous research.

With *confirmatory* factor analysis one uses previous research and relevant theory to decide in advance what the factors or constructs are that underlie the measures. Just as in path analysis, we propose a model that underlies the variables of interest. The fit statistics then provide feedback concerning the adequacy of the model in explaining the data. I hope it is obvious why the methods are termed exploratory versus confirmatory factor analysis. With the first, we examine the results and decide what the various scales are measuring, whereas with the second we decide what the various scales are measuring and then examine the results to find out how accurate our predictions were. This dichotomy is an obvious simplification—we can use exploratory factor analysis in a confirmatory fashion and can use confirmatory factor analysis in an exploratory fashion—but it is still a useful distinction.

The development of factor analysis is inexorably linked with development of theories of intelligence and intelligence tests. Early intelligence researchers developed the methods of factor analysis to understand the nature and measurement of intelligence, and factor analysis continues to be a major method of supporting and challenging the validity of intelligence tests. For this reason, I will illustrate the method of confirmatory factor analysis using intelligence test data. Note that this is one of two chapters on the topic of CFA; we will return to more advanced CFA topics after learning more about latent variable SEM.

AN EXAMPLE WITH THE DAS-II

The Differential Ability Scales, Second Edition (DAS-II; Elliott, 2007) is among the most commonly administered individual intelligence tests for children. The DAS-II includes a series of short verbal and nonverbal subtests and is appropriate for children and youth ages 2½ to 18. The DAS-II is a common portion of a broader psychological evaluation for children and adolescents who are having learning, behavioral, or adjustment problems. It may be used to help evaluate children for special programs (e.g., those for children with learning disabilities and gifted programs); diagnose learning, behavioral, and neurological problems; or provide information relevant to an intervention to ameliorate such problems.

Structure of the DAS-II

Although the DAS-II includes different tests for children at different ages, all 21 tests from the battery were standardized for children ages 5–8. We will analyze data for 12 of these tests designed to measure four underlying constructs. The test names and a portion of the theoretical structure of the DAS-II are shown in Figure 16.1. Although I will not describe the subtests in detail, they measure a variety of verbal and nonverbal skills. For example, the Word Similarities subtest requires children to explain the construct shared by three words. In contrast, Pattern Construction requires the child to construct, from pictures, geometric designs using two-colored foam squares and blocks. According to the author, the DAS-II measures verbal reasoning (Verbal Ability), nonverbal, inductive reasoning (Nonverbal Reasoning),

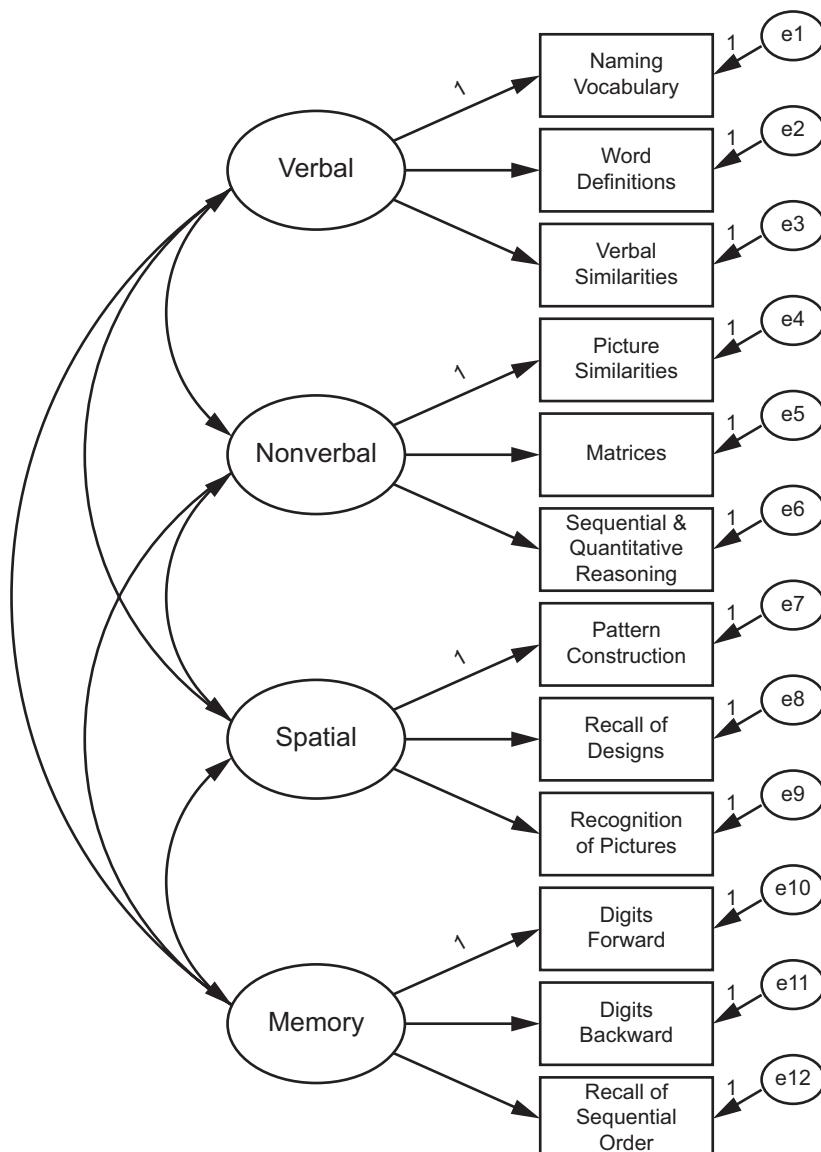


Figure 16.1 Initial DAS-II model. Does the DAS-II measure verbal, nonverbal, and spatial reasoning skills, along with short-term memory?

visual–spatial reasoning (Spatial), and short-term memory (Memory) (and some other skills not discussed here; you may also see these abilities referred to as Crystallized Intelligence, or Gc; Fluid Intelligence, Gf; Visual Processing, Gv; and Short-Term Memory, Gsm). The figure shows which subtests are designed to measure which skills. This structure is reflected in the actual scoring of the test. For children 7 and older, for example, scores for two tests per construct are added together to form Verbal, Nonverbal Reasoning, Spatial, and Working Memory composite scores.

The Initial Model

Figure 16.1 is also the setup for a confirmatory factor model (indeed, the figure is the input for analysis in Amos), with the constructs underlying the DAS-II shown in ovals as latent variables and the 12 subtests (the actual measurements we obtain) shown in rectangles as measured variables. The arrows in the figure make explicit the causal assumptions underlying such testing and models. The paths point from the constructs to the subtests in recognition of the implicit assumption that each person's level of verbal reasoning ability is the primary influence on his or her score on the Word Definitions subtest, for example, whereas each person's level of visual spatial ability is the primary influence on his or her score on the Pattern Construction subtest. Although the constructs the test is designed to measure are the primary influence on individuals' scores on the subtests, you know from the last chapter that individuals' scores on each subtest are also influenced by unreliability and by the unique characteristics of each test. This latter statement makes sense intuitively as well. Although Pattern Construction and Recall of Designs (in which children draw complex designs from memory) obviously both require visual and spatial skills, they also both obviously require different specific skills, such as the mental translation of a two-dimensional picture into three-dimensional form (Pattern Construction) versus visual and spatial memory skills (Recall of Designs). These unique skills and unreliability are represented by the small latent variables pointing to each subtest labeled e1 through e12. e7, for example, represents all influences on children's scores on the Pattern Construction subtest other than Spatial Ability.

You will recall that latent variables have no set scale, and we must set the scale of each latent variable to estimate the model. Recall also that one way to set the scale of a latent variable is to set a path from each latent variable to one measured variable at 1.0. This is done in Figure 16.1. The Verbal factor's scale is set to be the same as that for the Naming Vocabulary subtest. The choice of which measured variable to use is arbitrary; I have simply set the scale of each factor to be the same as the first variable that measures this factor. Without these constraints to set the scales of the latent variables the model would be underidentified. Kline (2016, p. 148) calls this method of setting the scales of latent variables "unit loading identification," or ULI. The scales for the unique-error variances are also set to the same scale as their corresponding subtests: e1 is set to have the same scale as Naming Vocabulary, e2 as Word Definitions, and so on. Alternatively, we could also set the scale of the factors by setting the variance of each factor to 1.0 (we will come back to this point).

The model shown in Figure 16.1 also includes correlations among each construct thought to be measured by the DAS-II. It is commonly recognized that cognitive tests and cognitive factors are positively correlated (Carroll, 1993). The model shown in the figure is on the Web site (www.tzkeith.com) in the folder for this chapter under the name "das 2 first order 1.amw"; Mplus script is also available.

The DAS-II manual includes tables of correlations among the subtests for each age level 2½ through 17 (along with means and standard deviations). The averaged covariance matrix for these subtests for children 5–8 is shown in Table 16.1; this matrix was produced as a by-product of CFA analyses designed to determine whether the DAS-II measures the same

Table 16.1 Average Covariance Matrix for the DAS-II for Ages 5 through 8

<i>rowtype_</i>	<i>varname_</i>	<i>wdss</i>	<i>vsss</i>	<i>sqss</i>	<i>soss</i>	<i>rpss</i>	<i>rdss</i>	<i>psss</i>	<i>nvss</i>	<i>mass</i>	<i>dfss</i>	<i>dbss</i>
cov	<i>wdss</i>	91.52										
cov	<i>vsss</i>	58.43	104.34									
cov	<i>sqss</i>	42.21	53.06	94.14								
cov	<i>soss</i>	50.34	54.85	54.15	113.43							
cov	<i>rpss</i>	27.88	36.19	44.16	40.00	102.09						
cov	<i>rdss</i>	31.19	44.29	49.98	48.46	48.19	99.74					
cov	<i>psss</i>	36.86	41.62	39.46	37.48	33.56	41.31	106.38				
cov	<i>pcss</i>	37.21	48.52	54.01	48.53	40.82	55.81	38.72	84.74			
cov	<i>nvss</i>	53.94	59.64	44.16	52.13	33.62	44.43	39.34	46.83	102.13		
cov	<i>mass</i>	41.67	47.50	60.40	54.75	40.01	41.38	39.48	47.34	40.05	104.59	
cov	<i>dfss</i>	44.45	51.76	46.54	61.56	32.91	46.32	37.07	44.28	49.54	39.66	121.52
cov	<i>dbss</i>	41.78	50.76	52.77	62.90	37.51	47.28	36.98	47.58	43.47	51.09	56.20
n		800	800	800	800	800	800	800	800	800	800	103.25
mean		50.03	50.21	49.99	49.94	50.01	49.75	49.92	50.07	50.22	50.02	49.65

Note: Variable names: wdss = Word Definitions; vsss = Verbal Similarities; sqss = Sequential & Quantitative Reasoning; soss = Recall of Sequential Order; rpss = Recognition of Pictures; rdss = Recall of Designs; psss = Pattern Construction; nvss = Naming Vocabulary; mass = Matrices; dfss = Digits Forward; dbss = Digits Backward.

constructs across its age levels (Keith, Low, Reynolds, Patel, & Ridley, 2010). The matrix of covariances among the twelve subtests was used to estimate the model shown in Figure 16.1. The covariance matrix is also contained in the Excel file “DAS 2 cov.xls” and the SPSS file “DAS 2 cov.sav.” The sample size for the analyses was 800.

Standardized and Unstandardized Results: The Initial Model

Figure 16.2 shows standardized results of the initial analysis of the DAS-II model. First, focus on the fit indexes. The Root Mean Square Error of Approximation (RMSEA) was .046, lower (better) than our rule of thumb for good models of .05. The Standardized Root Mean Square Residual (SRMR) was .027, meaning that the average difference between the

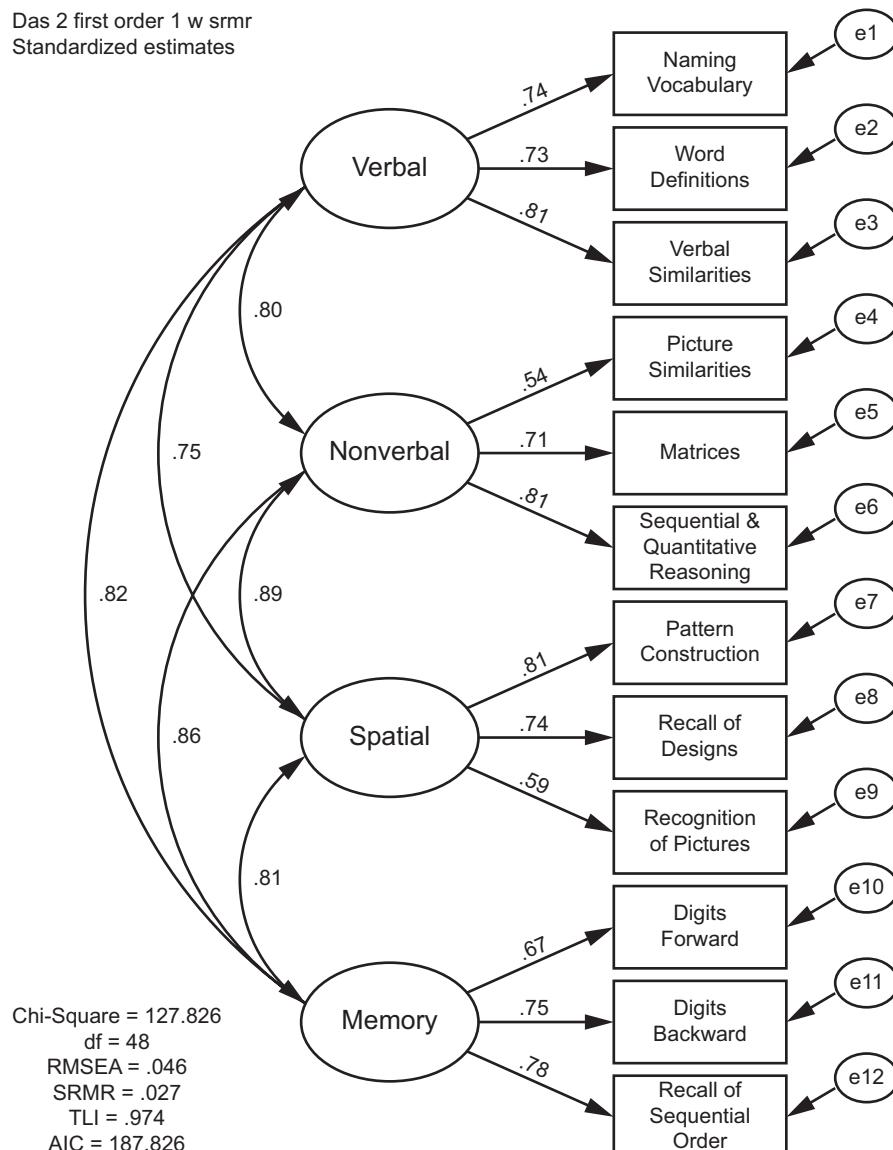


Figure 16.2 Standardized estimates for the initial DAS-II four-factor model

actual and the implied correlation matrices was only .027. The TLI (and the CFI, not shown) were above our target for a good model (.95). By these criteria, it appears that the DAS-II model fits the data well. In other words, the model that underlies the DAS-II indeed could have produced the correlations and covariances we observed among the DAS-II subtests, and the theoretical structure of the DAS-II is supported. Note, however, that the χ^2 is statistically significant (127.355 [47], $p < .001$), which, in contrast to the other indices, suggests a lack of fit of the model to the data. We will examine possible sources of misfit later in this chapter. Focusing on the model itself, it appears that most subtests provided relatively strong measures of the appropriate ability or construct; the factor loadings for most subtests on the Verbal, Nonverbal Reasoning, Spatial, and Memory factors were .6 or higher. The exceptions to these larger loadings were the Picture Similarities and Recognition of Pictures subtests on the Nonverbal and Spatial factors (loadings of .54 and .59, respectively). Although the detailed printout shows that these loadings were statistically significant, they are lower than for the other factors. Within factors, most subtests had fairly equivalent loadings on the factor they supposedly measure, although there are clearly subtests that have stronger loadings (e.g., Verbal Similarities, Sequential and Quantitative Reasoning, Pattern Construction). This difference in loadings suggests that the common construct measured by these tests is better measured by, for example, Sequential & Quantitative Reasoning than by Picture Similarities. The results also show that the latent factors correlate substantially with each other, with factor correlations ranging from .75 to .89.

Figure 16.3 shows the unstandardized estimates (“Regression Weights”) of the factor loadings, standard errors, z values (critical ratio, or CR), and p values (all less than .001). Note that the loadings used to set the scales of the latent variables, the ones that were set to 1, were not tested for statistical significance. Estimated values are tested for statistical significance; constrained parameters are not. In the second section of the figure are the standardized loadings (“Standardized Regression Weights”) followed by the covariances and correlations among latent factors. Note that all estimated paths (factor loadings) and covariances were statistically significant ($z > 2$), and that the standardized loadings match those in the figural display of the model.

Testing a Standardized Model

It is also possible to set the scale of the latent factors in the model by setting the factor variances to 1.0 (instead of setting one factor loading per factor to 1.0). The setup for such a standardized model—also known as unit variance identification, or UVI (Kline, 2016, p. 199)—for the DAS-II is shown in Figure 16.4. Although less consistent with SEM than the method of setting factor loadings, the factor variance method has two advantages. First, it results in the testing for statistical significance of all the factor loadings (which is often of prime interest in CFA studies). Second, this method produces *standardized* covariances (i.e., correlations) among the factors. Recall that a correlation matrix is a standardized covariance matrix, the result of standardizing the variables in the matrix (i.e., setting their variances to 1.0). Alternatively, you can think of a correlation matrix as just another variance–covariance matrix, but with all variances set to 1.0. Thus, when we set the variances of the factors in a CFA to 1.0, we have standardized the covariance matrix of factors. Figure 16.5 shows the *unstandardized* output for the UVI analysis just described. Note that the covariances (correlations) in this figure are the same as the correlations from the standardized output shown in Figure 16.2. The factor loadings, however, are still in an unstandardized metric (although a different unstandardized metric than previously).

The advantage of having the factor covariances standardized comes into play when we wish to compare competing models. Note the high correlation between the Nonverbal Reasoning and Spatial factors (.89). We may wonder if this correlation is statistically significantly

Regression Weights

		Estimate	S.E.	C.R.	P
nvss	<--- Verbal	1.0000			
wdss	<--- Verbal	.9418	.0489	19.2542	***
vsss	<--- Verbal	1.0996	.0526	20.8866	***
psss	<--- Nonverbal	1.0000			
mass	<--- Nonverbal	1.3056	.0926	14.0950	***
sqss	<--- Nonverbal	1.4059	.0936	15.0205	***
pcss	<--- Spatial	1.0000			
rdss	<--- Spatial	.9822	.0468	20.9828	***
rpss	<--- Spatial	.7949	.0485	16.3777	***
dfss	<--- Memory	1.0000			
dbss	<--- Memory	1.0346	.0576	17.9493	***
soss	<--- Memory	1.1187	.0609	18.3809	***

Standardized Regression Weights

		Estimate
nvss	<--- Verbal	.7395
wdss	<--- Verbal	.7333
vsss	<--- Verbal	.8052
psss	<--- Nonverbal	.5414
mass	<--- Nonverbal	.7102
sqss	<--- Nonverbal	.8082
pcss	<--- Spatial	.8139
rdss	<--- Spatial	.7370
rpss	<--- Spatial	.5906
dfss	<--- Memory	.6692
dbss	<--- Memory	.7535
soss	<--- Memory	.7780

Covariances

		Estimate	S.E.	C.R.	P
Verbal	<--> Nonverbal	33.4641	3.0790	10.8684	***
Verbal	<--> Spatial	42.0275	3.2851	12.7934	***
Verbal	<--> Memory	45.4283	3.6929	12.3016	***
Nonverbal	<--> Spatial	37.2672	3.2043	11.6305	***
Nonverbal	<--> Memory	35.2301	3.2871	10.7175	***
Spatial	<--> Memory	44.8726	3.5411	12.6721	***

Correlations

		Estimate
Verbal	<--> Nonverbal	.8049
Verbal	<--> Spatial	.7509
Verbal	<--> Memory	.8239
Nonverbal	<--> Spatial	.8922
Nonverbal	<--> Memory	.8561
Spatial	<--> Memory	.8100

Figure 16.3 Unstandardized and standardized text output for the initial DAS-II four-factor model.

Das 2 standardized fig 15-4
Model Specification

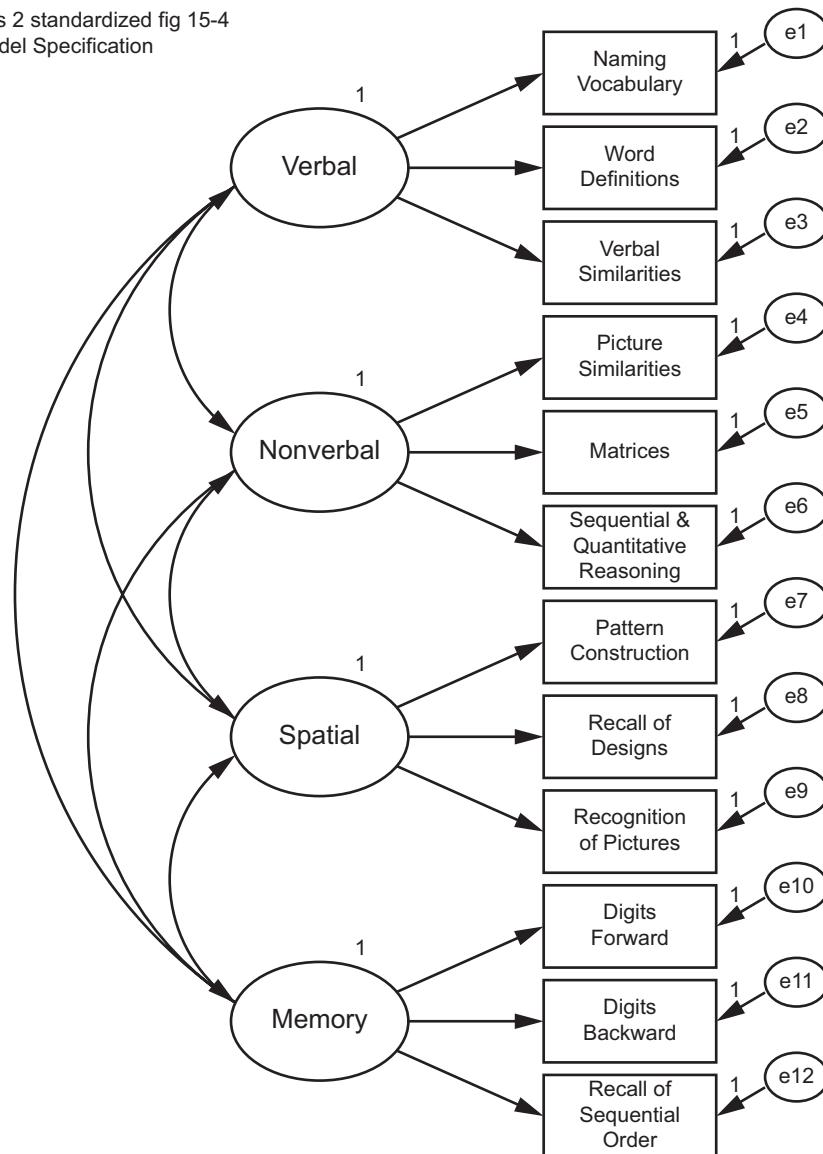


Figure 16.4 An alternative standardized method of specifying the initial DAS-II model. With this method, we set the scale of the latent variables by setting their variances to 1 instead of constraining factor loadings.

different from 1.0, meaning that the factors may be statistically indistinguishable. We could test this supposition by setting the factor correlation to 1.0 and comparing the fit of this model with the original model. However, model constraints apply to the *unstandardized* model only. Thus, if we wish to set a factor correlation to 1.0 (or some other value), we need to make the factor correlations equivalent to the factor covariances, using this standardized model. (As will be shown, a few other constraints are also needed to test the distinguishability of factors).

Although the primary results of a CFA—notably the fit indexes and the standardized output—will generally be the same whichever method is used, it is possible for some results

Das 2 standardized fig 15-5
unstandardized estimates

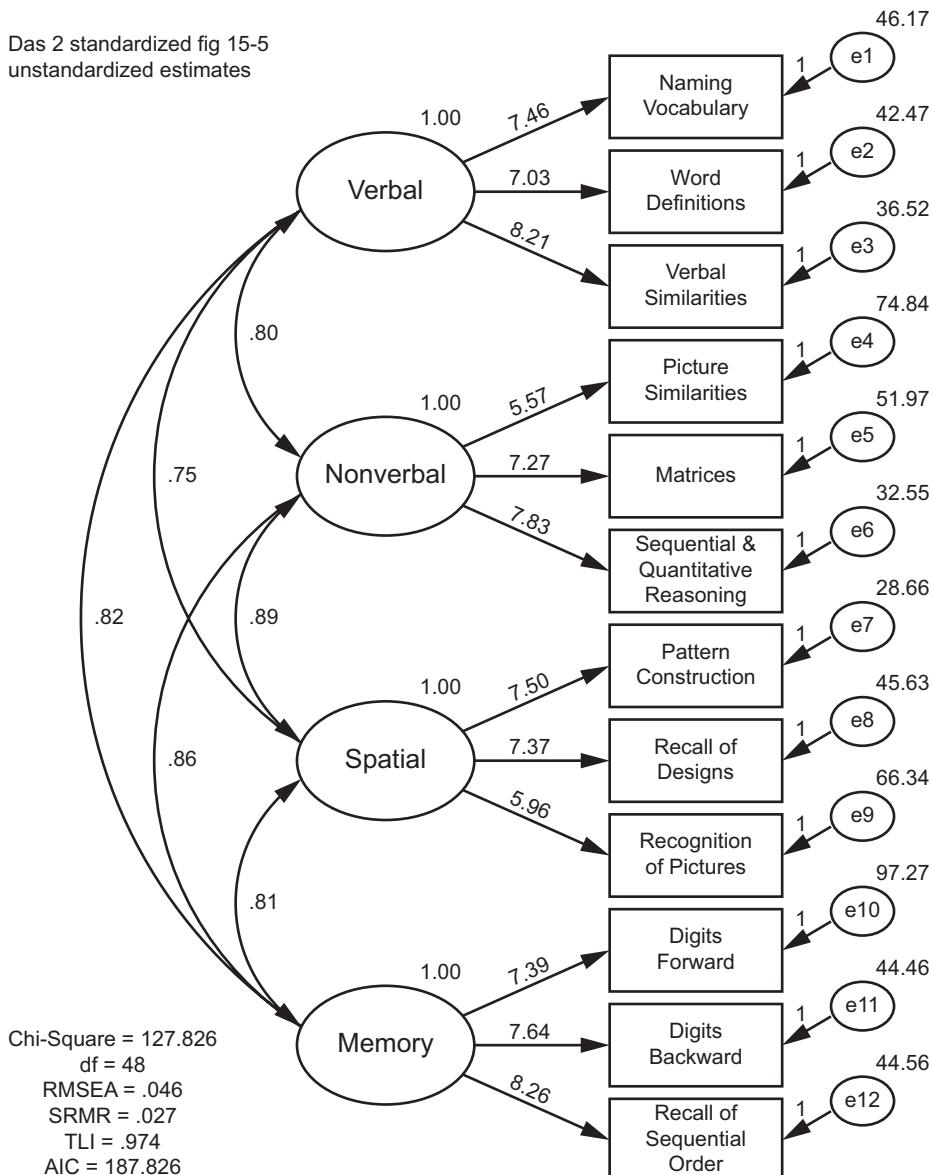


Figure 16.5 Unstandardized solution using the standardized model. Note that the factor covariances are now equivalent to the factor correlations from Figure 16.2.

to change slightly depending on whether the ULI (factor loading set to 1) or the UVI (factor variance set to 1) method is used. Likewise, results generally do not change—but sometimes do—depending on which factor loading is set to 1 using the ULI method. In particular, the unstandardized parameter estimates and the standard errors may change across the two methods, and the resulting z values (critical ranges) may change as well. What this means is that it is possible for a factor loading or factor covariance to be statistically significant using one method but not statistically significant using the other. (For more information, see Millsap, 2001. This article also shows that with complex models, where tests load on multiple factors, the fit of models can change depending on which factor-to-test path is set to 1.0.)

Before moving to the next topic, notice the numbers beside the unique and error variances: 46.17 for e1, 42.47 for e2, and so on. These numbers are the estimates of the combined unique and error variances of the various subtests. You can compare them to the variances of the variables shown in the diagonals of the variance–covariance matrix (Table 16.1). It appears that close to one half of the variation in the Word Definitions subtest is error and unique variance.

TESTING COMPETING MODELS

This initial example has tested the adequacy of a single confirmatory model. As in SEM, however, a more powerful use of the methodology is to compare alternative and competing models. I will briefly illustrate this method using the DAS-II example.

Note that in the models shown thus far each subtest has been assumed to measure one and only one underlying common ability or factor. But the constructs measured by tests may be and often are much more complex than this; indeed, it seems likely that some of the DAS-II subtests may measure more than one underlying ability. For example, the Recall of Designs subtest requires children to draw from memory designs they have seen a few seconds earlier. Doesn't it make sense to assume that this test requires short-term memory skills in addition to (or instead of) visual-spatial reasoning?

Testing Plausible Cross-Loadings

Figure 16.6 shows a model that tests this possible cross-loading by allowing Recall of Designs to load on both the Spatial and Memory factors. Note that this model and the initial model (e.g., Figure 16.1) are nested, because the initial model can be derived from this model by constraining the path (loading) from Memory to Recall of Designs to zero. Thus the model in Figure 16.1 is nested within the model shown in Figure 16.6. Figure 16.7 shows the standardized loadings for this model, along with some of the fit indices.

The DAS-II alternative cross-loading model fits the data well. Our primary stand-alone fit index, the RMSEA, suggests that the cross-loaded model explains well the test standardization data. The other stand-alone fit indexes (SRMR, TLI) also suggest a good fit of the model to the data. If we focus only on the fit of each model in isolation, we conclude that this model fits well, as does the earlier four-factor model. Our primary interest, however, is *relative* fit of the two models. In particular, we are interested in how this three-factor model compares to the initial model that did not include any cross-loadings. The cross-loaded model is less parsimonious than the initial model, with 47 degrees of freedom shown in Figure 16.7 versus 48 for the initial model in Figure 16.2. Degrees of freedom represent parameters that are constrained to some value, rather than freely estimated, and thus each additional degree of freedom means an increase in parsimony. Thus, if the two models fit equally well, we will prefer the initial (more parsimonious) model. Do the models fit equally well? To answer this question, we need to focus on the fit indexes appropriate for comparing competing models.

In Chapter 14 I argued that $\Delta\chi^2$ was a good method for comparing competing models that were nested, that is, when one model can be derived from the other by fixing one or more parameters. The two models are indeed nested; to derive the model shown in Figure 16.1 from that in Figure 16.6 we would only need to constrain the loading of Recall of Designs on the Memory factor to 0.

Table 16.2 shows the $\Delta\chi^2$ comparing these two models. According to the χ^2 , the initial model fit slightly worse than did model 2 (the model with Recall of Designs loaded on two factors). But if two models are nested, the more constrained model (the model with the larger df) will always fit worse than the less constrained model according to χ^2 . The question,

Alternative w crossloading fig 15-6
Model Specification

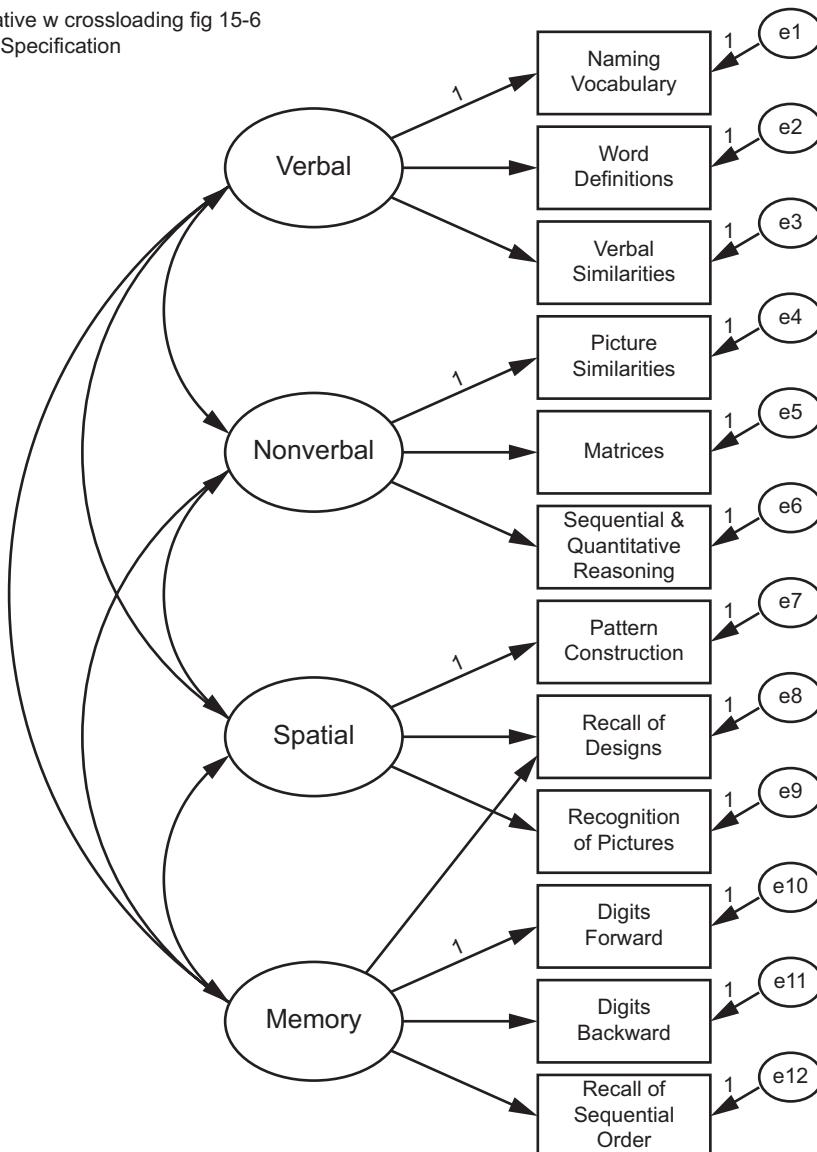


Figure 16.6 An alternative model testing whether Recall of Designs measures both visual-spatial and short-term memory skills. This model and the initial model are nested.

then, is how much worse is the fit? Is it trivial or is it large enough so that we say that it is not worth the extra degrees of freedom we gain? The common way to judge whether the fit-worsening constraint is “worth it” is to test the $\Delta\chi^2$ for statistical significance. This has also been done in the table. As shown, when the extra path/loading was added to the second model, χ^2 decreased by only .491, and this difference is not statistically significant ($p = .483$) (And recall that we need a $\Delta\chi^2$ of approximately 3.9 for statistical significance with 1 df and $p < .05$.) What does this mean? Recall also our rule that if $\Delta\chi^2$ is not statistically significant that we prefer the more constrained model, the one with more df . This means that we would tentatively accept the Initial four-factor model over the cross-loaded model, and that we

Alternative w crossloading
Standardized estimates

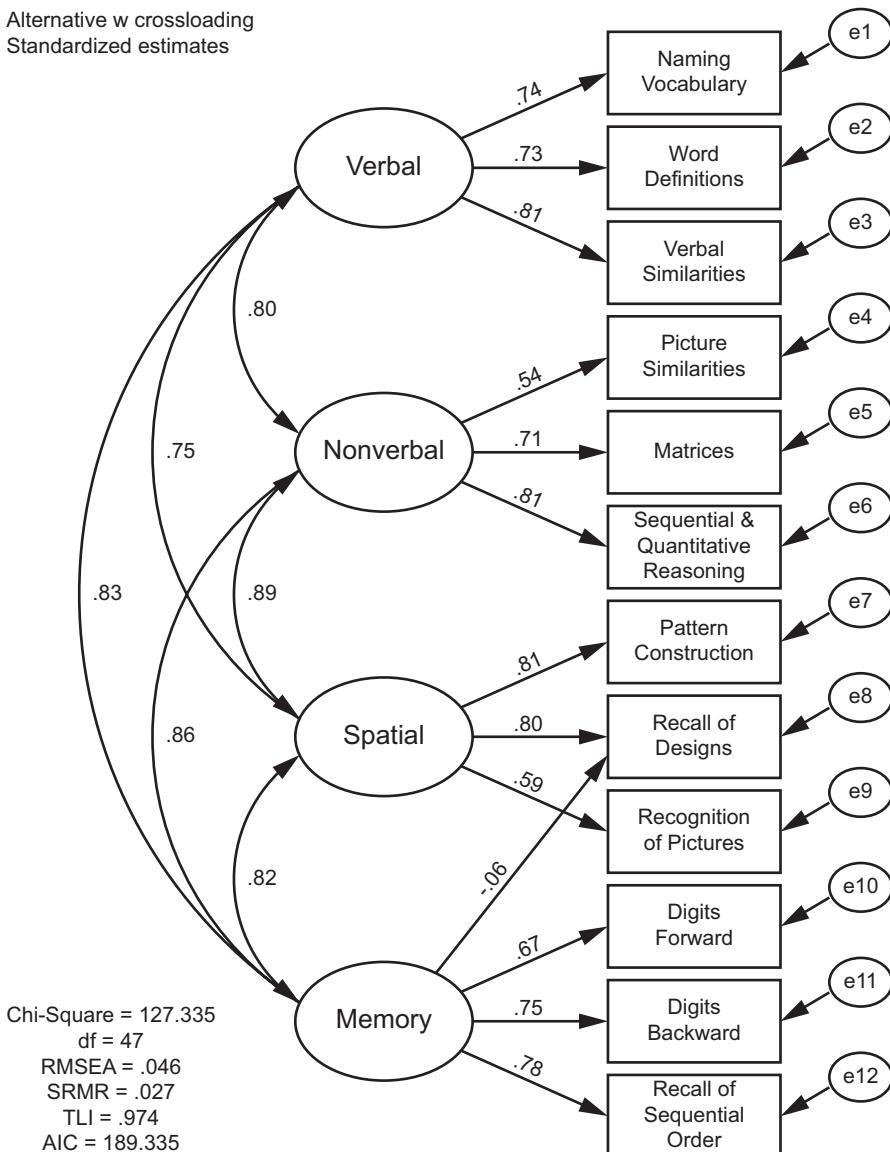


Figure 16.7 Standardized estimates and fit for the cross-loaded model

would reject the hypothesis that is personified by the difference between the two models. In other words, no, the data do not support the cross-loading of Recall of Designs on both the Spatial and the Memory factors; it appears that Recall of Designs indeed measures visual-spatial reasoning skills, not short-term memory.

A Three-Factor Combined Nonverbal Model

Although I have argued that the DAS-II should measure four underlying constructs, we have already noted the very high correlation (.89) between the Nonverbal Reasoning and the Spatial factors. Perhaps these two factors really are equivalent, meaning that we could collapse them into one? We could easily argue that the Spatial and the Nonverbal Reasoning subtests

Table 16.2 Comparison of Fit Indexes for Alternative Models of the Structure of the DAS-II

Model	χ^2	df	$\Delta\chi^2$	df	p	AIC	aBIC	RMSEA	TLI	CFI	SRMR
1. Initial four-factor	127.826	48				187.826	233.023	.046	.974	.981	.027
2. Recall Designs cross-loaded	127.335	47	.491	1	.483	189.335	236.038	.046	.974	.981	.027
3. Three-factor (Figure 16.8)	163.651	51	35.825	3	<.001	217.651	258.328	.053	.966	.974	.029
4. Nonverbal-Spatial correlation = 1	156.698	49	28.872	1	<.001	214.698	258.388	.052	.966	.975	.028
5. Equivalent correlations	163.651	51	6.953	2	.031	217.651	258.328	.053	.966	.974	.029

Note: All models are compared to Model 1 with the exception of Model 5. The $\Delta\chi^2$ for Model 5 is a comparison to the previous model (Model 4).

should be considered as measuring a single underlying ability. After all, most of these tests require some degree of spatial awareness and nonverbal reasoning; why separate the two factors? Thus, we have both a priori logical as well post hoc data-driven reasons for suggesting another plausible model, one that combines these two factors. Figure 16.8 shows such a plausible three-factor model. Although it is not obvious, the model is nested with the model in Figures 16.1 through 16.5. This three-factor model is equivalent to the model shown in Figure 16.4 (the standardized model) with the following constraints:

1. Set the Nonverbal Reasoning–Spatial correlation to 1.0 (in the standardized model). This constraint essentially equates the factors.
2. Constrain other factor correlations to be equal to one another across these factors. That is, constrain the Memory–Spatial factor correlation to be equal to the Memory–Nonverbal Reasoning correlation, and then constrain the Verbal–Spatial factor correlation to be equal to the Verbal–Nonverbal correlation. The most direct way to do this in Amos is to constrain the correlations to an alphabetical value (e.g., *a* for the first two correlations and *b* for the second two). The result of this constraint is that the values will be freely estimated, but all values with the same letter will be constrained to be equal. Other SEM programs will have other methods of constraining values to be equal.

Because the models are nested, $\Delta\chi^2$ can be used to compare the competing models. This model is more parsimonious than the initial four-factor model. (Make sure you understand why this three-factor model is more parsimonious than the initial model.) Thus, if the two models have an equivalent fit, we will favor the more parsimonious three-factor model with the combined Nonverbal factor.

As shown in Figure 16.8, the three-factor combined Nonverbal model showed a good fit to the data according to most of the stand-alone fit indexes (with the exception of RMSEA), yet the χ^2 also increased substantially for this model. The four-factor model had a χ^2 of 127.826 ($df = 48$) versus 163.651 ($df = 51$) for the three-factor Combined Nonverbal model. Change

Das 2 three-factor
Standardized estimates

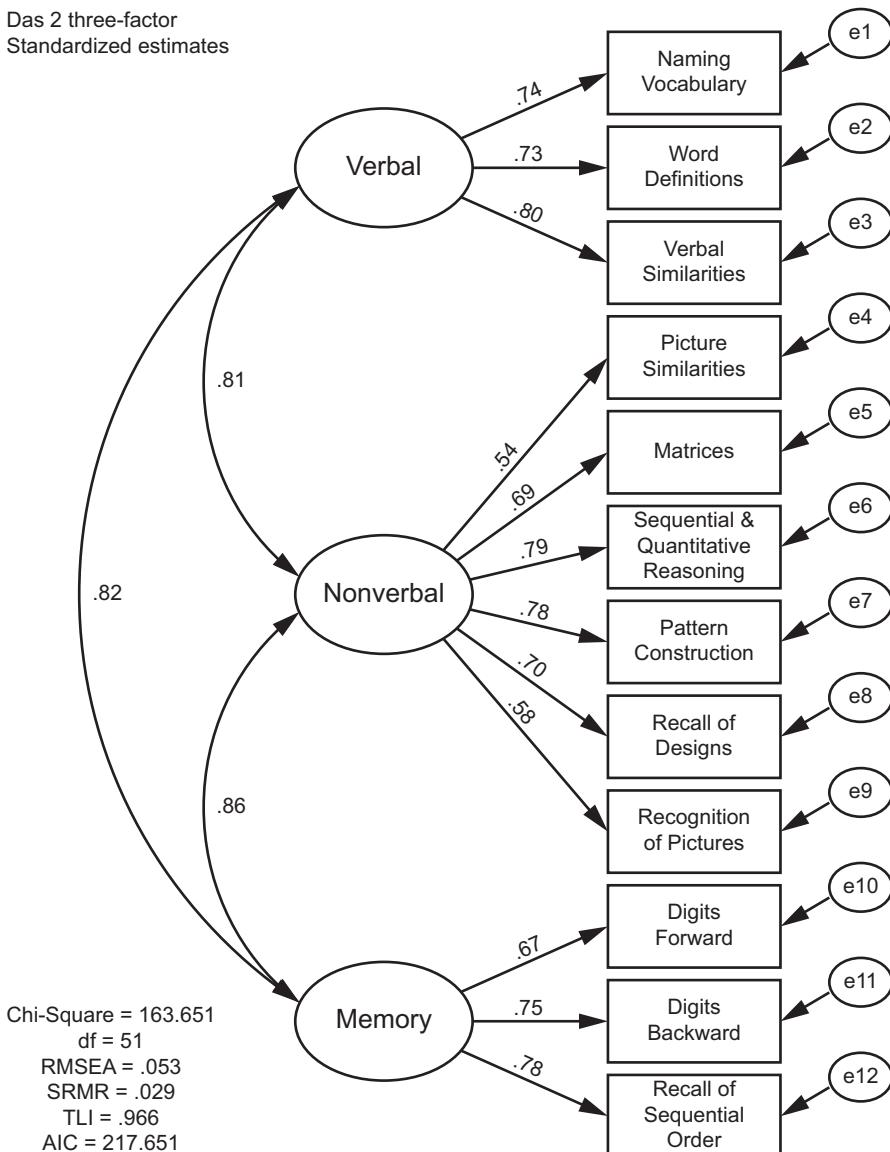


Figure 16.8 Another competing model of the DAS-II. This model combines the Nonverbal Reasoning and Spatial factors into a single Nonverbal factor.

in $\Delta\chi^2 = 35.825$ ($df = 3$), a value that is statistically significant ($p < .001$). This means that the three-factor combined Nonverbal model, although more parsimonious than the four-factor model, does not explain the relations among the DAS-II subtests, the DAS-II structure, as well as does the four-factor model. Said differently, the Nonverbal Reasoning and Spatial factors are indeed statistically distinguishable. The models shown in Figures 16.1 through 16.4 provide a better “theory” for understanding the DAS-II than does the model shown in Figure 16.8. Thus this analysis suggest that the DAS-II tests used in this analysis should be interpreted as measuring four, rather than three, underlying abilities. The fit indexes for this model are also shown in Table 16.2.

Although $\Delta\chi^2$ is our primary method for comparing competing, nested models, it is also worth noting the other fit indexes we discussed as useful for comparing (non-nested) models, the AIC and the aBIC. The rule of thumb for the AIC (and aBIC) is that they favor the model with the lower value; again the four-factor model appears superior if we use the AIC or aBIC to compare models. Again, according to our primary criteria, the four-factor model provides a better fit than does the three-factor model. Table 16.2 also includes fit indexes for the two steps outlined earlier for turning the standardized four-factor model into the three-factor model. (I will not show the analyses or models here, but I encourage you to conduct them.) In the first step, the factor correlation (standardized covariance) between the Nonverbal and Spatial factors was constrained to 1. In the second step, the Nonverbal-Verbal and the Spatial-Verbal factor correlations were constrained to be equal, as were the Nonverbal-Memory and Spatial-Memory factors. Note that the fit statistics associated with this second step are identical to those from the three-factor model as shown in Figure 16.8. As an aside, I don't believe it is necessary to conduct this analysis in two steps, but it does help to understand what is being done.

Before we move to the next topic consider why this approach is equivalent to that in which we simply combined the two factors into one. What we are testing is whether the Nonverbal and Spatial factors should really be considered as the same factor. What would be required for them to be "the same" factor? First and obviously, they should be perfectly correlated with one another. But a perfect correlation is not enough. If the Nonverbal and Spatial factors are really "the same," then they should also have the exact same relation (correlation) with other factors. The second step, constraining factor correlations to be the same value, fulfills this part of the requirement that the factors be "the same factor."

MODEL FIT AND MODEL MODIFICATION

A common response when a model does not fit well is to examine more detailed aspects of fit with an eye toward modifying the model. I won't try to dissuade you from this practice, because it is indeed useful and necessary, but I encourage you to do so sparingly, unless you are primarily involved in model development and exploration (as opposed to testing a priori models). I am not alone in this ambivalence concerning model modification: "As a statistician, I am deeply suspicious of modification indices. As a data analyst, however, I find they are really great" (Dag Sörbom, one of the creators of LISREL, quoted in Wolfe, 2003, p. 32). There are several aspects of the printout that may help in this process.

Modification Indexes

To illustrate the use of the more detailed fit indexes, let's examine the combined three-factor Nonverbal model from Figure 16.8. If we had started with this model—if we had not compared this model with the initial four-factor model (e.g., Figure 16.1), could we have figured out that the four-factor model was better? Would the modification indexes or the other detailed fit statistics have led us to what we have concluded was a better model? And are there other changes we need to make in our models?

Figure 16.9 shows the modification indexes from the Amos output for this model. With some programs, all modification indexes are printed; with Amos you can request modification indexes above a certain level. The figure shows the default, indexes greater in magnitude than 4.0 (recall that 3.9, or approximately 4, is the value of $\Delta\chi^2$ that is statistically significant with 1 *df*). When models do not fit well, you may be able to improve the fit by freeing parameters in the model. Recall that freeing a parameter reduces the degrees of freedom (parsimony) of the model and improves the $\Delta\chi^2$ to some degree. The question we ask with such

Modification Indices***Covariances:***

			M.I.	Par Change
e11	<-->	Nonverbal	5.334	2.320
e11	<-->	Verbal	5.116	-3.381
e8	<-->	Verbal	4.521	-3.288
e8	<-->	e7	21.895	7.878
e8	<-->	e9	14.549	8.641
e6	<-->	e8	5.886	-4.243
e5	<-->	e7	6.381	-4.414
e5	<-->	e12	7.027	5.542
e5	<-->	e10	7.128	-6.454
e5	<-->	e8	21.672	-9.739
e5	<-->	e6	21.896	8.493
e4	<-->	Verbal	10.847	5.969
e4	<-->	e6	4.884	-4.551
e2	<-->	Nonverbal	5.909	-2.399
e2	<-->	e7	4.741	-3.500
e2	<-->	e8	15.275	-7.488
e2	<-->	e5	4.273	4.107
e2	<-->	e4	4.568	4.787
e1	<-->	e7	6.716	4.340
e1	<-->	e11	6.197	-4.838
e1	<-->	e8	4.259	4.120
e1	<-->	e6	4.944	-3.870
e1	<-->	e5	4.929	-4.596

Regression Weights:

			M.I.	Par Change
pcss	<---	rdss	10.309	.072
dfss	<---	mass	4.597	-.064
rpss	<---	rdss	6.734	.078
rdss	<---	pcss	7.334	.079
rdss	<---	rpss	9.224	.081
rdss	<---	mass	10.490	-.085
rdss	<---	wdss	9.454	-.086
sqss	<---	mass	10.718	.075
mass	<---	rdss	10.090	-.088
mass	<---	sqss	6.989	.076
psss	<---	wdss	5.161	.074
wdss	<---	rdss	10.836	-.083

Figure 16.9 Modification indexes for the 3-factor combined model.

relaxations in the model is whether the decrease in $\Delta\chi^2$ is worth the reduction in the df . The modification indexes estimate the minimum decrease in $\Delta\chi^2$ that will result from freeing the listed parameter. Modification indexes are shown for covariances and for regression weights (the first row, for example, lists a modification index of 5.334). Although the actual output

also had a table for variances, there were no modification indexes associated with variances greater than 4.0 so the table was blank and is not included in the figure.

Note the modification index for the covariance between e5 and e6: a value of 21.896. This modification index suggests that $\Delta\chi^2$ can be reduced by at least 21.896 by freeing the covariance between e5 and e6. Although this is a statistically significant decrease in $\Delta\chi^2$ with a *df* of 1, we need to consider whether this change makes theoretical sense. The variables e5 and e6 represent the unique variances of Matrices and Sequential & Quantitative Reasoning. The column marked “Par Change” shows the expected value of this parameter (covariance) if we were to free this constraint, that is, if we were to allow these two unique variances to correlate. Note that the expected parameter change is positive (this shows the expected value of this parameter in the unstandardized solution if it, and it alone, were freed). If we were to free this covariance (correlation), it would suggest that we think the unique variances of the Matrices and Sequential & Quantitative Reasoning subtests are related above and beyond the effect of Nonverbal Reasoning on each subtest. The factors correlate with each other because they are both affected by Nonverbal Reasoning, but could they be correlated for other reasons, as well? Stated differently, do Matrices and Sequential & Quantitative Reasoning measure something in common other than the factor Nonverbal Reasoning? Given our other analyses, it is fairly easy to answer this question: yes, these two subtests likely measure a more narrow Nonverbal Reasoning factor that is separate from Spatial ability. Note also the modification index for the covariance between e7 and e8 (21.895), suggesting that we free the covariance between the unique variances of the Pattern Construction and Recall of Designs subtests. Again, given our knowledge of the four-factor solution, we can say that yes, these two tests indeed do measure something in common, a Spatial factor that is separate from the Nonverbal reasoning factor.

If we had started with this three-factor (combined Nonverbal) model and IF we were skilled in reading the modification indexes, or IF we had some knowledge of the theory underlying the DAS-II, then the modification indexes may have suggested to us to split this factor into two factors. This example also illustrates that the modification indexes are not always easy to interpret!

The other large modification index in Figure 16.9 is between e5 and e8 (21.672). This modification index suggests that freeing the covariance between the unique variance for Matrices subtest and that of the Recall of Designs would result in a $\Delta\chi^2$ of at least 21. This “suggestion” by the modification indexes would seem to be in the opposite direction from the previous ones, because we know from the four-factor model that Matrices and Recall of Designs measure separate abilities. But note also that the expected parameter change is negative. This finding, in turn, suggests that these two tests measure less in common than our three factor model would predict. Again, given our additional knowledge, this finding also suggests the possibility of placing these two subtests on separate factors. The question is whether they would have suggested this possibility if we did not have this additional knowledge!

Here are common rules of thumb for using modification indexes. Examine the larger values of the modification indexes. Note that in actual practice you may have even more modification indexes to examine than those shown for this model. What is large? Modification indexes, like χ^2 , are sample-size dependent; if our model fit much worse or if we had a larger sample size, we would have larger modification indexes and more modification indexes greater than 4.0. Thus, you should examine the larger values of the modification indexes relative to the other values. Again, the modification indexes show the expected minimum reduction in χ^2 if the listed parameter is freed, at a cost of 1 *df*. Next, consider whether each change is justifiable through theory and previous research. Make the single change that makes the most theoretical sense and results in the largest improvement in model fit, and then re-estimate the model. You can then repeat the process. Generally we don’t use the

modification indexes to make several changes at a time, because with each additional change the modification indexes are likely to differ. I remind you to use the modification indexes cautiously. You will find it is all too easy to justify model modifications *after* examining modification indexes; do so sparingly and with an eye toward theory and previous research. If you see the modification index and smack yourself in the head because you should have thought of that model change *a priori*, then the model change is probably reasonable. If you find yourself having to do mental gymnastics to justify freeing a parameter, then you probably should not.

One final note on the modifications indexes. None of the MIs for the second table (Regression Weights) were particularly large, but if they had been, and if they were between a subtest and a factor, they would have suggested the possibility of allowing for cross-loadings of tests on other factors.

Residuals

Another aspect of fit to examine to understand why a model does not fit well is the matrix of standardized residuals (Standardized Residual Covariances) shown in Table 16.3 (this matrix is also from the output for results of the model analyzed in Figure 16.8). Recall from Chapter 14 that the various fit statistics examine the consistency between the actual covariance matrix and the covariance matrix implied by the model. The difference between these two matrices is the matrix of residual covariances; the matrix of standardized residual covariances simply puts these residuals on the same standardized scale so that they can be compared. That matrix is shown in Table 16.3.

Standardized Residual Covariances

For this matrix, as well, we are looking for relatively larger values, regardless of sign. One rule of thumb suggests examining standardized residual covariances (commonly referred to as standardized residuals) greater in absolute magnitude than 2.0; but the standardized residuals are also sample-size dependent, so with larger samples you may have many values greater than 2, whereas with smaller samples there may be few or no standardized residuals that

Table 16.3 Standardized Residual Covariances for the Three-Factor Combined Nonverbal Model.

	<i>pcss</i>	<i>soss</i>	<i>dbss</i>	<i>dfss</i>	<i>rpss</i>	<i>rdss</i>	<i>sqss</i>	<i>mass</i>	<i>psss</i>	<i>vsss</i>	<i>wdss</i>	<i>nvss</i>
<i>pcss</i>	.000											
<i>soss</i>	-.548	.000										
<i>dbss</i>	.192	-.040	.000									
<i>dfss</i>	-.455	.173	-.130	.000								
<i>rpss</i>	-.226	-.383	-.111	-1.014	.000							
<i>rdss</i>	1.614	-.411	.264	.353	1.910	.000						
<i>sqss</i>	-.196	-.095	.693	-.411	-.125	-.815	.000					
<i>mass</i>	-.891	1.163	1.169	-1.123	-.238	-1.981	1.609	.000				
<i>psss</i>	-.288	-.680	.050	.322	.377	.533	-.936	-.052	.000			
<i>vsss</i>	.420	-.207	-.166	.447	-.655	-.531	.650	.387	1.291	.000		
<i>wdss</i>	-1.036	.552	-.581	.295	-1.372	-2.297	-.316	.610	1.480	.108	.000	
<i>nvss</i>	1.012	.227	-1.052	1.005	-.316	.489	-.542	-.609	1.366	-.348	.365	.000

reach this level. Again, focus on the relatively larger values; these are bolded and italicized in the table. For the present example, the combined Nonverbal DAS-II model, there is only one value greater than 2.0, between the Recall of Designs and the Word Definitions (-2.297).

What does this value mean? Recall how this matrix is created: the implied covariance matrix is subtracted from the actual covariance matrix to create the residuals. The residuals are then standardized to create this matrix. This means that for positive values the *actual* correlation between two measured variables is larger than the *implied* correlation. For negative residuals, just the opposite is the case: the implied correlation is larger than the actual correlation. This means that positive standardized residuals suggest that the model does not adequately account for the observed correlation between two variables, whereas for negative residuals the model more than accounts for the original correlation between variables. Positive residuals are thus generally more informative for purposes of model modification in that they suggest ways the model can be modified to improve the fit.

In the current example, the highest value, -2.297, is between Recall of Designs (rdss) and Word Definitions (wdss). The value is negative, which suggests that the model—in which these subtests load on the Nonverbal and Verbal factors, factors which correlate .81—more than accounts for the correlation between Word Definitions and Recall of Designs. Given the loadings of these subtests on their factors, and given the correlation between the factors, we would expect these two subtests to be more highly correlated than they are. This standardized residual thus seems to hint a different aspect of local misfit than we saw with the modification indexes, although it is not clear what this means or if there is anything we should do about it.

The other larger standardized residuals (those with values greater than 1.5 are highlighted) tell the same story as did the modification indexes. There are high positive values for Recall of Designs with Pattern Construction and with Recognition of Pictures. The model does not adequately explain the correlations between Recall of Designs and the other two Spatial tests. Likewise, the model does not adequately explain the correlation between the Matrices and the Sequential & Quantitative Reasoning subtests (both measures of Nonverbal Reasoning), but more than accounts for the correlation between Matrices and Recall of Designs (which, in the four-factor model measure two different underlying abilities). Again, if we were skilled and theoretically savvy, we might have taken these as hints that these six subtests should be split into two factors rather than loaded all on one. Or maybe not. The other thing the pattern of higher loadings suggests is that the Recall of Designs subtest is a general source of misfit in this model.

Residual Correlations

Table 16.4 shows a related but potentially useful matrix, the matrix of residual correlations. As noted in chapter 14, this matrix shows the residuals for the actual and implied correlation matrices. The downside is that many SEM programs do not produce this matrix (Amos does not, as least as of this writing). But the matrix is easy to produce; I simply copied and pasted the sample correlation matrix and the matrix implied by the model into Excel and subtracted the latter from the former. Again I have highlighted the higher values in this matrix (here, values greater than .06 in absolute value). Note that the subtests highlighted the same as in the previous table, which should always be the case. This table, however, shows differences in correlations, so the values are readily interpretable. The value for Recall of Designs and Word Definitions (-.088) means the actual correlation between these two subtests is .088 lower than that predicted by the model. If you focus on the model (Figure 16.8), it shows Word Definitions with a standardized loading of .73 on the Verbal factor and Recall of Designs with a loading of .70 on the Nonverbal factor, with these factors correlating .81 with each other. The expected or implied correlation between these two subtests would thus

Table 16.4 Residual Correlations for the Three-Factor Combined Nonverbal Model.

	<i>pcss</i>	<i>soss</i>	<i>dbss</i>	<i>dfss</i>	<i>rpss</i>	<i>rdss</i>	<i>sqss</i>	<i>mass</i>	<i>psss</i>	<i>vsss</i>	<i>wdss</i>	<i>nvss</i>
<i>pcss</i>	0											
<i>soss</i>	-.022	0										
<i>dbss</i>	.008	-.002	0									
<i>dfss</i>	-.018	.007	-.005	0								
<i>rpss</i>	-.009	-.015	-.004	-.038	0							
<i>rdss</i>	.065	-.016	.010	.014	.073	0						
<i>sqss</i>	-.008	-.004	.028	-.016	-.005	-.033	0					
<i>mass</i>	-.036	.045	.045	-.043	-.009	-.078	.065	0				
<i>psss</i>	-.011	-.026	.002	.012	.014	.020	-.036	-.002	0			
<i>vsss</i>	.017	-.008	-.007	.017	-.025	-.021	.026	.015	.048	0		
<i>wdss</i>	-.040	.022	-.023	.011	-.051	-.088	-.012	.023	.055	.004	0	
<i>nvss</i>	.040	.009	-.041	.038	-.012	.019	-.021	-.023	.051	-.014	.015	0

be $.73 \times .70 \times .81$, or .41 using the tracing rule. In fact, the actual correlation between these two subtests was .32, a difference of -.09 (rounded). Again, this model predicts a higher correlation between these two subtests than was found in the actual data.

For both the standardized residuals and the residual correlations, consider whether the larger positive values share some characteristic in common (you can do the same for the larger negative values, which may suggest additional constraints to the model). Although the residuals are somewhat more difficult to interpret than the modification indexes, they also sometimes show a pattern, and thus may be very useful in suggesting additional paths, correlations, or even minor factors to add to a model.

The residual correlations should highlight the same sources of misfit as the standardized residuals. The advantage is that these residuals are on a scale with which we are familiar, that of a correlation coefficient. As a result, we can devise informal rules of thumb for problematic values. Kline, for example, suggests “correlation residuals” greater than .10 as potentially problematic (2016, p. 240). Kline also suggests examining the residual correlations whenever the χ^2 for the model is statistically significant; I would simply add that this is a good idea when any of the fit indexes suggest a lack of fit.

Adding Model Constraints and *z* Values

You can modify a model by relaxing constraints to the model (estimating a parameter that was previously set to zero), as discussed previous. Model relaxations will always improve χ^2 , but will make the model less parsimonious. Sometimes the relaxation of constraints is worth the improvement in fit. Another direction in modifying models is to add constraints, generally by constraining a previously estimated value to zero (or some other value). If, for example, some of the factor loadings had been statistically not significant according to the critical ranges (*z* values), we might have constrained these values to zero (i.e., removed the path) in subsequent models. Adding constraints to the model will always lead to a larger (worse) χ^2 , but a more parsimonious model. If the $\Delta\chi^2$ is not statistically significant, the constraint makes sense. These same rules apply to many other fit indexes, as well: relaxations will improve fit,

constraints will degrade fit. The exception to this rule is with fit indexes that take model parsimony into account; these indexes may improve with constraints and degrade with relaxations. Of the indexes we have discussed, the TLI, RMSEA, aBIC, and (commonly) the AIC also take parsimony into account. Indeed, the AIC and related fit indexes (e.g., BIC) are designed to prevent “overfitting,” or making small, sample-specific changes solely to improve fit.

Cautions

I again encourage you to be cautious when making model modifications. Extensive model modifications will take you far afield from the supposedly confirmatory, theory-testing nature of SEM and CFA and can even lead to erroneous models (MacCallum, 1986). Some authors make the useful distinction between the use of SEM and CFA in a theory-testing versus a more exploratory matter (Jöreskog & Sörbom, 1993). I believe this is a useful distinction, and encourage you to know where you are along this continuum. If you make more than minor changes to your model, you should not think of what you are doing as theory testing unless you have retested the model with new data.

HIERARCHICAL MODELS

Higher-Order Model Justification and Setup

The analyses so far have pointed to the model in Figure 16.1 as a more valid representation of the structure of the DAS-II than the models in Figures 16.7 and 16.8. But the model in 16.1 is not complete, either. In addition to measuring the four abilities shown in Figure 16.1, the DAS-II is also designed to measure overall general intelligence. The model shown in Figure 16.10, then, is probably a more accurate reflection of the intended structure of the DAS-II: rather than simply having the first-order factors correlated, these factors are shown as reflections of second-, or higher-order factor, general intelligence, usually symbolized as g , in a hierarchical model. Note that this type of hierarchical model (with higher-order factors) is generally referred to as a higher-order model; another type of hierarchical model (the bifactor model) is also discussed (also see Keith & Reynolds, 2018 or Reynolds & Keith, 2013).

There are several reasons for developing and estimating higher-order models. In the arena of intelligence, higher-order models are more consistent with commonly accepted theories of intelligence (e.g., three-stratum or Cattell–Horn–Carroll theory, Carroll, 1993) than are first-order models and are more consistent with the actual structure of most intelligence tests. Higher-order and other hierarchical models may be equally relevant in many other areas of research. Higher-order models can also lead to a better understanding of the first level of factors. Just as the first level of factors helped us understand what the subtests measured, the second-order factor(s) may help us better understand the first-order factors.

The mechanics of estimating a higher-order CFA also need comment. Note that the scale of the second-order factor (g) is set in the same way as the first-order factors, by fixing one path from it to one of the first-order factors to 1.0. We could also set the scale by fixing the variance of g to 1.0 (in this case we would still need to set the scale of the first-order factors by setting a path to 1.0, and thus the first-order factor solution will not be standardized). The higher-order model differs from the first-order model in that the first-order factors have small latent variables pointing toward them, labeled $uf1$ through $uf4$ (for unique factor variance). These latent variables have the same essential meaning as other disturbances/residuals: they represent all influences on the first-order factors (Verbal, Nonverbal Reasoning, etc.) other than g . To put it another way, *any variable—whether measured or latent—that has an arrow pointing to it must also include a latent disturbance/unique variable to represent all other influences on the variable*. Finally, the model shown here includes three levels—measured

Das 2 hier no fit
Model Specification

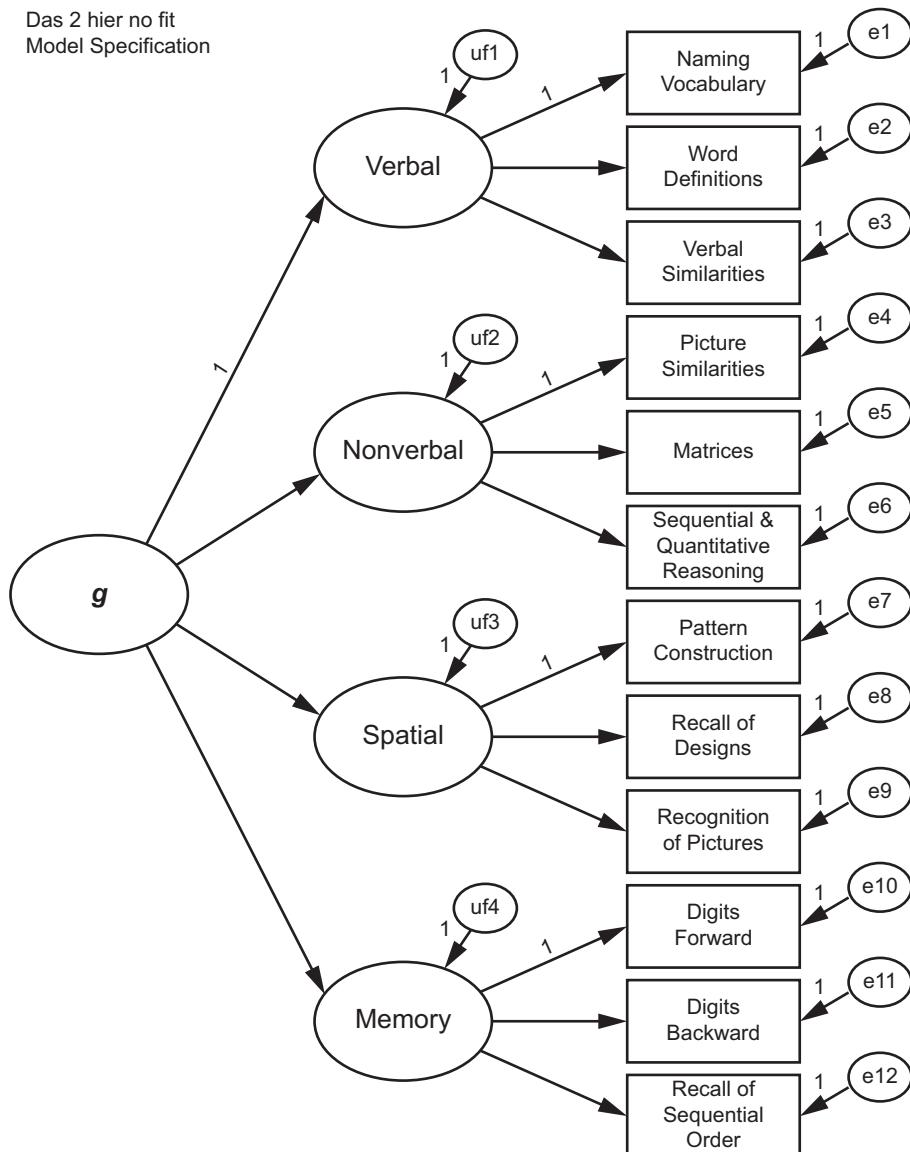


Figure 16.10 Higher-order model of the DAS-II. The model specifies that the DAS-II measures general intelligence in addition to the four broad cognitive ability factors.

variables, first-order factors, and a second-order factor—but additional levels are possible and are capable of estimation using these same methods.

Higher-Order Model Results

Figure 16.11 shows the fit statistics and standardized estimates for the higher-order analysis. Note that the first-order factor loadings are the same as they were for the initial, first-order analysis (Figure 16.2; these will not always be identical but should be very similar). The equivalence is because the essential difference between the higher-order and the first-order model is that the higher-order model explains the correlations (covariances) among

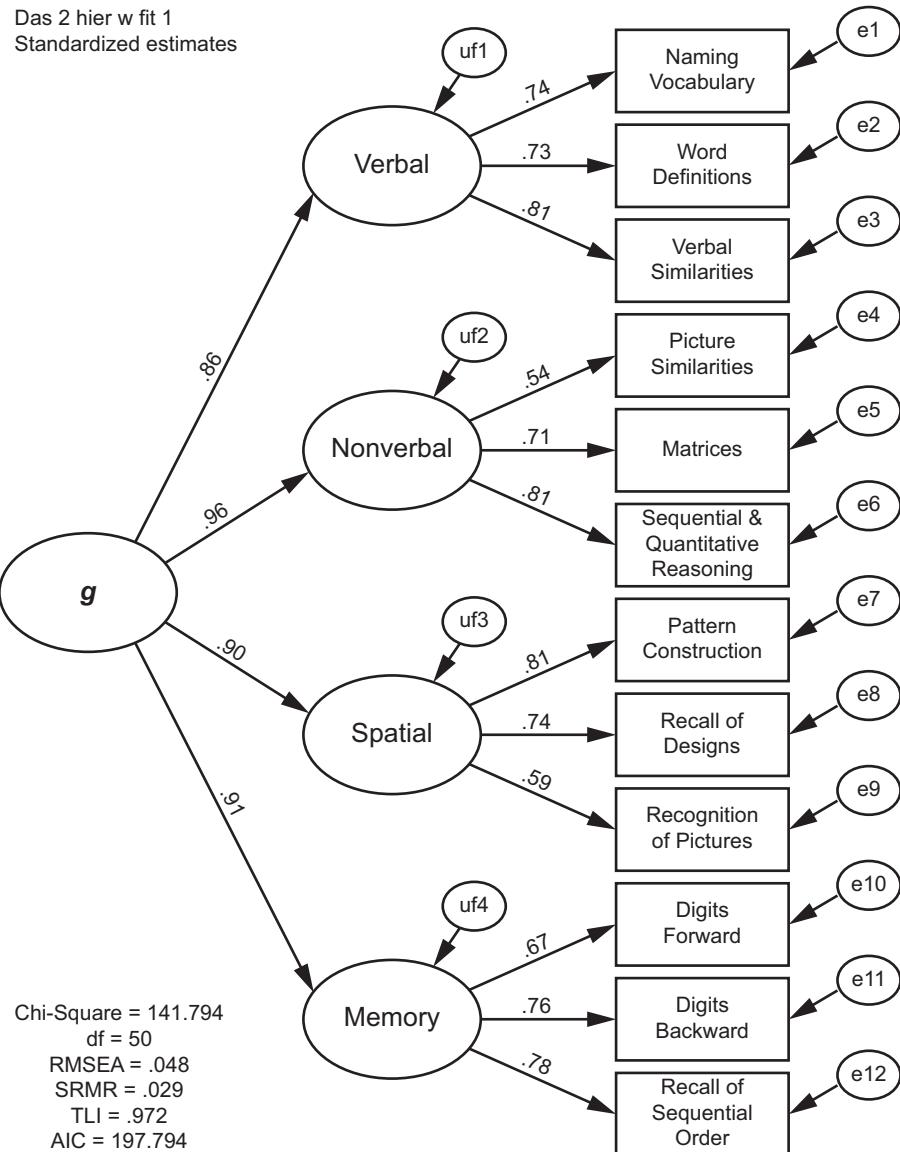


Figure 16.11 Standardized estimates for the higher-order DAS-II model.

the first-order factors with a specific structure. The first-order factor model helps explain why the *subtests* correlate with each other; because there are four abilities that partially cause students to perform at a certain level on the eight subtests. The second-order model adds to that a possible explanation of the reason for the correlations among the four *factors*: because there is one general intellectual ability factor that influences, in part, the four more narrow abilities. Conceptually, the factor analysis of latent variables (second-order) is equivalent to the factor analysis of measured variables (first-order).

The fit of the model looks good; with the exception of the statistically significant χ^2 , all indexes suggest a good fit of the model to the data. An examination of the modification indexes, standardized residuals, and correlation residuals show similar results as for the initial four-factor model, and suggest no major problems.

Given that the higher-order model is the same as the initial four-factor model with the addition of paths explaining the correlations among factors, it should be clear, then, that the higher-order model may be considered a more constrained, more parsimonious version of the first-order model (Rindskopf & Rose, 1988). The first-order model places no constraints on the factor correlations, whereas the higher-order model says that these correlations are the product (in this case) of another latent variable, g . Given this similarity, you may consider the two models as nested, and thus we could use $\Delta\chi^2$ to compare the two models. If we were to do so, $\Delta\chi^2 = 13.968 [2], p = .001$, and we would likely reject the higher-order model as not worth the increase in χ^2 compared to gain in parsimony. Likewise, the AIC and aBIC are worse for the higher-order model compared to the four-factor first-order model. I generally don't compare first-order with higher-order models in this way (as nested models), however. It seems to me that at least in the area of intelligence, such models are justified on purely theoretical grounds, without reliance on fit indexes to compare them to agnostic, non-higher-order models. In addition, theorists recognize the likelihood of there being intermediate factors between the first-order factors and g , (Carroll, 1993, chap. 16) and such factors, if accurate, would improve the fit of the higher-order models. With the DAS-II, for example, allowing the unique variances of the Nonverbal and Spatial factors to correlate would lead to a higher-order model that fit as well as the first-order model (Keith et al., 2010). Allowing this "correlated error" is statistically equivalent to specifying that Nonverbal Reasoning and Visual-Spatial skills are reflections of an intermediate factor between them and g (can you figure out why this would be the case?). Or perhaps I just have a soft spot in my heart for higher-order models of intelligence.

Let's be sure we understand where these two extra degrees of freedom come from as you look over the model. For the first-order model, there were six covariances among the first-order factors and four first-order factor variances. This is calculated as $\frac{p \times (p+1)}{2}$ "moments" in the variance/covariance matrix where p = the number variables, in this case, first-order factors, and thus $\frac{4 \times (5+1)}{2} = 10$. The higher-order model uses up eight of these free parameters to estimate three of the second-order to first-order factor loadings (recall that one path was set to 1), along with the variance of the g factor and the variances of the new disturbances (uf1 through uf4), leaving two extra df . This means that if there are only three first-order factors the higher-order portion of the model will be just-identified; the two models will then have identical fit and cannot be compared statistically. If we try to add a higher-order factor to a model with only two first-order factors, the higher-order portion of the model will be underidentified and estimation will be impossible unless we make additional constraints (e.g., constraining the two second-order loadings to be the same). You need to pay attention to the identification status of the higher-order portions of such models (identification was discussed in Chapter 12).

Recall that one reason for investigating higher-order models is to help understand the first-order factors. Indeed, the second-order factor loadings are interesting. The highest loading (near 1.0, unity) was by the Nonverbal Reasoning factor. Nonverbal Reasoning thus appears to be the most intellectually laden of the first-order factors. This finding suggests that the deductive and inductive reasoning that underlies the tasks on this factor is close to the essence of general intelligence.

Total Effects

Psychometric researchers are often also interested in understanding which of the subtests are most highly related to the global general intelligence factor. We can calculate these loadings of the subtests on the second-order factor by multiplying paths (e.g., the loading of Word Definitions on g would equal $.86 \times .73 = .63$). If this process sounds familiar, it should; we are simply calculating the indirect effect of g on each subtest. Because there are no direct effects in this model (all the effects from g are mediated by the first-order factors), these indirect effects are also the total effects. Figure 16.12 shows the total standardized effects of g

Standardized Total Effects, Higher-Order Model

	g	Memory	Spatial	Nonverbal	Verbal
Memory	.913	.000	.000	.000	.000
Spatial	.903	.000	.000	.000	.000
Nonverbal	.955	.000	.000	.000	.000
Verbal	.858	.000	.000	.000	.000
pcss	.736	.000	.815	.000	.000
soss	.709	.776	.000	.000	.000
dbss	.693	.758	.000	.000	.000
dfss	.608	.665	.000	.000	.000
rpss	.531	.000	.588	.000	.000
rdss	.667	.000	.738	.000	.000
sqss	.770	.000	.000	.806	.000
mass	.682	.000	.000	.714	.000
psss	.516	.000	.000	.540	.000
vsss	.692	.000	.000	.000	.807
wdss	.626	.000	.000	.000	.730
nvss	.635	.000	.000	.000	.740

Figure 16.12 Standardized total effects for the higher-order model. The bolded coefficients are the total effects of *g* on the subtests. These may also be considered the loading of the subtests on the higher-order *g* factor.

on the first-order factors and subtests (for some reason, the order of subtests in Figure 16.12 is almost reversed from the order of the subtests in Figure 16.11). The total effects from *g* to subtests are shown in boldface. As shown in the figure, the Sequential and Quantitative Reasoning subtest had the highest total effect from *g* (.770). Thus, this subtest is most closely related to *g*, or *g* has a stronger effect on this subtest than on any of the other subtests.

Bifactor Model Justification and Setup

There is another type of hierarchical model, often known as the bifactor model. You may see this model referred to by other names, as well, including the nested-factors or direct hierarchical model. A bifactor version of the DAS-II is shown in Figure 16.13. This model, like the higher-order one, includes both Verbal, Nonverbal, and the other first-order factors, and it also includes a more general factor, here symbolized as *G*. With the bifactor model, however, both the narrow and the general factor are first-order factors, whereas in the higher-order model the general factor is a higher-order one designed to explain the correlations/covariances among the first-order factors. Because the general factor in a bifactor model is also a first-order factor, it is often symbolized in intelligence models as *G* as opposed to *g* (used for a second- or higher-order factor).

Note several other aspects of the bifactor model. First, note that the more narrow factors (the “broad abilities” in intelligence lingo) are often specified as uncorrelated with one another, and as uncorrelated with the general factor. This is done because if we allowed all of these to be correlated with one another the model would be underidentified and thus we could not estimate it. Because models imply theories, the bifactor model thus says that the broad abilities and *G* are unrelated to one another. This model also says that each of the DAS-II tests measures two things: a general ability shared by all the DAS-II tests, and one of four other underlying constructs. Note also that the scales of both the broad abilities and *G* were set using ULI (unit loading identification). It is also possible to use UVI to set the scale for the broad abilities, or for *G*, or for both.

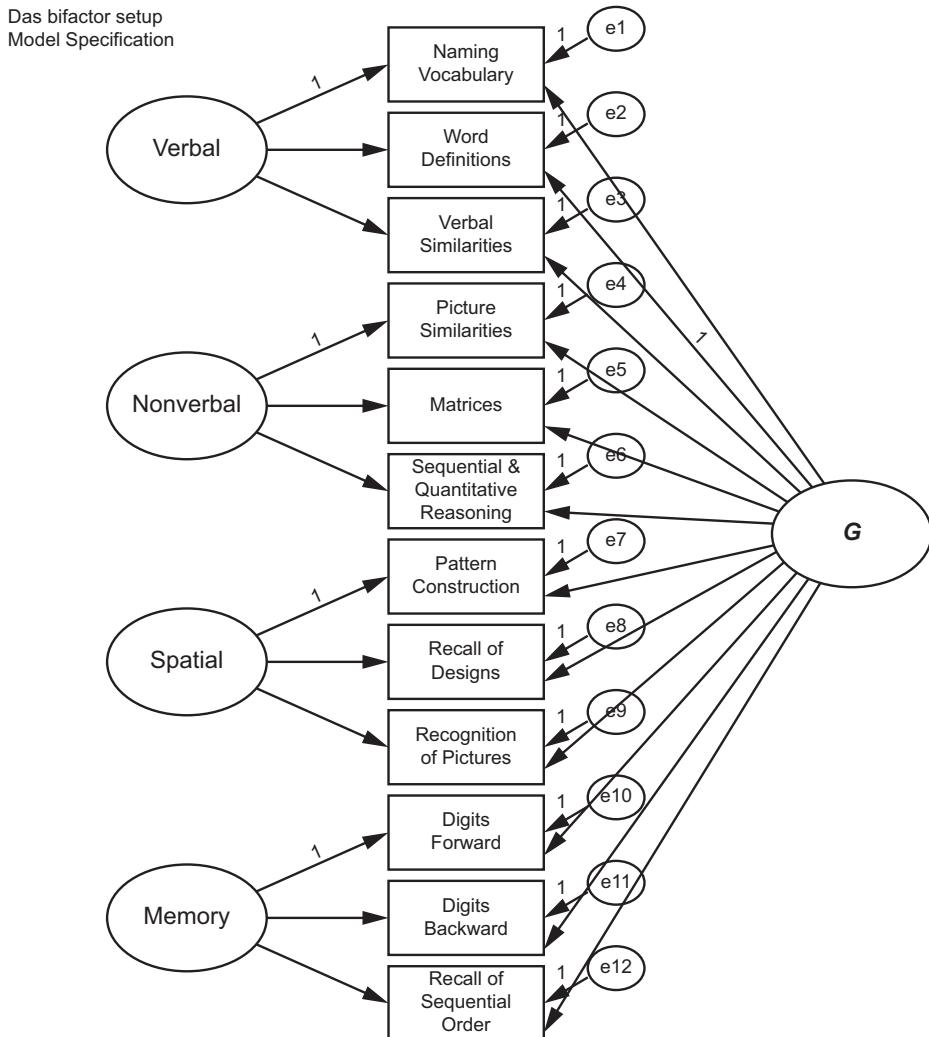


Figure 16.13 A bifactor hierarchical model for the DAS-II. This model has both *G* and the broad abilities as first-order factors.

Bifactor Model Results

The initial analysis of the bifactor model returned the error message that the variance associated with *e*6 was negative, as shown in Figure 16.14. Variances, which are squared terms (one way of thinking of them is that they are the standard deviation squared), cannot be negative, so the model would not run. This problem is common enough in factor analysis that it has a name, a “Heywood case.” In CFA, a Heywood case generally shows up as a negative error variance resulting from the path or paths to a variable explaining 100% or more than 100% of its variance. In the present example, the Nonverbal and *G* factors, together, explain more than 100% of the variance in the Sequential and Quantitative Reasoning test. It is worth noting that Heywood cases are not unique to bifactor models; they also show up in higher-order models (one should always check the first-order residual/disturbance variances carefully in higher-order models), and even in first-order models. One common method of dealing with a negative variance is to set the offending value to zero.¹

The following variances are negative. (Group number 1 - initial model)

	e6
	-35.264

Figure 16.14 Error message for the initial bifactor model. Variances cannot be negative.

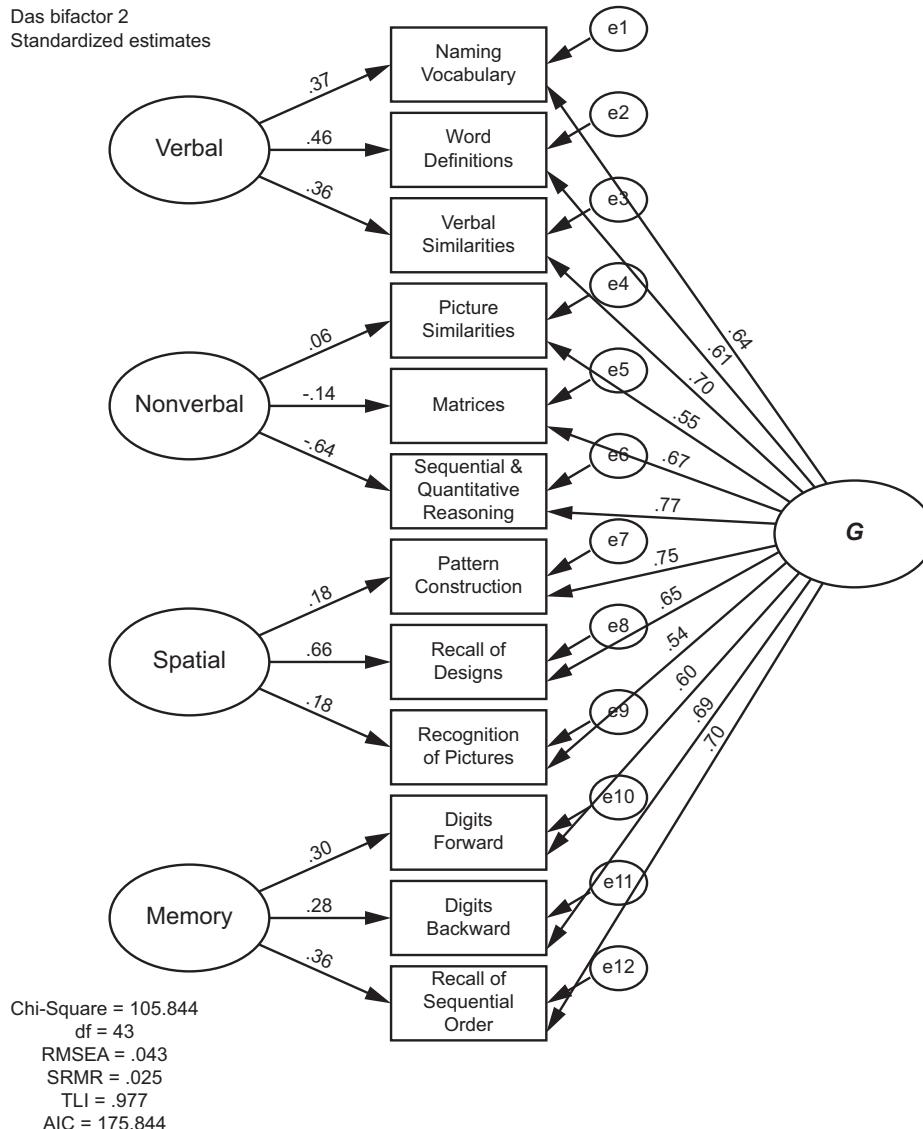


Figure 16.15 Standardized bifactor model results. The variance associated with residual e6 was set to zero to allow estimation. See the text for the explanation of the negative factor loadings for the Nonverbal broad ability.

Figure 16.15 shows the standardized results for the bifactor analysis with the error variance for Sequential and Quantitative Reasoning (e6) constrained to zero. As shown in the figure, the model fit the data well, with RMSEA = .043, SRMR = .025, and TLI = .977. Indeed, the bifactor model fit better than did the higher-order model (AIC = 175.844 versus 197.794 for the higher-order DAS-II model). We will return to this issue of fit momentarily.

Beyond fit, you probably noticed a few curious aspects of the model results, like the negative loadings of two of the tests on the Nonverbal factor. These exist because the Picture Similarities test was chosen as the reference variable for the unit loading identification. If the loading from Nonverbal to Matrices or to Sequential and Quantitative Reasoning (SQR) had instead been set to one, the Picture Similarities tests would have shown a small negative loading on the Nonverbal factor ($-.06$) and the Matrices and SQR loadings would have been positive (.14 and .64, respectively).

The loadings of each test on G were large and statistically significant. Note also how similar these values are to those shown as the total effects of g on the subtests for the higher-order model (Figure 16.12). Although the rank order changes slightly, the subtests that were the best measures of g for the higher-order model are also the best measures of G for the bifactor model, and the worst for one are also the worst for the other. In contrast, note how much lower are the loadings for the subtests on the four broad factors in the bifactor model compared to all previous models. Indeed, although not shown in the figure (but would be in the detailed output), some of these paths/factor loadings are not statistically significant. Why, you may wonder? The short answer is that the broad abilities have a different meaning in one model versus the other. The first-order loadings in the higher order model show the effect of these abilities on the subtests. In the bifactor model, the broad abilities represent the effect of these abilities on the subtests, *controlling for G*. In other words, with the bifactor model, these are the unique effects of the broad abilities, or the effects of the broad abilities with G statistically removed. Both types of effects—with and without the effects of G removed—may be of interest, but they do have different meanings and interpretations.

These two models also imply quite different theories about the nature of intelligence. The higher-order model says that the primary reason that the 12 tests shown correlate with one another is that they measure four underlying cognitive abilities. g , in turn, affects these broad cognitive abilities, and g affects the specific tests only indirectly. The bifactor model, in contrast, says that there are two reasons for the correlations among these 12 tests: first, they all measure G , and second, they all measure some other broad cognitive abilities that are independent from one another. In the bifactor model, G has direct effects on the specific tests. In the higher-order model, then, g can be understood by the nature of the broad cognitive abilities that underlie it, and those cognitive abilities can be understood as more or less related to g . For the bifactor model, the nature of G can be referenced to specific tests.

Comparing the Hierarchical Models

It is possible to obtain similar (smaller) loadings for the higher-order models as for the bifactor model, and doing so also aids in understanding their differences (or similarities). One way of doing so is illustrated in Figure 16.16. For the previous higher-order models I specified that the paths from the disturbances for the first-order factors were equal to 1 and that the unique factor variances were estimated (i.e., ULI specification for uf1 through uf4). In Figure 16.16, in contrast, UVI was used for identification of the disturbances. With this setup it is possible to calculate the indirect effect of uf1 through uf4 on the various DAS-II subtests. These indirect effects are generally quite similar to those for the loading of the subtests on the broad abilities in the bifactor model. Consider what this means. uf1 through uf4 represent all other influences on the broad abilities, once g is taken into account. These indirect effects, then, represent the unique effects of the broad abilities on the subtests, *once g is removed*. Thus if it is of interest it is possible to obtain estimates of the effects of the broad abilities on the subtests with g removed. Such estimates may indeed be of interest, and are in fact equivalent to the Schmid-Leiman transformation that is a popular method for interpreting higher-order exploratory factor results.²

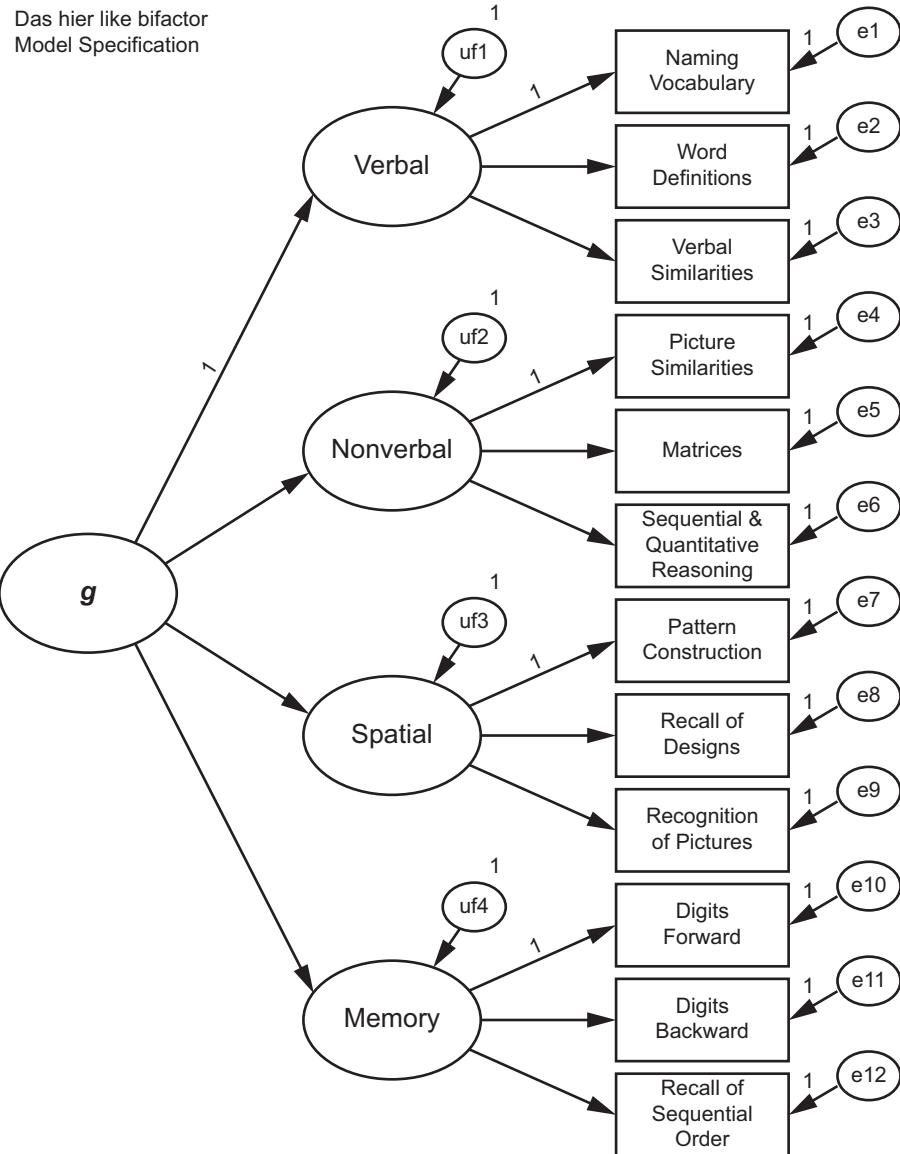


Figure 16.16 Model setup allowing the comparison of the broad ability loadings from the higher-order to the bifactor models.

Although it is common to treat the bifactor and higher-order models as non-nested (as we have done here), it is possible to go from the bifactor model to a model that is equivalent to the higher-order model. Note the df difference between the two models. Before the $e6$ variance was constrained to zero, the bifactor model had 42 df versus 50 for the higher-order model, for a Δdf of 8. If we were to allow factor loadings in the higher-order model for 8 subtests directly on g —two per each of the four broad abilities—the model would be equivalent to, and have the same fit as the bifactor model. (For the present example we would also need to once again constrain the $e6$ variance to zero.) In other words we can add a path from g to Word Definitions and Verbal Similarities, to Matrices and Sequential & Quantitative Reasoning, and so on, and obtain an equivalent (although likely uninterpretable) model.

Alternatively, note that because g affects the subtests only through the broad abilities in the higher-order model, this places constraints on the relative loadings on the subtests on g . It would be possible, then, to go from the bifactor to the higher-order model by adding proportionality constraints (a topic beyond the scope of this text) (Yung, Thissen, & McLeod, 1999). What is important to realize at this stage of understanding is that higher-order model is equivalent to a more constrained version of the bifactor model. Thus, the bifactor model will generally fit as well or better (using χ^2) than the more constrained higher-order model.

The bifactor model is popular these days (my colleague Tiffany Whitaker calls it the “little black dress” of CFA), and it does indeed have some advantages over a higher-order model (Chen, West, & Sousa, 2006; Reise, 2012). Chief among these is that it fits as well or better than does a higher-order model (based on $\Delta\chi^2$; see the previous example). One could consider its lack of specification of a relation between G and the broad factors as agnostic (not sure how they are related) rather than well-defined (it actually specifies that they are unrelated). With this change in thinking, the bifactor model would seem to be a good choice for a hierarchical model when there is no theory specifying how general and broad factors are related, or when that theory is undefined on this point. I will note that this is not the case in the area of intelligence, but it may be the case in many other areas where hierarchical CFA is of interest.

The bifactor model also has some disadvantages. As should be obvious by now, models imply theories, and the model you choose should be consistent with the theory you wish to test. Although some researchers treat the higher-order and the bifactor models as interchangeable, even our cursory explanation shows that they imply different theories. If one of these models is more consistent with the theory you wish to test, then that is the one you should use. If one theory says the structure of your construct of interest is one way (e.g., a bifactor-type model), and another theory says it is another way (e.g., a higher-order-type model), then you should compare the two (with knowledge that the bifactor model will fit as well or better based on $\Delta\chi^2$; it may or may not fit better with fit indexes that take parsimony into account). Such comparisons should make reference to the underlying theory being tested.

Another problem with the bifactor model is that it is not always easy to estimate, and the results can be quirky. With the present example you saw that we had to make an additional constraint to one error variance in order for the model to work. The fact that this model specifies that each measure is a reflection of two underlying factors sometimes leads to problems with estimation and convergence of the factor solution. As a result, it is not unusual to have to specify “start values,” or initial guesses of what parameters might be (this is easy to do in most SEM programs, although you may have to do some digging to find out how). If the broad ability factors are uncorrelated, they must be referenced by three or more measures or the model will be underidentified.² Sometimes you may get different results for the bifactor model depending on how you go about estimating your model. In the present example, when I constrained the second test on each factor as the reference variable (i.e., Matrices loading set to 1 instead of Picture Similarities), the various standardized loadings showed the same magnitude but a different pattern of nonsignificance (e.g., the unstandardized SQR on Nonverbal loading was nonsignificant for the initial analysis but statistically significant for this one). When a UVI specification was used (factor variances set to 1), there were many fewer nonsignificant factor loadings. Finally, when I analyzed the initial model in Mplus, it suggested a negative variance for e5, whereas Amos suggested a negative variance for e6. All these differences are likely related to the fact that the variances for some of broad abilities were quite small and, depending on estimation method, nonsignificant. But whatever the reason, finding such differences is disconcerting (cf. Millsap, 2001). Such anomalies happen with the higher-order model as well, but in my experience they are more common with a bifactor as opposed to a higher-order factor model.

A final disadvantage of the bifactor model is that it may lend support to an incorrect model (Maydeu-Olivares & Coffman, 2006; Murray & Johnson, 2013). Simulation studies show that, for example, the bifactor model may fit the data better than a higher-order model,

even when a higher-order model is the correct model (Murray & Johnson, 2013). Bifactor models may also fit random and unlikely data (Bonifay & Cai, 2016; Mansolf & Reise, 2017; Reise, Kim, Mansolf, & Widaman, 2016). For these reasons, although possible, $\Delta\chi^2$ comparisons between higher-order and bifactor models may not be useful (Mansolf & Reise, 2017).

My current take on the bifactor model, as compared to a higher-order model, is that the bifactor model may indeed be a useful model when one is agnostic or unclear about how the most general factor should relate to the more specific factors. Likewise, if one believes that the structure of the underlying data conform to something like a bifactor model, then it should be used in those cases as well. When the guiding theory specifies a higher-order relation between the most general and more specific factors, however, the bifactor model results may be misleading. I also think it is useful to interpret the two models in combination, and note similarities and differences in findings (see Reynolds & Keith, 2017 for an intelligence example). I am not sure if these tentative conclusions will be supported five years from now, however. Despite the long history of the bifactor model, we are still learning about it! For more detailed comparisons of the two models see some of the references already listed (Chen, West, & Sousa, 2006; Mansolf & Reise, 2017; Murray & Johnson, 2013; Reise, 2012). My colleagues and I have compared the two models with intelligence data (e.g., Keith & Reynolds, 2018 and Reynolds & Keith, 2013, 2017), and Mansolf and Reise (2017), Mulaik and Quartetti (1997), and Yung and colleagues (1999) show some important statistical comparisons.

ADDITIONAL USES OF MODEL CONSTRAINTS

Occasionally, it is useful to be able to specify single-indicator factors. This may seem impossible, given that we earlier noted that we needed to have multiple measures of each construct to have a latent variable model. As you will see, with single indicators the portion of the measurement model is underidentified, but there are ways of working around this problem.

Pretend for this example that the DAS-II only included a single measure of short-term memory skills, the Digits Forward subtest. Is there some way we could model a Memory factor despite this weakness in the data? There are several ways we could do so. One method is shown in Figure 16.17, which shows a Memory factor with a single indicator, Digits Forward. This sort of model is more difficult to estimate because, without further constraints, this portion of the model is underidentified. We can work around this problem of estimating a single-indicator latent variable in SEM (and CFA) by fixing the value of the unique-error variance to some value; this brings this portion of the model back into a just-identified state.

We could, of course, constrain the value of the unique/error variance to zero. This approach tacitly suggests that we believe the measured variable is measured without error, that the measured variable and the factor are exactly the same. Whether we realized it or not, this is what we were doing when we were analyzing path models (and when we were doing multiple regression): we assumed that a single measure was a perfectly valid and reliable indicator of the constructs we were interested in.

Another approach is to use information about the estimated *reliability* of the measured variable in the model, if we know it or can estimate it. One minus the reliability provides an estimate of the proportion of error in the measured variable; if this value is multiplied by the variance of the variable, the result is the variance in the measured variable that can be attributed to error. Figure 16.17 shows a model that uses this methodology. The estimated (internal consistency) reliability for the Digits Forward test, across ages 5–8, is .91 (Elliott, 2007), and the variance of Digits Forward for the present sample is 121.523 (from the variance/covariance matrix). The estimate used for the error variance for Speed of Processing (u9) is thus 10.94:

$$V_e = (1 - r_{tt})V = (1 - .91) \times 121.523 = 10.937.$$

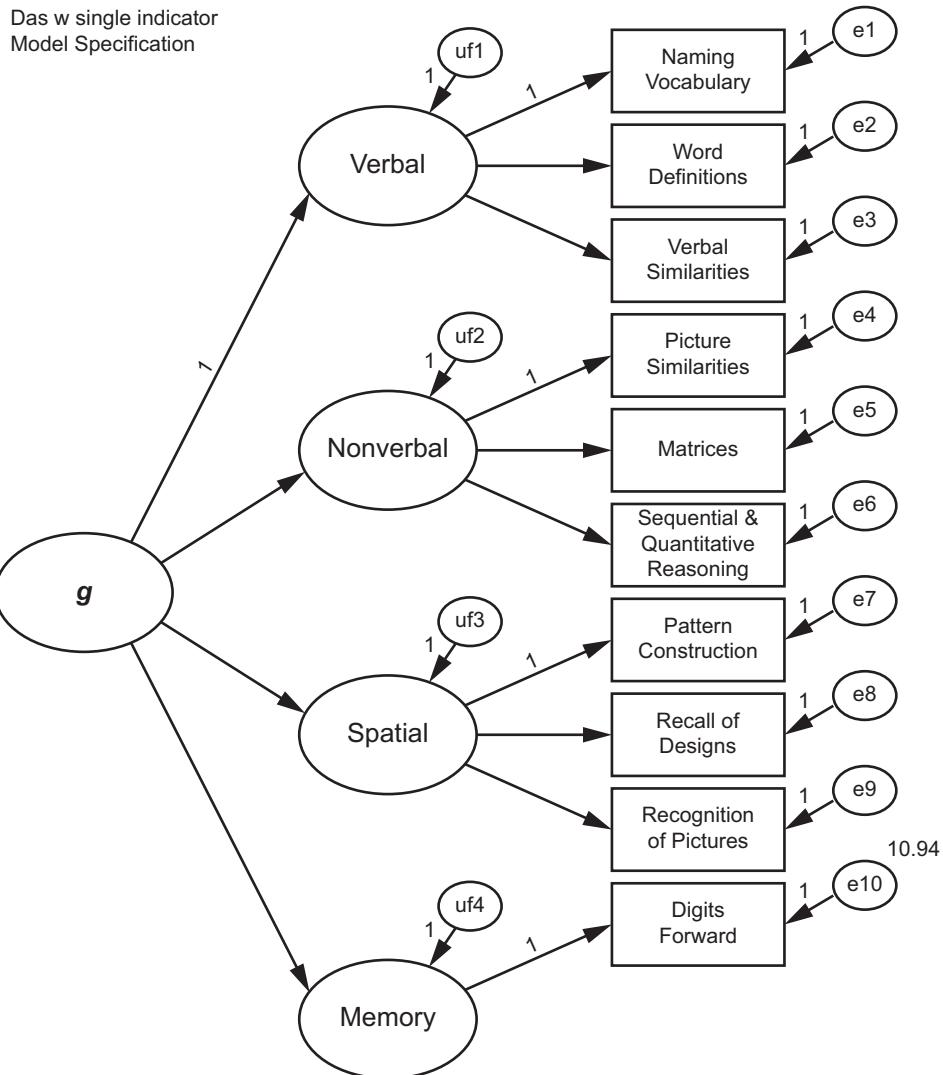


Figure 16.17 Modeling a single-indicator factor. In this model the memory factor has only a single measured variable.

Study this portion of the model. As for all other factors, one path from the latent to the measured variable is set to 1 in order to set the scale. The only difference is that there is only one path from the factor to the measured variable. The path from the unique variance to the subtest is also set to 1, again to set the scale. Recall when we discussed estimating path models via SEM programs we noted we can either estimate the path from the disturbance or estimate the variance of the disturbance. It is the same with the unique and error variances. Normally, we set the path from the unique-error variance to 1 and estimate the unique and error variance. With only a single measured variable, we have to fix the unique-error variance as well as the path to allow model estimation. The value 10.94 beside e_{10} shows that we have done so, and with this constraint we can estimate the model successfully. Again, this is a common method for dealing with single-indicator latent variables; for more detail, see Hayduk (1987, chap. 4). In fact, this was the method I used to estimate the models showing

the effects of different degrees of error in the previous chapter (although I did not show these aspects of the models in the figures). It is also possible to use estimates of validity to account for both unreliability and invalidity. The use of reliability probably provides a very conservative (lower-bound) estimate for the unique and error variance (e_{10}). In the complete higher-order model (the model in Figure 16.11), the estimate for Digit Forward's unique and error variance was 67.91 (this information is contained in the text output or the unstandardized estimates, neither of which are shown here). Some writers recommend using a range of values in such single-indicator analyses to make sure the estimates obtained for loadings and paths are reasonable.

The results of this analysis are shown in Figure 16.18. With this approach the Memory factor had a considerably lower loading on the g factor than did the other first-order factors (and it was

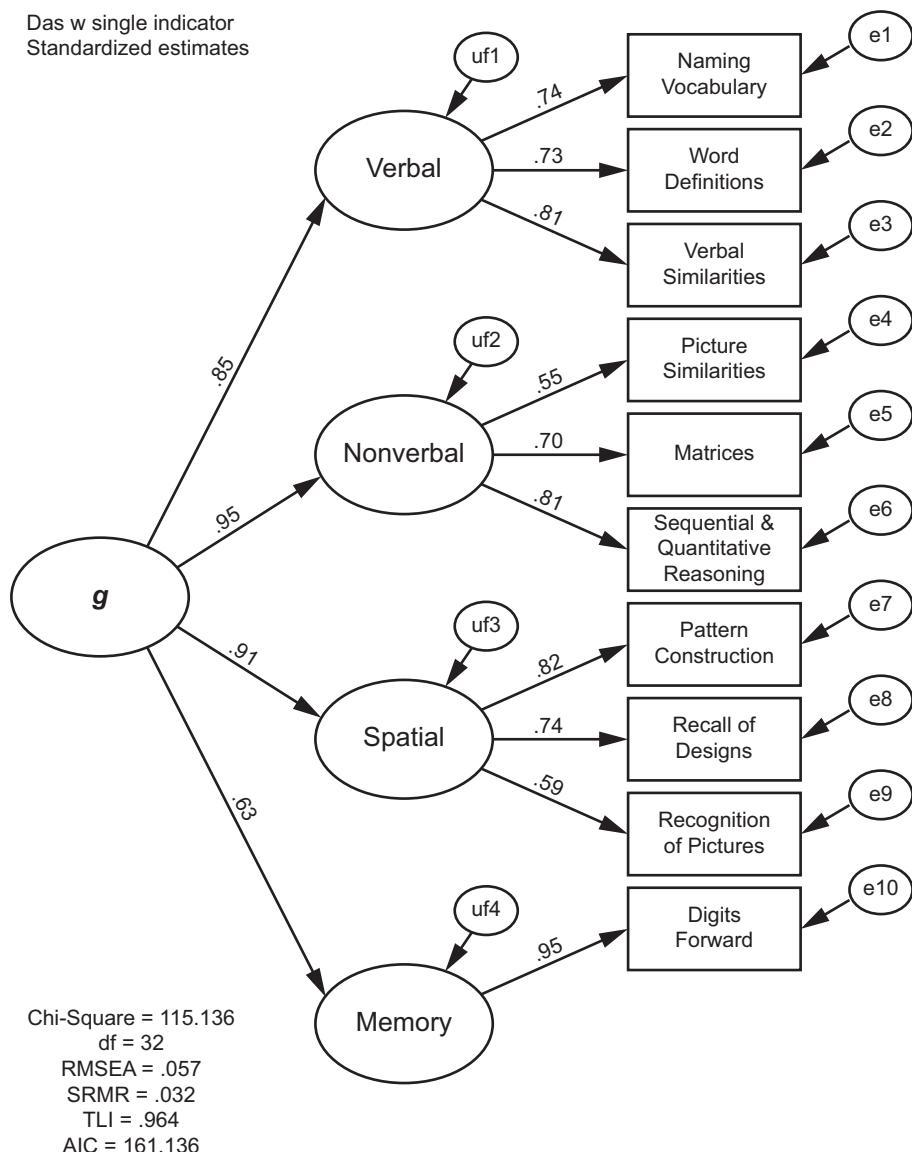


Figure 16.18 Standardized solution for the model with a single indicator for the memory factor.

considerably lower than in higher-order model with three memory indicators). Although this method allows us to estimate a model with single-indicator factors, it obviously provides less information about these factors than do factors defined by multiple measured variables. For the current example, the model tells us the relative effect of g on Memory (with Memory defined as very closely related to the Digits Forward subtest), but it provides little additional information concerning the nature of the Digits Forward subtest or the Memory factor. Although many SEM users regard this method for dealing with single indicators as a trick to allow estimation, Hayduk has argued persuasively for advantages for this approach in path analysis and SEM (1987).

Let's briefly review two alternative methods for dealing with single indicator factors. Figure 16.19 shows the results of a model in which the unique and error variance for Digits

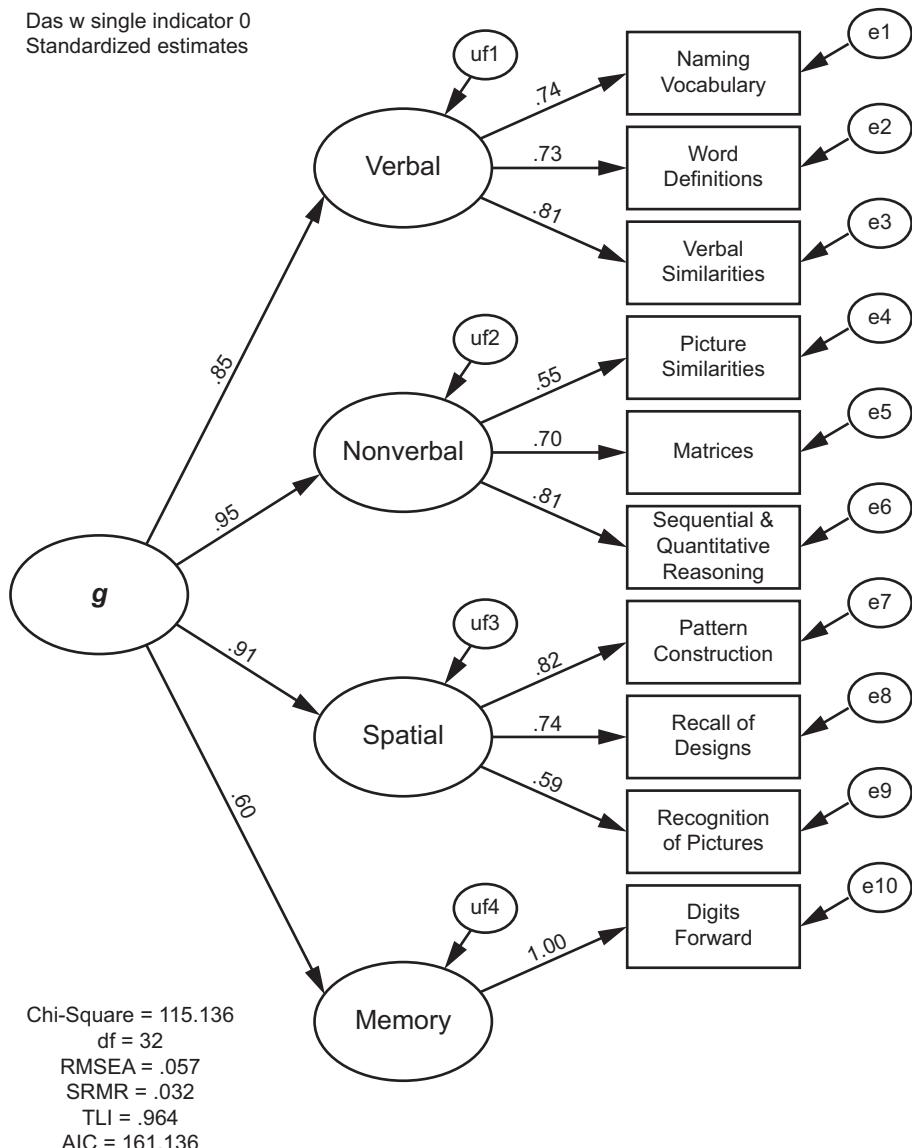


Figure 16.19 An alternative specification with a single indicator. Here, we have constrained the error variance for the Digits Forward test to zero, which essentially says that the subtest is perfectly reliable and that the memory factor and the subtest are equivalent.

Forward (e10) was set to zero. Note that the fit indexes for this model are the same as those shown for Figure 16.18, but that the estimates of the first and second-order factor loadings for Digits Forward and Memory are different. A third possible method is shown in Figure 16.20, in which Digits Forward is loaded directly on the *g* factor; here we essentially say that we don't know what the Digits Forward test measures other than general intelligence. It may not be immediately obvious, but this model is statistically and conceptually equivalent to the previous one. Note that in Figure 16.19 by setting e10 to zero we essentially said that the Memory factor and the Digits Forward subtest are the same "thing." Note also that the loading of Digits Forward on the second-order *g* factor are identical in the two models (Figures 16.19 and 16.20). Whether you think you will ever use single-indicator latent variables or not, I encourage you to try estimating these three models. You will learn a lot about latent

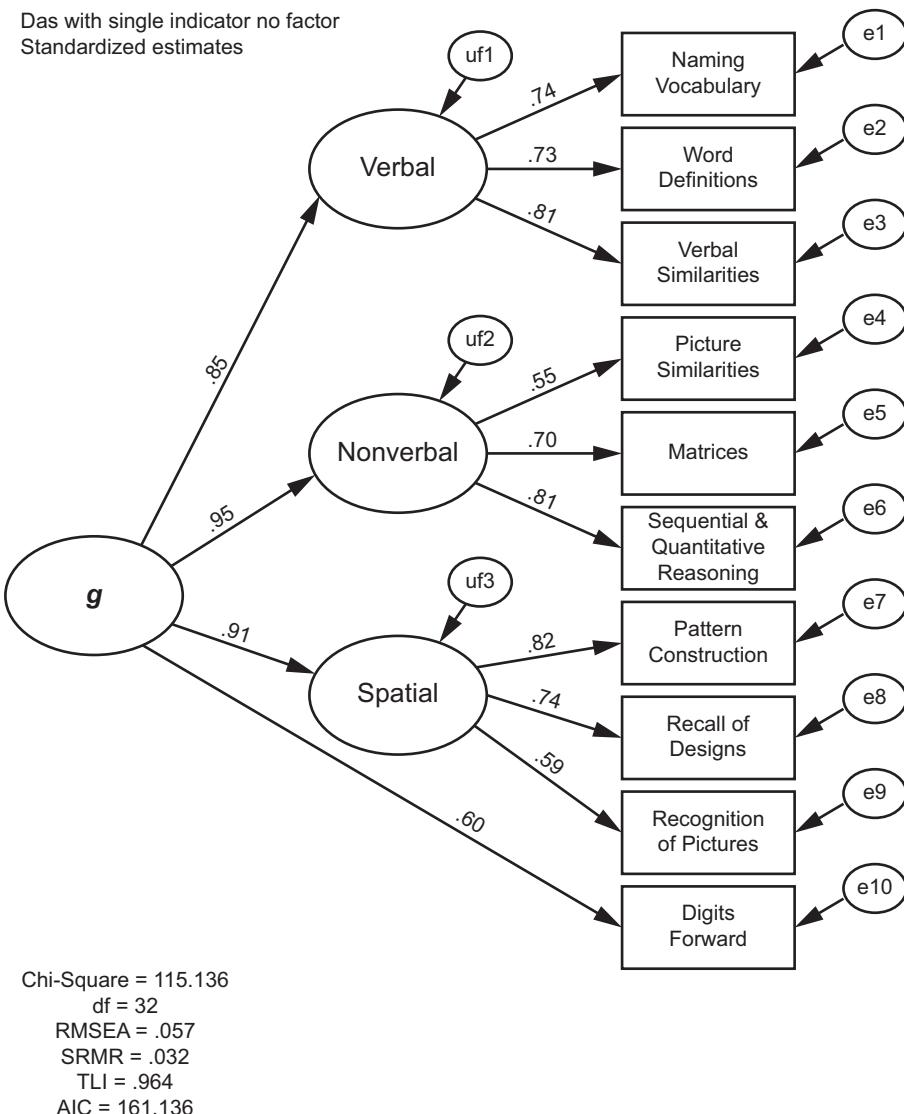


Figure 16.20 Yet another method for dealing with a single-indicator. Although it seems quite different, this model is interchangeable with the previous one.

variables and alternative models in the process. Make sure you carefully examine the unstandardized estimates in addition to the standardized values shown here.

If the DAS-II indeed only included a single measure of short-term memory, there is a more powerful method for better understanding the nature of the constructs being measured by the DAS-II Digits Forward test and the Memory factor (in addition to the other factors). This more powerful method would be to factor analyze the DAS-II with another test that includes known measures of short-term memory along with other related factors. For example, Stone (1992) analyzed the original DAS along with another intelligence test, the Wechsler Intelligence Scale for Children—Revised (Wechsler, 1974) to better understand the constructs measured by both tests.

The examples in this chapter have focused on testing the validity of existing measures. CFA can also be used to test theories. I have mentioned three-stratum theory in the area of intelligence. The DAS-II, it appears, measures several important constructs from three-stratum theory, and thus we can use three-stratum theory to develop a better understanding of what the DAS-II measures. We can turn this process around, as well, to examine the validity of the guiding theory. If we develop multiple measures of the constructs in three-stratum theory, CFA can be used to determine whether a three-stratum-derived model fits the data better than do plausible alternative theories (see Keith & Reynolds, 2018 for more information).

SUMMARY

In the preceding chapter we introduced the full latent variable SEM model. In this chapter we focused on the measurement portion of this model. As it turns out, the measurement model portion of SEM is a useful methodology of its own, generally termed Confirmatory Factor Analysis (CFA). Because the history of factor analysis is so intertwined with the history of intelligence testing, the chapter illustrated CFA through the analysis of a common measure of intelligence, the Differential Ability Scales, Second Edition (DAS-II).

The example used 12 subtests of the DAS-II that supposedly measure four underlying constructs. We drew a model that shows the relations among the factors and subtests (latent and measured variables) (Figure 16.1). The model specifies, with paths drawn from factors to subtests, which subtests load on, or measure, which factors. Consistent with our rules for other path models, each subtest also has a small latent variable pointing to it that represents all other influences on the subtest beyond the four latent factors. With CFA/measurement models, these other influences represent a combination of errors of measurement along with unique or specific influences. With the addition of constraints to set the scales of the latent factors and the unique variances and correlations among the factors—the conceptual model underlying the DAS-II is a testable confirmatory factor model.

We estimated the DAS-II model with data derived from the DAS-II standardization sample. The initial model fit the data well according to the stand-alone fit indexes that we have used in previous chapters (e.g., RMSEA = .046, SRMR = .027), and most of the subtests appeared to measure their corresponding factors strongly. That is, the paths from factors to measured variables, or factor loadings, were generally high. Another way of interpreting these loadings is that the latent constructs (e.g., verbal ability, spatial ability) had strong effects on the corresponding subtests. The factors, or latent constructs, also correlated substantially with each other; all correlations were .75 or larger. This finding suggests that these latent, broad abilities are substantially related to each other.

The common method of setting the scale of latent variables is to set one path from each latent variable to 1, which sets the scale of the variable to be the same as that of the measured variable (the Unit Loading Identification, or ULI, approach, Kline, 2016). An alternative

method is to set the variance of the latent variable to 1 (the UVI, or unit variance identification approach). When done with first-order factors, this method turns the factor covariances in the unstandardized solution into factor *correlations*, because a correlation matrix is simply a covariance matrix among standardized variables. This methodology may be useful to test hypotheses about factor correlations. It also results in tests of all factor loadings for statistical significance.

Just as we can test competing path models using fit statistics, so can we test alternative competing CFA models. We illustrated the testing of competing models by comparing the initial four-factor DAS-II model with a model with a cross-loading and with an alternative three-factor model. In both cases, the initial model fit the data better than did the competing models.

When we wish to use information from the model results to revise the model, several aspects of the SEM program output may be useful. Modification indexes and standardized residual covariances may suggest relaxations in the model that will lead to a better fit. Residual correlations may also be used and have an easier-to-understand metric than standardized residual covariances. Residual correlations are not displayed as output in many SEM programs but are easy to compute. Using these data for model modifications will result in less parsimonious but presumably better fitting models. Using the z (CR) values may lead to values that can be constrained and thus should lead to more parsimonious but equivalent fitting models. You should use such methods to modify models sparingly or else recognize that you are using CFA in an exploratory rather than a theory-testing manner. Model modifications should also be justifiable based on logic, theory, and previous research.

We are often interested in higher-order or other hierarchical models. The field of intelligence is replete with higher-order models, but such models may be relevant in other fields, as well. For the DAS-II example, we hypothesized that a more general factor, often symbolized as g for general intelligence, affects each of the four latent variables, which, in turn, affect the subtests. Said differently, our higher-order model explains that the correlations among the latent factors is a product of their each being affected, in part, by another, more general factor.

An alternative hierarchical model, commonly known as the bifactor model, was also illustrated and tested against the DAS-II data. The bifactor model has shown renewed popularity in recent years and is sometimes considered as a more agnostic version of a hierarchical model. As always, I urge you to consider your underlying theory carefully and allow that theory to guide your model.

It is possible to model latent variables or factors when some of these latent variables include only a single measured variable by constraining the unique–error variance (i.e., $e10$ in Figure 16.14) to some value. A common method of estimating that unique–error variance uses estimates of the reliability of the measured variable (and thus really only models the error variance, not the specific variance). This may prove a useful method when we only have a single indicator, but we recognize that the variables are not error free. The method can be used in both CFA and SEM models. The chapter ended with a hint of some other uses of CFA.

EXERCISES

1. Conduct the analyses outlined in this chapter. The initial four-factor model is on the accompanying Web site (www.tzkeith.com) as the file “DAS-II first 1.amw,” and the data are in the file labeled “das 2 cov.xls” or “DAS 2 cov.sav.”
2. The files “DAS 5–8 simulated 6.sav” and “DAS 5–8 simulated 6.xls” include 500 cases of simulated data for the DAS-II.

- a. Conduct the first-order factor analyses from this chapter using the simulated data. Interpret the findings. How do the results compare with those in this chapter (and in Exercise 1)? Would you come to different conclusions following these analyses than we did in the chapter?
 - b. Note the fit indexes. Which changed the most from the analyses in the chapter? Why do you think this may be?
 - c. As you examine your analyses, are any other hypotheses or models suggested by the findings? If so, conduct these analyses and interpret the findings.
3. The NELS data include a series of items (ByS44a to ByS44m) designed to assess students' self-esteem and locus of control. Conduct a CFA for the self-concept and locus of control items. First, check the descriptive statistics for the items in the raw data file (ByS44a through ByS44m). Note that some of the items are worded positively and others negatively. Given the scaling (1 = strongly agree to 4 = strongly disagree), for positively worded items larger numbers actually represent worse self-concept or locus of control. I recommend you analyze the matrix data in the files "sc locus matrix.sav" or "sc locus matrix.xls" where the positively worded items have been reversed so that for all items high scores represent better psychological health. Items that have been reversed end with an "r" in the matrix and in Figure 16.21. If you analyze the raw data I recommend you reverse these items.

Figure 16.21 shows the model that I recommend you start with. You will find that the model does not fit well (I obtained $\chi^2 = 768.855$, $df = 64$, and $CFI = .780$). First, examine the item wording (shown in Table 16.5). How might you modify the model to improve its fit (think of this as an informal theory method of revising the model)? Do those modifications improve the fit to a statistically significant degree? Now take

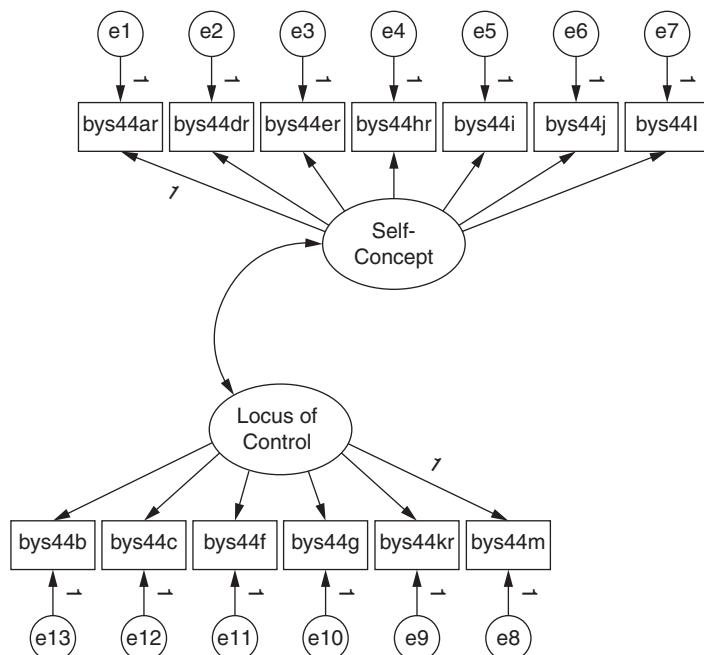


Figure 16.21

Table 16.5 Variable labels for self-concept and locus of control items.

Variable	Label
bys44a	I FEEL GOOD ABOUT MYSELF
bys44b	I DON'T HAVE ENOUGH CONTROL OVER MY LIFE
bys44c	GOOD LUCK MORE IMPORTANT THAN HARD WORK
bys44d	I'M A PERSON OF WORTH, EQUAL OF OTHERS
bys44e	I AM ABLE TO DO THINGS AS WELL AS OTHERS
bys44f	EVERY TIME I GET AHEAD SOMETHNG STOPS ME
bys44g	PLANS HARDLY WORK OUT, MAKES ME UNHAPPY
bys44h	ON THE WHOLE, I AM SATISFIED WITH MYSELF
bys44i	I CERTAINLY FEEL USELESS AT TIMES
bys44j	AT TIMES I THINK I AM NO GOOD AT ALL
bys44k	WHN I MAKE PLANS I CAN MAKE THEM WORK
bys44l	I FEEL I DO NOT HAVE MUCH TO BE PROUD OF
bys44m	CHANCE AND LUCK IMPORTANT IN MY LIFE

a look at the modification indexes and the standardized residuals. Produce a table of correlation residuals (the sample correlation matrix minus the implied correlations). What modifications might you make based on these various hints? How does the model fit now?

How many modifications did you make? Have you crossed the line from confirmatory analysis to exploratory analysis? Take a step back and think of your model more broadly. Might the model be better conceived as having more than two factors? Might it be worth deleting some messy items? Discuss your models and your thoughts in class.

Notes

- 1 In higher-order intelligence models, Heywood cases often show up in connection with Fluid Reasoning factors (Gf, in the DAS-II represented by the Nonverbal Reasoning factor). When this happens, the g to Gf path may approach or exceed 1 and the associated unique factor variance become negative. Note in Figure 16.11 that the g to Nonverbal Reasoning loading approached 1. One implication of such a finding is that g and Gf factors are not separable. Some researchers use this not-uncommon finding to argue that the Gf factor is redundant with g, whereas others argue that this shows that g is redundant. As noted, one common method for dealing with negative variances is to set the value to zero. This makes sense if the value is fairly close to zero but is less defensible if it is a large negative value (which likely indicates problems with the model). There are also other possible ways to deal with negative variances, including constraining the value to be positive.
- 2 Here is an interesting conundrum. When factors are correlated, it is possible (although not desirable) to have factors referenced by only two measured variables each. So, for example, a correlated two-factor, four-measured variable model would have one degree of freedom. But when factors are uncorrelated, each factor requires a minimum of three measured variables for identification, and with three measured variables each factor is just-identified (as in the present bifactor example). That means that if a bifactor model includes fewer than three variables for a factor, the researchers will need to either make additional constraints (e.g., constrain the two factor loadings to be equal) or, counter-intuitively, relax constraints (e.g., allow that factor to be correlated with another factor).

As you are reading research using the bifactor model and you notice only two measured variables on a factor, make sure the researchers tell you what they have done to solve this problem! This conundrum of identification also occasionally leads to a phenomenon known as “empirical underidentification” in which a model allows factors to be correlated, but that correlation is small and nonsignificant. If one of the offending factors involves fewer than three measured variables, it will thus be underidentified. The phenomenon of empirical underidentification applies to first-order factor models as well (Kenny, 1979).

17

Putting It All Together

Introduction to Latent Variable SEM

Putting the Pieces Together	389
An Example: Effects of Peer Rejection	391
<i>Overview, Data, and Model</i>	391
<i>Results: The Initial Model</i>	395
Competing Models	399
<i>Other Possible Models</i>	400
Model Modifications	402
Summary	404
Exercises	405
<i>Note</i>	408

Let's review our progress in our adventures beyond MR. You know how to conduct path analysis using MR. This experience includes the estimation of standardized and unstandardized paths, the calculation of disturbances ($\sqrt{1 - R^2}$), and the calculation and comparison of direct, indirect, and total effects using two different methods. We transitioned into estimating path models using Amos and other SEM programs and focused again on the estimation of both standardized and unstandardized effects and direct, indirect, and total effects. With Amos, we switched from the estimation of the paths from disturbances to estimating the variances of the disturbances, although either is possible. We have defined just-identified, overidentified, and underidentified models, and I suggested that you use a SEM program to estimate overidentified models but use either MR or an SEM program if your models are just-identified. We have examined fit indexes for overidentified models and have highlighted a few that are useful for evaluating a single model and those that are useful for comparing competing models. We briefly focused on equivalent models, nonrecursive models, and longitudinal data. We focused on the effects of measurement error on path analysis, MR, nonexperimental research, and research in general and began considering the use of latent variables as a method of obviating this threat. We expanded our knowledge of latent variables, their meaning, and estimation via confirmatory factor analysis.

PUTTING THE PIECES TOGETHER

In this chapter, we will begin putting all these pieces together in latent variable structural equation modeling. As noted in Chapter 15, you can consider latent variable SEM as a

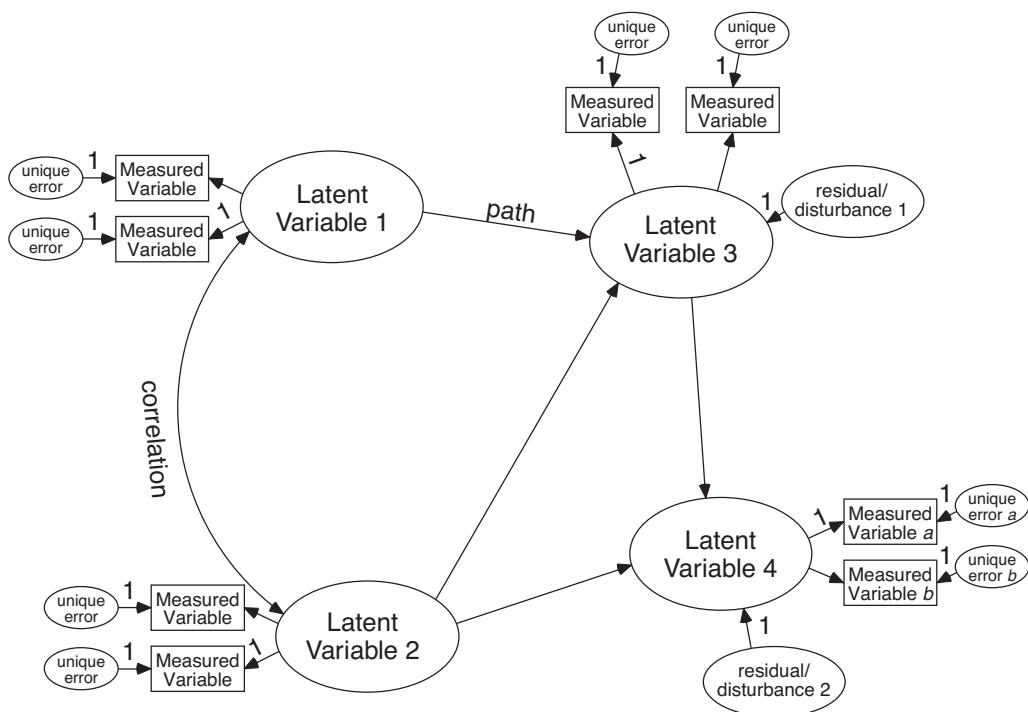


Figure 17.1 Full latent variable SEM model

confirmatory factor analysis of the constructs involved in the research project, along with a path analysis of the effects of these constructs on each other. For this reason, many writers refer to these as the measurement model and the structural model, respectively (e.g., Mulaik & Millsap, 2000), to denote the conceptual distinctions between components of latent variable SEMs. Although this separation of measurement and structural portions is not necessary statistically, it can be very useful conceptually, especially at this stage of learning.

Figure 17.1 displays, for review, the components of a latent variable SEM. The measurement model consists of the estimation of the four latent variables from eight measured variables. The structural model consists of four paths and one correlation among the four latent variables. Note that each variable that has a path pointing to it also has a residual-disturbance-error term pointing to it, representing all other influences on the variable other than the variables pointing to it. Some of these residuals represent the unique and error variances of measured variables, the remaining influences on these measured variables other than the latent variable underlying it. Some residuals represent disturbance terms for latent variables, meaning all remaining influences on these latent variables other than the other latent variables. Although I refer to some of these as unique-error variances and others as disturbances, the terms error and residual are used fairly interchangeably.

Why, you may wonder, doesn't Latent Variable 1 have a disturbance pointing to it? Because Latent Variable 1 has no paths pointing to it; it is exogenous. Note also that each latent variable (including the unique-error variances and the disturbances) has its scale set by fixing a single path from it to another variable to 1. So, for example, the latent variable labeled residual/disturbance 2 has its scale set to the same value as the latent variable labeled Latent Variable 4, which in turn is set to the same value as Measured Variable *a*. Note that the biggest difference between this model and the CFA models from the last chapter is that some correlations among latent variables are replaced by paths. As a result, the latent variables

with paths pointing to them also have disturbances pointing to them. Of course, this is akin to the difference between a correlation matrix of variables and a path model specifying that one variable influences another. Take some time studying the model to make sure you understand it.

AN EXAMPLE: EFFECTS OF PEER REJECTION

Overview, Data, and Model

Eric Buhs and Gary Ladd used SEM to examine the effects of peer rejection on Kindergarten students' academic and emotional adjustment (2001). A portion of the model they analyzed is shown in Figure 17.2. The latent variables in the model, along with the measured variables used to estimate them, were these:

1. Rejection was indexed by averaged sociometric ratings for each child by the other children in the class (Averaged Rating; the scale of this variable was reversed to make it consistent with the negative [Rejection] name of the latent variable) and by the number of times each child was nominated negatively (as someone other children did not want to play with; Negative Nominations).
2. Change, from a previous rating, in Classroom Participation. This variable was estimated from teacher ratings of Cooperative Participation (e.g., accepts responsibility) and Autonomous Participation (e.g., self-directive).

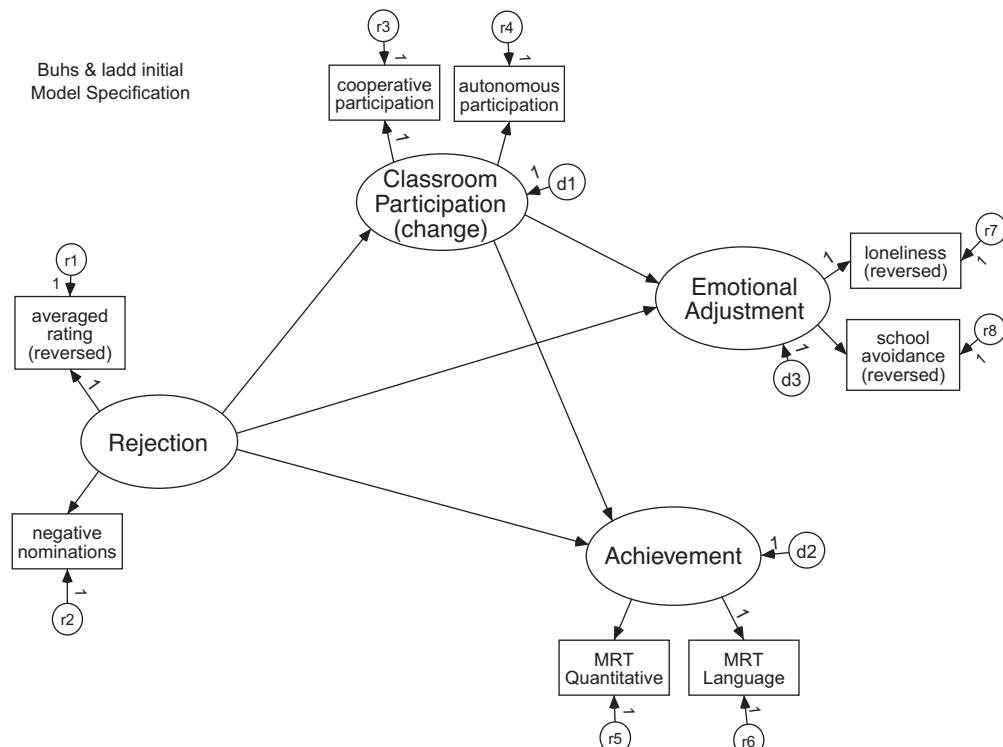


Figure 17.2 Effects of peer rejection on Academic and Emotional Adjustment, initial model. The model was derived from Buhs and Ladd, 2001.

3. Achievement, which the authors considered one aspect of adjustment, was estimated from the Language and Quantitative subtests from a standardized school readiness test (the Metropolitan Readiness Test, Nurss & McGauvran, 1986).
4. Emotional Adjustment, as indexed by self-ratings of students' Loneliness at school and their desire to avoid school (School Avoidance). These two variables were reversed to make the latent variable consistent with the positive name (Adjustment).

Buhs and Ladd's article included an additional intervening variable (Negative Peer Treatment) and an additional indicator of Rejection. These variables were not included here to simplify the model. The model is longitudinal; the Rejection variables were collected in the fall, the other variables in the spring (for more detail, see Buhs & Ladd, 2001).

Recall that with our earlier path models (e.g., the homework models in Chapter 14) many of the variables in the model were composites (e.g., Achievement was a composite of four scores). Buhs and Ladd (2001) could have done the same thing here, but instead of adding Quantitative and Language into an achievement *composite* variable, for example, the authors used these two measures as indicators of an Achievement latent variable. Recall our discussion in Part 1 about multiple regression predicting an outcome variable from an *optimally weighted combination* of the independent variables. Conceptually, the latent variables in SEM are similar: they are optimally weighted combinations of the measured variables.

The model will be estimated from the *measured* variables. A portion of the data is shown in Table 17.1 (and is saved as data files on the Web site under the label "buhs & ladd data.sav" and "buhs & ladd data.xls"). Note there are no variables in the data file corresponding to the latent variables. This is because the latent variables, or factors, are estimated from the measured variables. If this is still confusing, think of the latent variables as *imaginary variables* that we estimate from the measured variables. (In the actual data file, the variable names are shortened versions of the variable labels used in the table and the Amos model, but they should be self-explanatory. Note that the data included here and on the Web site are not the actual data but rather simulated data created to be consistent with the correlation matrix, means, and standard deviations reported in the article. $N = 399$. Three of the measured variables were reversed to make them consistent with the variable names and thus more easily interpretable.)

Table 17.1 Sample Data: Measured Variables for the Peer Rejection Example

Child	Averaged Rating	Negative Nominations	Cooperative Participation	Autonomous Participation	Quantitative	Language	Loneliness	School Avoidance
1	-1.33	-1.09	1.19	.69	7.47	6.30	2.09	2.48
2	1.32	.55	-.13	-.07	2.72	2.76	1.42	2.16
3	-.64	-1.09	-.29	-.126	6.40	5.39	1.59	1.38
4	1.42	-.36	-.19	-.56	.99	1.05	.94	2.13
5	.58	-.01	-.36	-.13	2.80	3.56	.36	2.20
6	-1.20	-1.51	.04	.07	7.07	7.79	1.37	2.03
7	.42	.39	-.25	.40	3.68	3.47	2.08	3.00
8	-.40	-.81	.78	1.03	7.03	4.94	2.03	2.61
9	1.99	1.89	-.45	-.66	1.51	5.08	.66	.99

	Descriptive Statistics									
	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis		
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error	
ave_rat	399	-2.33	3.10	.1200	.95000	.101	.122	.011	.244	
neg_nom	399	-2.37	2.64	-.1000	.90000	.186	.122	-.140	.244	
coop	399	-1.38	1.88	.0000	.60000	.145	.122	-.256	.244	
auto	399	-1.71	1.96	.0000	.62000	.135	.122	.043	.244	
quant	399	-.86	11.91	5.3800	1.98000	-.150	.122	.472	.244	
lang	399	.51	10.47	5.3600	1.78000	-.010	.122	-.302	.244	
lone	399	-.02	3.12	1.5100	.56000	-.047	.122	-.182	.244	
schavoid	399	.12	4.11	2.0500	.67000	.133	.122	-.128	.244	
Valid N (listwise)	399									

Figure 17.3 Descriptive statistics for the simulated rejection data.

Just because our analyses have gotten more sophisticated does not mean we should ignore the mandate from Part 1: Always, always, always, always, always, always check your data prior to conducting analyses! This command is just as important—maybe even more so—as our analyses become more complex. So before conducting the SEMs here, make sure you check means, *SDs*, minimums and maximums of the variables in this file. As we conduct SEM, you should also get in the habit of examining skew and kurtosis. Note that with the current data, few of the measured variables had meaningful scales, and many had both positive and negative values. The averaged ratings, for example, were standardized within classroom. The descriptive data are shown in Figure 17.3.

For the current model, I have symbolized the unique–error variances of the measured variables as *r1* through *r8* and the disturbances of the latent variables as *d1* through *d3*. Recall that we can consider the unique–error variances as all other influences on the measured variables beyond the influence of the latent variable, just as the disturbances are all other influences on a latent variable beyond those of the other latent variables.

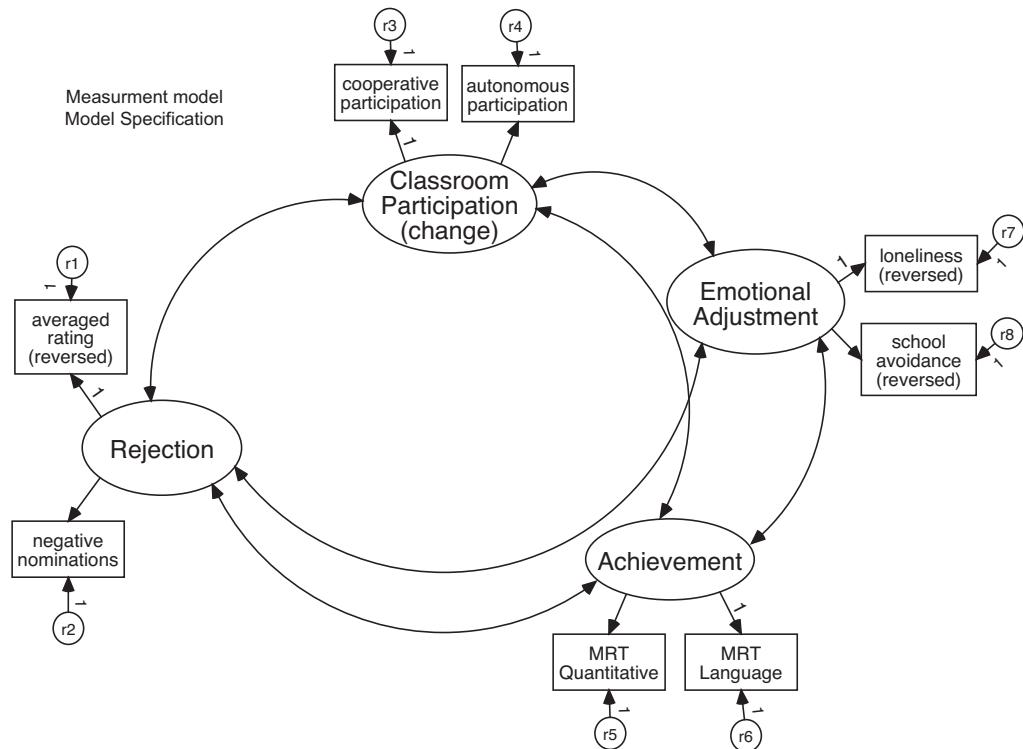
Measurement Model

For the sake of clarity, the measurement model, without the structural model, is shown in Figure 17.4. Except for its placement of variables (in a circular fashion instead of in a line), the model is similar to the confirmatory factor models from the last chapter. The model simply delineates the estimation of the four latent variables (Rejection, Adjustment, etc.) from the eight measured variables (Averaged Rating, Negative Nominations, etc.).

Note that each latent variable had its scale set by a single factor loading (path from the latent to measured variable) set to 1. Each error–unique (residual) variable had its scale set by setting the path from it to its corresponding measured variable to 1.

Structural Model

The structural portion of the model is shown in Figure 17.5, a figural representation of the hypotheses of the effects of one latent variable on another, and includes the disturbances for the endogenous latent variables in the model. The model examines the effect



Figures 17.4 Measurement model portion of the initial peer rejection model.

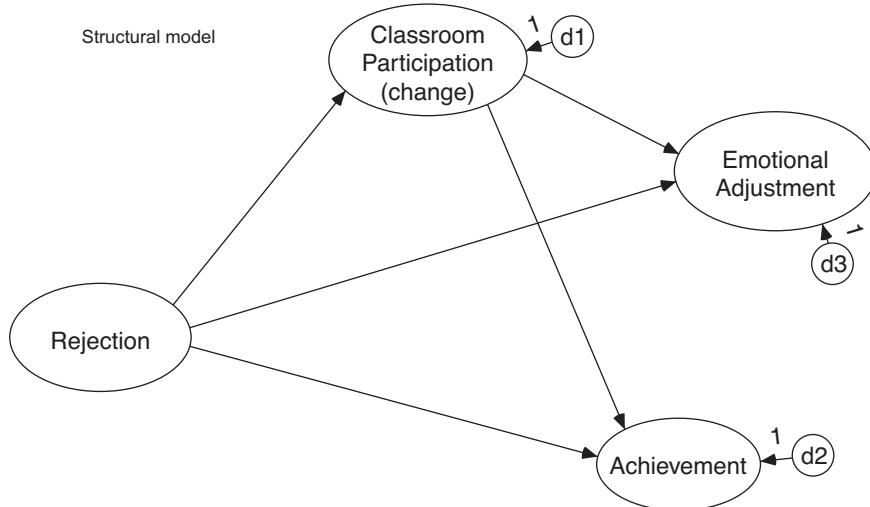


Figure 17.5 Structural model portion of the initial peer rejection model.

of Rejection on Adjustment, both directly and indirectly, through the class participation of the students.

The full SEM model (Figure 17.2) has 15 degrees of freedom. Fourteen degrees of freedom are from the measurement portion of the model (Figure 17.4). Note that all the factor

loadings that could be included in the model (e.g., a path from Rejection to Cooperative Participation or Loneliness) are not included; these constraints are the source of this 14 *df*. The structural model (Figure 17.5) includes one additional *df*, resulting from the omission of a path between Achievement and Adjustment. The model is saved on the Web site (www.tzkeith.com) in the file “Buhs & Ladd model 1.amw.” Note 1 at the end of the chapter shows the calculation of the degrees of freedom.¹

Results: The Initial Model

The model (Figure 17.2) was analyzed using the raw data (Table 17.1 and the file “buhs & ladd data.sav” or “buhs & ladd data.xls”) via Amos. Figure 17.6 shows relevant fit indexes, along with the standardized output. The model shows an adequate, but not good, fit to the data. The RMSEA was above .05 (.067, 90% confidence interval = .043 to .092), but was below .08. The SRMR was below the cutoff of .08 or .06 (.046). The CFI was above .95, but the TLI was below our informal cutoff for a good fit of .95. Although not shown in the figure the χ^2 was also statistically significant ($p < .01$), further suggesting a lack of fit. Again, the model shows an adequate, but not good, fit. The full array of fit indexes is shown in Figure 17.7. Because the model had an adequate fit, we’ll first interpret these results. Later in the chapter we’ll take a look at the more detailed fit information and consider how the model might be modified.

Figure 17.8 shows more detail concerning the paths and factor loadings, including the unstandardized coefficients, their standard errors, and critical ranges (*z* statistics). All the parameters that were estimated were statistically significant (*z* greater than approximately 2).

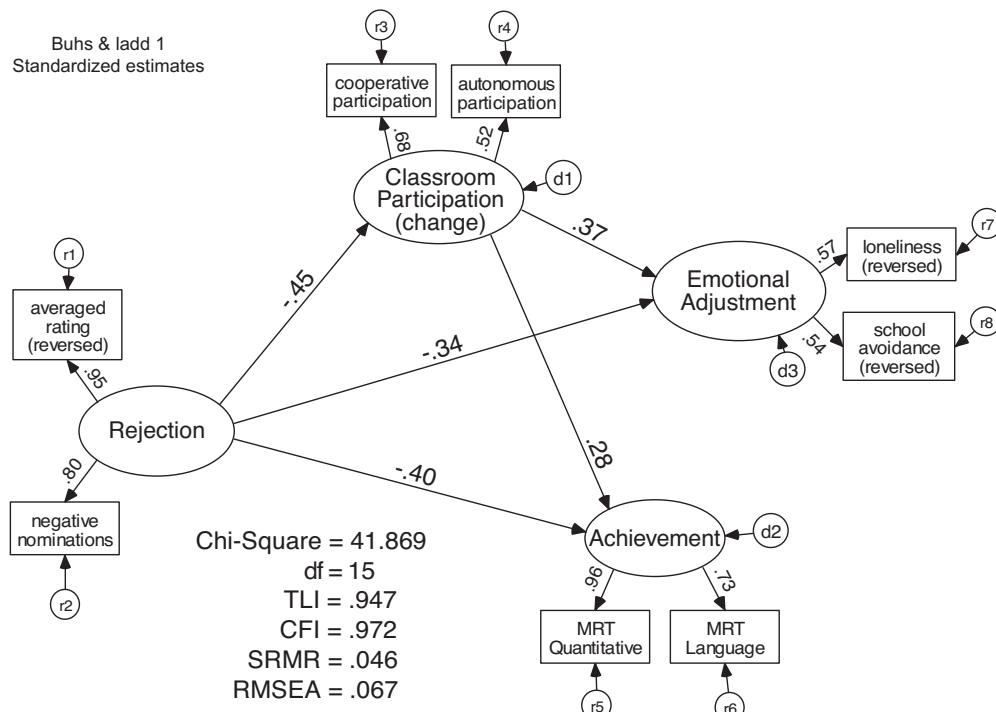


Figure 17.6 Standardized estimates from the initial peer rejection model. The model has an adequate, but not good, fit to the data.

Model Fit Summary***CMIN***

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	21	41.869	15	.000	2.791
Saturated model	36	.000	0		
Independence model	8	972.032	28	.000	34.715

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	.047	.974	.938	.406
Saturated model	.000	1.000		
Independence model	.504	.574	.453	.447

Baseline Comparisons

Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	.957	.920	.972	.947	.972
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

Parsimony-Adjusted Measures

Model	PRATIO	PNFI	PCFI
Default model	.536	.513	.520
Saturated model	.000	.000	.000
Independence model	1.000	.000	.000

FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	.105	.068	.028	.126
Saturated model	.000	.000	.000	.000
Independence model	2.442	2.372	2.125	2.637

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.067	.043	.092	.110
Independence model	.291	.276	.307	.000

AIC

Model	AIC	BCC	BIC	CAIC
Default model	83.869	84.841	167.637	188.637
Saturated model	72.000	73.666	215.603	251.603
Independence model	988.032	988.403	1019.944	1027.944

Figure 17.7 Fit indexes for the initial rejection model.***Standardized Results***

Let's now focus on the meaning of the results (Figure 17.6). Our primary interest was in the effects of Rejection on kindergarten students' academic Achievement and Emotional Adjustment. The standardized direct effect of Rejection on Achievement was $-.40$, whereas the direct effect on Emotional Adjustment was $-.34$. Both effects were statistically significant and

large. Given the adequacy of the model, for each *SD* change in the latent Rejection variable, Emotional Adjustment should decrease by .34 of a standard deviation, and Achievement should decrease by .40 of a *SD*, other things being equal. These findings, in turn, suggest strong effects for Rejection on kindergarteners' subsequent Adjustment, both academically and emotionally. Obviously, Rejection can have deleterious effects.

Unstandardized Findings

Focus on the unstandardized coefficients (Figure 17.8). The unstandardized direct effect of Rejection on Emotional Adjustment was $-.118$, meaning that for each 1-unit change in the latent Rejection variable Emotional Adjustment decreased by .118 points. To understand the meaning of this statement, we need to understand the scales involved. The Rejection latent variable was set to have the same scale as the measured Averaged Ratings variable, whereas the Emotional Adjustment latent variable was set to the same scale as the Loneliness scale. The Averaged Ratings variable was originally based on a 3-point scale but was each child's average rating on this 3-point scale by all of his or her classmates. In addition, these ratings were standardized separately by classroom (Buhs & Ladd, 2001). This seems a good approach, but it means

Regression Weights

		Estimate	S.E.	C.R.	P
Classroom_Participation_(change) <--- Rejection		-.205	.034	-6.055	***
Emotional_Adjustment <--- Classroom_Participation_(change)		.289	.098	2.944	.003
Achievement <--- Rejection		-.578	.105	-5.526	***
Achievement <--- Classroom_Participation_(change)		.886	.274	3.236	.001
Emotional_Adjustment <--- Rejection		-.118	.034	-3.430	***
NEG_NOM <--- Rejection		.802	.057	14.157	***
AVE_RAT <--- Rejection		1.000			
COOP <--- Classroom_Participation_(change)		1.000			
AUTO <--- Classroom_Participation_(change)		.788	.141	5.596	***
LONE <--- Emotional_Adjustment		1.000			
SCHAVOID <--- Emotional_Adjustment		1.140	.223	5.104	***
LANG <--- Achievement		1.000			
QUANT <--- Achievement		1.465	.134	10.901	***

Standardized Regression Weights

		Estimate
Classroom_Participation_(change) <--- Rejection		-.451
Emotional_Adjustment <--- Classroom_Participation_(change)		.372
Achievement <--- Rejection		-.403
Achievement <--- Classroom_Participation_(change)		.281
Emotional_Adjustment <--- Rejection		-.335
NEG_NOM <--- Rejection		.803
AVE_RAT <--- Rejection		.949
COOP <--- Classroom_Participation_(change)		.682
AUTO <--- Classroom_Participation_(change)		.520
LONE <--- Emotional_Adjustment		.567
SCHAVOID <--- Emotional_Adjustment		.540
LANG <--- Achievement		.726
QUANT <--- Achievement		.957

Figure 17.8 Unstandardized and standardized paths and loadings, standard errors, and critical ratios.

that the Averaged Ratings unstandardized metric and thus the metric of the Rejection latent variable are not readily interpretable. According to the authors, the Loneliness scale is a five-item composite (Buhs & Ladd). Although not explained further, it appears from the means and standard deviations that this scale is also a mean of the item scores. Without further detail, the unstandardized metric of this variable and thus the Emotional Adjustment latent variable are also not interpretable. The unstandardized coefficients, although useful for other purposes (e.g., comparisons with other research), are not readily interpretable, and thus the previous interpretation of the standardized paths is probably our best approach.

Mediation

Many more interesting findings are contained in the model beyond the direct effects. One primary interest of the researchers was to determine whether classroom participation *mediated* the effect of Rejection on Adjustment. In other words, what were the *indirect* effects of Rejection on Adjustment through Classroom Participation? Note in Figure 17.8 that Rejection had a powerful effect on Participation (-.45): rejected children showed less participation than did their nonrejected peers. Classroom Participation, in turn, had a strong effect on both Achievement (.28) and on Emotional Adjustment (.37); children who participated evidenced higher achievement and better adjustment. Thus, it certainly seems that the indirect effects of Rejection on the two adjustment variables were also substantial and that Classroom Participation partially mediates the effects of Rejection on Adjustment.

Indirect and Total Effects

Figure 17.9 shows the standardized direct, indirect, and total effects of the latent variables on each other. Rejection had moderate and negative indirect effects on Achievement (-.126) and Emotional Adjustment (-.168). Although not shown in the figure, these effects were also

Standardized Total Effects

	Rejection	Classroom_Participation_(change)	Achievement	Emotional_Adjustment
Classroom_Participation_(change)	-.451	.000	.000	.000
Achievement	-.529	.281	.000	.000
Emotional_Adjustment	-.503	.372	.000	.000

Standardized Direct Effects

	Rejection	Classroom_Participation_(change)	Achievement	Emotional_Adjustment
Classroom_Participation_(change)	-.451	.000	.000	.000
Achievement	-.403	.281	.000	.000
Emotional_Adjustment	-.335	.372	.000	.000

Standardized Indirect Effects

	Rejection	Classroom_Participation_(change)	Achievement	Emotional_Adjustment
Classroom_Participation_(change)	.000	.000	.000	.000
Achievement	-.126	.000	.000	.000
Emotional_Adjustment	-.168	.000	.000	.000

Figure 17.9 Standardized total, direct, and indirect effects for the initial rejection model.

statistically significant (tested via bootstrapping in Amos or another SEM program). Although these effects are smaller than the direct effect of Rejection on each variable, they are meaningful and show that students' participation in class partially mediates the effects of rejection on adjustment. Children who are rejected by their peers show less participation, which, in turn, results in lower levels of school emotional adjustment and achievement. Because the direct and indirect effects of Rejection on the academic (Achievement) and Emotional Adjustment variables were both negative, the total effects were even larger ($-.529$ on Achievement; $-.503$ on Emotional Adjustment). (Of course we could have calculated these indirect and total effects by hand. For example, the indirect effect of Rejection on Achievement via Participation = $-.451 \times .281 = -.127$. The total effect = $-.127 - .403 = -.530$ [the same value as the figure, within errors of rounding]. With more complex figures, of course, such calculations become considerably more complex.)

COMPETING MODELS

We may wonder if the model, as drawn, is correctly specified. Is it reasonable, for example, to assume that the *only way* Achievement and Adjustment are related to each other is by their both being affected by Rejection and Participation? Or does Achievement affect Adjustment, as well (or Adjustment affect Achievement)?

Figure 17.10 shows an alternative model in which Achievement affects Adjustment. The logic behind this competing model is simple: children who are successful academically, a major component of the orientation of kindergarten, will, as a result, be better emotionally

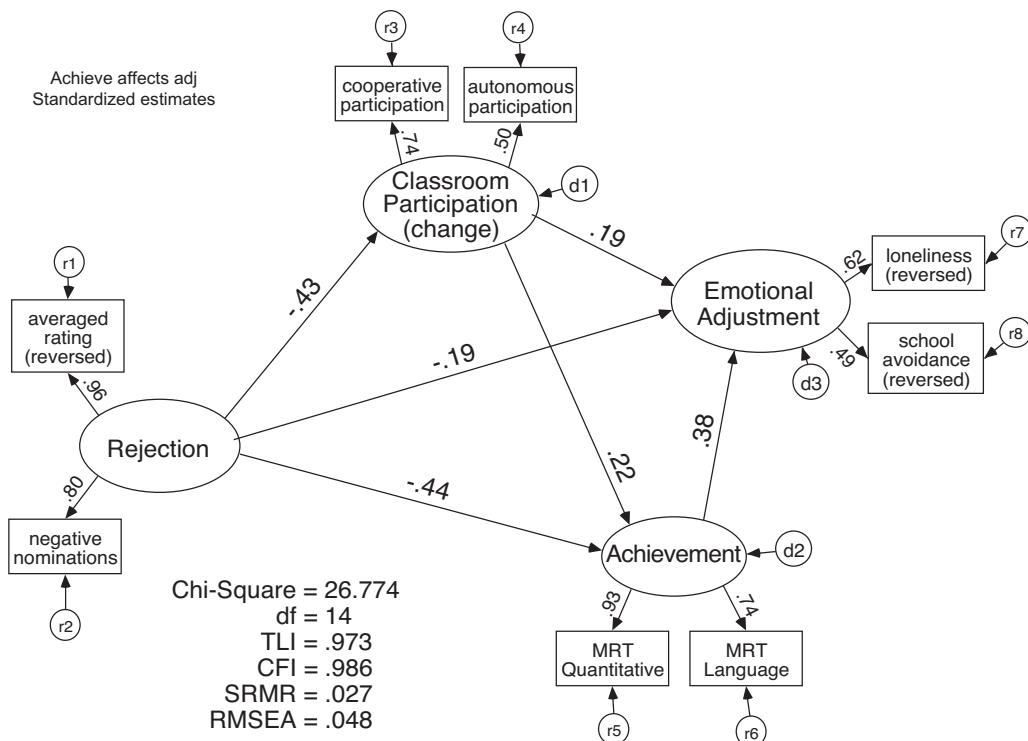


Figure 17.10 Alternative Achievement Effect model of the effects of rejection on educational and emotional adjustment. The model includes a path from Achievement (educational adjustment) to Emotional Adjustment.

Table 17.2 Comparison of the Fit of Alternative Peer Rejection Models

Model	χ^2	df	$\Delta\chi^2$	df	p	AIC	TLI	CFI	SRMR	RMSEA (90% CI)
1. Initial	41.869	15				83.869	.947	.972	.046	.067 (.043-.092)
2. Achievement Effects	26.774	14	15.095	1	< .001	70.774	.973	.986	.027	.048 (.018-.075)

adjusted than will children who have difficulty with the academic aspects of kindergarten. As shown in the figure, this model had a good fit to the data. In particular, the RMSEA was .048, and the TLI and CFI were above .95.

More directly, we can compare the fit of this model with the initial model. Because the two models are nested, we can use $\Delta\chi^2$ to compare the two models. The fit statistics for this Achievement Effect model are shown in Table 17.2, along with those from the initial model. As can be seen in the table, the model in which Achievement was allowed to affect Adjustment resulted in a smaller χ^2 than did the initial model, and this $\Delta\chi^2$ was statistically significant ($\Delta\chi^2 [1 df] = 15.095, p < .001$). Although the initial model was more parsimonious, our rule of thumb is that when $\Delta\chi^2$ is statistically significant we will reject the more parsimonious model in favor of the better fitting model. In this case, the model shown in Figure 17.10 is the better fitting model; the decrease in parsimony is worth the decrease in χ^2 .

Given our acceptance of the Achievement Effect model over the Initial Model, what are the implications for this new model? The results shown in Figure 17.10 suggest that Achievement has a powerful effect on Emotional Adjustment ($\beta = .38$). If this model is correct, then it appears that Achievement is an important mediating variable between Rejection and Adjustment: children who are rejected suffer academically, and this academic difficulty, in turn, results in lower levels of adjustment in school.

This change in the model also substantially reduced the direct effect of both Rejection and Participation on Emotional Adjustment (compare the models shown in Figures 17.6 and 17.10). If you compare the total effects for Rejection on Adjustment in the two models, however, you will find them to be similar. Take a few minutes to consider why this is the case. As long as you are pondering models, it is also worth noting that with the Achievement Effect model (Figure 17.10), the structural portion of the SEM (the paths among the latent variables) is just-identified. That is, for a measurement model there are six correlations among the latent variable; for the model shown in Figure 17.10, all six of those correlations are used to estimate the six paths among the latent variables. Finally, please note that these results are with simulated data. I do not know if the addition of this path would have led to such an improvement in fit in the actual data.

Other Possible Models

You may question why I drew the path from Achievement to Adjustment rather than the reverse. The decision was based primarily on logic. I reasoned that the types of skills and abilities assessed by the Achievement measured variables are more stable than the ratings of loneliness and school avoidance assessed by the Adjustment latent variable. Given what

is meant by these two latent variables, it seemed to me that it was more likely that Achievement would affect Adjustment than it was that Adjustment would affect Achievement. What do you think? Should the path go in this direction or the reverse? It is interesting to conduct this exercise, but if we examine this model as more than an exercise, we will need to examine relevant theory and previous research to see which of these possibilities is more likely. We would use such theory and research to design the study and to draw the path in the appropriate direction.

Why not, you may wonder, just estimate a model with the path drawn in the opposite direction and see how that model fits? Recall the rules for equivalent models in Chapter 14. Unfortunately, these two models are statistically equivalent; their fit is identical. Although this alternative Adjustment Effect model will have very different implications for interpretation, the data cannot tell us which model is correct. It is also inappropriate to run this alternative model, interpret it, and then decide which interpretation we like more. Perhaps, then, we can draw the paths in both directions, a nonrecursive model, and see which path is stronger? This solution will not work either; the structural model will be underidentified. If differentiation between these two models is one of the purposes of the research, the researchers could build in noncommon causes of the two outcome variables and thus test nonequivalent or nonrecursive models; likewise, longitudinal data will help. With the current model and data, we must rely on theory and previous research to make this decision.

What if theory and previous research do not inform this decision; what if you cannot decide in which direction to draw the path? One option, an agnostic option, is shown in Figure 17.11. In this model, we have allowed the disturbances of Achievement and Adjustment

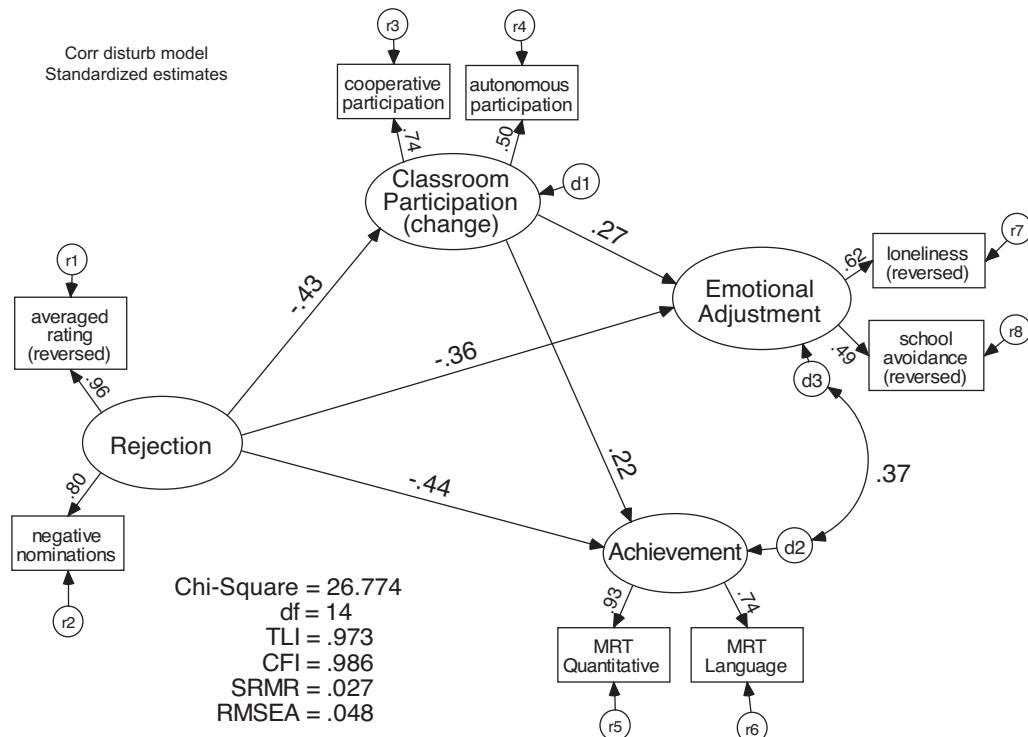


Figure 17.11 Another alternative model of the effects of rejection. This agnostic model specifies an unknown causal relation between Emotional Adjustment and Achievement. The model is equivalent to and statistically indistinguishable from the previous Achievement Effect model.

to be correlated. Note that this model is also equivalent to the model in Figure 17.10; the fit indexes are therefore the same, and the data cannot tell us which of the two models is correct. But consider what this model with the correlated disturbances says about our assumptions of the causal process underlying these variables. The disturbances represent all other influences on the latent variables other than the variables in the model that are pointing to the latent variable. To allow the disturbances to be correlated means that we recognize that these other causes may be related. In other words, we recognize that Emotional Adjustment and Achievement may be related in other ways beyond the paths shown in the model, but we're not really sure what these other relations may be. Practically, these correlated disturbances may mean that the two variables are causally related, but we don't know the direction. The correlated disturbances may also mean that there is some other variable, not included in the model, that affects both Adjustment and Achievement (an unmeasured common cause). If you think about it, this correlation means what any correlation may mean: *a* may cause *b*, *b* may cause *a*, or there may be a third variable, *c*, that causes both *a* and *b*. Again, the models are equivalent, so we can't decide which is correct based on the data. As a general rule, however, I prefer to make the causal statement (Figure 17.10) than to be noncommittal (Figure 17.11), but I want a more solid grounding in relevant theory and research than I now have before making the decision of causal direction. We will return to the topic of causal direction in the next chapter.

MODEL MODIFICATIONS

The competing model discussed above was developed based on logic rather than analysis of the detailed fit information. You may wonder, if we had not thought of this competing model, would the modification indexes (MIs) or the standardized residuals (or the correlation residuals) have hinted at it? Figure 17.12 shows the modification indexes greater than 4.0 for the initial model (from Figure 17.6). Although many of the modification indexes do

Modification Indices

Covariances

	M.I.	Par Change
d3 <--> d2	12.075	.086
r7 <--> d2	16.425	.120
r7 <--> r5	4.590	.073
r4 <--> r5	7.147	.101
r4 <--> r6	10.334	-.117

Regression Weights

		M.I.	Par Change
Achievement	<--- Emotional_Adjustment	4.714	.522
Emotional_Adjustment	<--- Achievement	7.159	.046
QUANT	<--- AUTO	5.448	.250
LANG	<--- AUTO	8.107	-.294
LONE	<--- Achievement	9.785	.065
LONE	<--- QUANT	9.925	.041
LONE	<--- LANG	9.972	.046
AUTO	<--- LANG	4.345	-.033

Figure 17.12 Modification indexes for the initial rejection model.

not make a lot of sense, several are worth noting. The largest index suggests that χ^2 could be reduced by at least 16.425 by freeing the correlation–covariance between the residual for Loneliness (r7) and the disturbance for Achievement (d2). This modification makes little sense. The next largest modification index (12.075 for the covariance between d3 and d2) does, however. This MI suggests that the model will fit statistically significantly better if this covariance is freed. Focus on the MIs for the regression weights (the paths). Although they are not the largest MIs, the first two listed also suggest that the fit of the model could be improved by focusing on the relation between Achievement and Emotional Adjustment. Thus, although the modification indexes do not point directly to our Achievement Effect model, they certainly hint in that direction.

Table 17.3 shows the standardized residual covariances and the correlation residuals among the variables. These residuals show that the Initial Model did not adequately account for the correlations between Loneliness and the MRT Quantitative and Language scores and also between Language and School Avoidance. The table of residual correlations also shows that these residuals are substantial. The model predicts a correlation of .144 between the MRT Language test and the Loneliness scale, whereas the actual correlation between these measured variables was .301, a difference of .157 (the actual correlation and the implied correlation are not shown in the table but are easily accessible in the text output from Amos or

Table 17.3 Standardized residual covariances and residuals correlations for the initial rejection model.

Standardized Residual Covariances

	QUANT	LANG	SCHAVOID	LONE	AUTO	COOP	AVE_RAT	NEG_NOM
QUANT	0							
LANG	0	0						
SCHAVOID	.682	1.260	0					
LONE	2.968	3.096	0	0				
AUTO	.321	-1.821	-.008	-1.089	0			
COOP	-.608	-.488	-.128	-.444	.313	0		
AVE_RAT	.006	-.235	.236	-.053	.760	-.286	0	
NEG_NOM	.433	-.104	-.536	.741	.333	-.983	.014	0

Residual Correlations

	QUANT	LANG	SCHAVOID	LONE	AUTO	COOP	AVE_RAT	NEG_NOM
QUANT	0							
LANG	0	0						
SCHAVOID	.035	.064	0					
LONE	.152	.157	0	0				
AUTO	.016	-.093	-.001	-.055	0			
COOP	-.031	-.025	-.007	-.023	.016	0		
AVE_RAT	.001	-.013	.012	-.003	.039	-.015	0	
NEG_NOM	.024	-.005	-.028	.038	.017	-.051	.001	0

other SEM programs). Again, the residuals *might* lead you in the direction of the Achievement Effect model if you had not thought of it previously.

As long as we are cleaning up our models, we might reexamine the statistical significance of the various parameter estimates to see if all paths are statistically significant, with the idea that if any are not it will be okay to remove them. As shown earlier in Figure 17.8, all paths were statistically significant. Although not shown here, all paths are also statistically significant in the Achievement Effect model. It is worth reiterating a previous point: models that are extensively modified based on modification indexes and other tools for model modification should be considered exploratory, tentative models until tested against new data.

SUMMARY

This chapter is the first to focus on latent variable structural equation models. Such SEM models may be considered as a confirmatory factor analysis of the various constructs involved in the research, with a simultaneous path analysis of the effects of these constructs on each other. The chapter reviewed the components of latent variable SEMs and illustrated the methodology with an extended example from the research literature.

Conceptually, you may consider latent variable SEM as a confirmatory factor analysis of the constructs underlying the measured variables in the research, along with a path analysis of the latent variables. The measurement model includes the latent variables, constructs, or factors that underlie the measured variables in the research as causes of these measured variables. The measurement model also includes latent variables, one per measured variable, representing the unique and error variances of each variable, or all other causes of that measured variable other than the construct/latent variable. The structural model includes the paths and covariances among the latent variables, along with the disturbances for the endogenous latent variables (all other causes of the latent variables other than those with arrows pointing to the latent variables). It is often confusing to those new to the SEM methodology to know which variables require latent variables representing disturbances or unique/error variances. At the most mechanical level, any variable that has an arrow pointing to it must also include a latent variable representing all other influences on this variable. For measured variables, these other influences are unique and error variances. For latent variables, these other influences generally represent disturbances along the lines of the disturbances from path analysis or the residuals from multiple regression analysis. In fact, you can, and some methodologists recommend that you do, analyze the model separately as a measurement (confirmatory factor) model, and then add the structural model. We have not used this process here, but it can be useful, especially for complex models or in the beginning stages of research.

The research example used in the chapter was based on research on the effects of peer rejection on kindergarten students' academic and emotional adjustment (Buhs & Ladd, 2001). The example analyzed models similar to (but smaller than) those analyzed in the actual research, with data simulated to mimic the actual data. The initial model included four latent variables with two measured variables indexing each latent variable (more good measures per latent variable are preferable in practice, but our interest was in a smaller, more manageable example). We split apart the measurement model from the structural model for conceptual purposes but not for analysis. The initial model was fairly parsimonious ($15\ df$), with most of the degrees of freedom a result of constraints in the measurement model (undrawn factor loadings from latent to measured variables).

The initial model had an adequate fit to the data and suggested that Rejection by peers resulted in lower subsequent Achievement and school-related Emotional Adjustment. A

portion of these effects were indirect, or mediated, through Class Participation: rejected students had lower rates of participation, which resulted in lower achievement and adjustment. Thus, all three types of effects—direct, indirect, and total—were interesting and interpretable.

An alternative model, which included an additional path from Achievement to Emotional Adjustment, was also estimated. This change resulted in a statistically significant improvement in χ^2 , which we interpreted as meaning that the alternative Achievement Effect model was a better explanation of the data than the initial model. The alternative model led to different interpretations of direct, indirect, and total effects. As an aside, this change (in the structural portion of the model) used up the 1 degree of freedom that was due to the structural portion of the model.

Any complacency we may have garnered that we had now found the correct model was quickly shattered, however. The chapter discussed two alternative models that are equivalent to our preferred Achievement Effect model. Although these two models are statistically indistinguishable from the Achievement Effect model, they have very different interpretations and implications. The chapter included the standardized figural output from one of these alternative models to demonstrate its statistical, but not conceptual, equivalence to the Achievement Effect model. This fuzziness served as another reminder of the importance of theory, logic, and previous research in the construction of models. The equivalent models also served as a reminder of the importance of planning the research so that you can indeed answer the questions of interest.

In the final section of the chapter we examined some of the more detailed fit statistics from the SEM program output. The modification indexes and the standardized residual covariances and correlations for the initial model hinted at the change we made in the Achievement Effect model (although they also suggested the other equivalent, indistinguishable models). Although we might have arrived at the same place had we constructed the alternative Achievement Effect model based on these hints, alternative models devised prior to the examination of the data and results should generally be given more credence than models derived from extensive data-driven model modifications. There were no statistically not-significant paths or factor loadings that we might have constrained in subsequent models.

Although not discussed in detail, there are always equivalent possible models, and their veracity must be tested against these (theory, etc.) standards, not through complex statistical analysis. We can test and reject some models, but we can rarely (maybe never) test and evaluate all possible models that would result in alternative interpretations. Some we don't think of, and some are indistinguishable. At the most basic level, our models always come back to this need for theory, thought, and previous research. “The study of structural equation models can be divided into two parts: the easy part and the hard part” (Duncan, 1975, p. 149). The hard part is developing sound, theory-grounded models. Again, welcome to the dangerous world of SEM.

EXERCISES

1. Analyze the simulated Buhs and Ladd data (“Buhs & Ladd data.sav” or Buhs & Ladd data.xls”) using a structural equation modeling program (if you are using Amos, the initial model is saved as “Buhs & Ladd 1.amw”; the Mplus script is also online).
 - a. Estimate the models discussed in this chapter. Study the parameter estimates and standard errors, the fit statistics, modification indexes, and standardized residuals.
 - b. Interpret the model. Be sure to interpret the indirect and total effects in addition to the direct effects.

- c. Compare the initial model with the competing model discussed in this chapter (the Achievement Effect model). Do you agree that this model is a better alternative? What theoretical, logical, and research evidence can you offer in support of this model? What evidence argues against this model?
- d. Are there other alternative models that you are interested in testing? Do so; be sure to evaluate the relative fit of the model and to interpret your findings.
- e. Are there any common causes that the authors may have neglected? How could you investigate the possibility of unmeasured common causes more completely?
2. Figure 17.13 shows a model to test the effects of participation in Head Start on children's cognitive ability. This example is a classic reanalysis of a controversial quasi-experiment; I have seen variations of it presented in Kenny (1979) and Bentler and Woodward (1978), among others. The measured background variables in the model include measures of mother's and father's educational attainment, father's occupational status, and family income. Head Start was hoped to improve participants' cognitive skills, and the latent Cognitive Ability outcome was indexed by scores on two tests: the Illinois Test of Psycholinguistic Abilities (ITPA) and the Metropolitan Readiness Test (MRT). The Head Start variable is a dummy variable coded 1 for those who participated in Head Start and 0 for children in the control group. The data are shown in Table 17.4. These are data from 303 white children from an early Head Start evaluation, 148 who attended Head Start in the summer and 155 who did not. To understand why the example is so controversial, note the correlation between Head Start and the two cognitive outcomes: both are negative ($-.10, -.09$), suggesting that Head Start may have negative effects on Ability! The model is one of several possible models designed to determine what the outcomes of Head Start are after taking the family's background characteristics into account. The correlations and *SDs* are also included here and in the Excel file "head start.xls" (the *SDs* are not included in most presentations of these data. I estimated these and the means from data presented in Magidson & Sörbom, 1982). All continuous variables are standardized.
- a. Draw (set up) and estimate the model. Is the structural portion of the model just-identified or overidentified? Evaluate the fit of the model and, if adequate, focus on parameter estimates. Interpret the model. According to these results, does Head Start have a positive effect on cognitive ability, a negative effect, or no effect at all? Interpret the other aspects of the model.

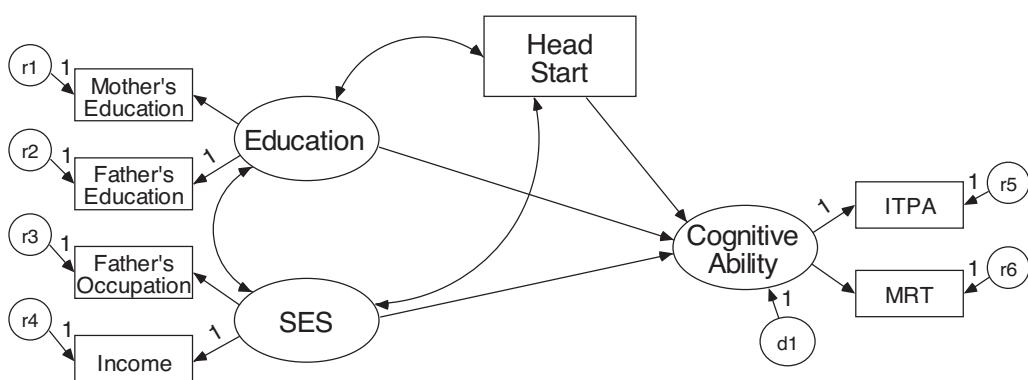


Figure 17.13 Model testing the potential effects of Head Start participation on children's cognitive ability

- b. Fix the path from Head Start to Cognitive Ability to zero; compare the fit of this model to the initial model. Do you still come to the same conclusion as before?
- c. Are there other alternative models that you are interested in testing? Are they equivalent to the initial model? Test these models; be sure to evaluate the relative fit of the model and to interpret your findings.
- d. Are there any common causes that the research may have neglected? How could you investigate the possibility of unmeasured common causes more completely?
3. Kimmo Sorjonen and colleagues used SEM to estimate the relative effects of intelligence, family of origin SES, and emotional capacity (at the time of their conscription into the military) on Swedish men's occupational status at ages approximately 35-40 (Sorjonen, Hemmingsson, Lundin, Falkstedt, & Melin, 2012). The authors were interested in the relative effects of these variables as well as the extent to which their effects were mediated by educational attainment. Figure 17.14 shows the authors' model (minus one correlated error). A dataset of 1000 cases, simulated to give similar findings to the article, are on the website in the file "Sorjonen et al simulated 7.sav" (the actual research had an N of over 48,000). Note that while the simulated data are designed to mimic the means and variances of the original data, I have not been strict in the scaling; thus there are items that have (impossible) negative values. A brief explanation of the variables in the analysis are shown in Table 17.4. Estimate the model shown. Create a table of direct, indirect, and total effects on the final outcome (Attained Occupation). Which variables are the most important influences on these men's eventual occupations? Which variables are less important? Interpret your findings. Is there anything unusual about this model?

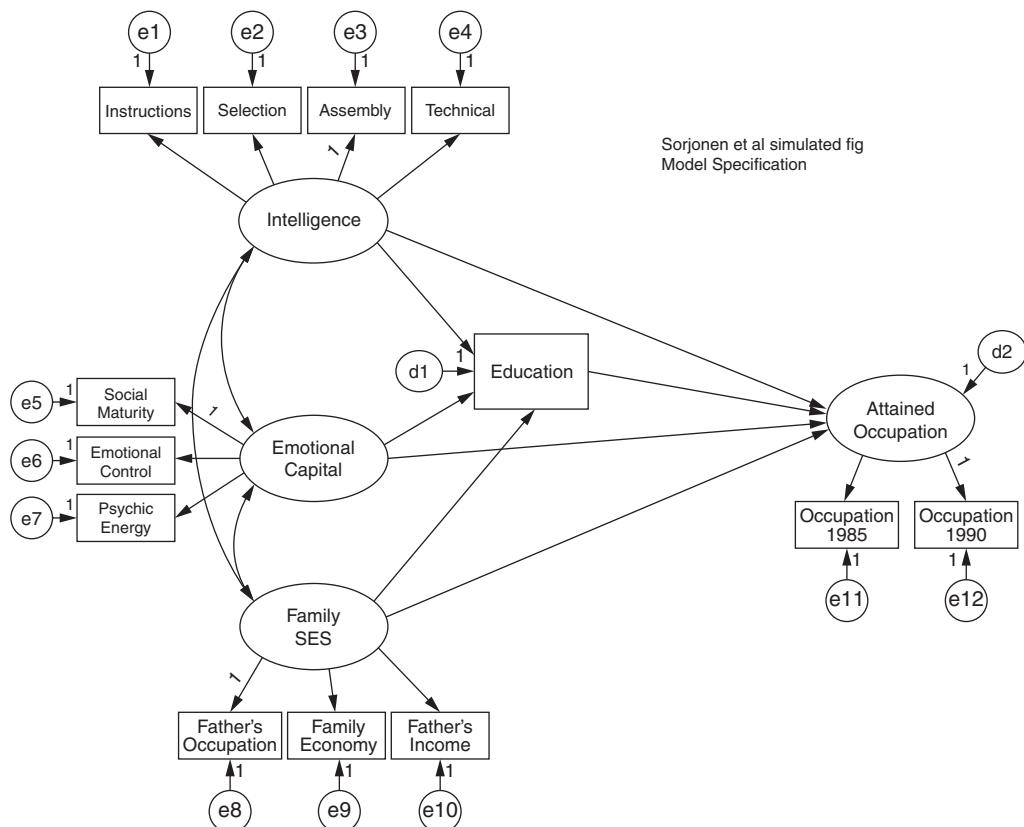


Figure 17.14 Model for the Sorjonen and colleagues (2012) exercise.

Table 17.4 Variables in the Sorjonen et al. (2012) example.

Variable Name	Label in Figure	Description
Instructions		Short measure of verbal intelligence & inductive reasoning
Selection		Short measure of verbal intelligence & inductive reasoning
Assembly		Short measure of visual-spatial reasoning
Technical		Short measure of “mechanical ability” and “technical understanding” (p. 270)
Pop Occ	Father's Occupation	Occupation status on a 5-point scale from census
Fam Economy	Family Economy	Participant's ratings of their family's economic standing from very poor (1) to very good (5), rated in 1969/70 at time of conscription
Pop Income	Father's Income	Natural log of participant's father's income, from census data, for 1970
Maturity	Social Maturity	Psychologist's ratings in 1969/70 irresponsibility and maladjustment versus “responsibility . . . independence, . . . and extraversion” (p. 271)
Control	Emotional Control	Psychologists' ratings of nervousness and anxiousness versus calmness
Energy	Psychic Energy	Psychologists' ratings of a lack on initiative versus initiative and ideas
Occ 85	Occupation 1985	Occupational status from 1985 census
Occ 90	Occupation 1990	Occupational status from 1990 census
Education		Level of education (7 point scale) from 1990 Census data

Note

- 1 Here is the calculation of df for the measurement and structural models: with eight measured variables, there are 36 elements in the variance/covariance matrix: $\frac{p \times (p+1)}{2} = \frac{8 \times 9}{2} = 36$. For the measurement model we estimate 22 parameters: 6 correlations/covariances among the factors, 4 factor loadings (recall that for each factor one factor loading is set to one to set the scale), 4 factor variances, and 8 unique/error variances (r1 through r8). $36 - 22 = 14 df$ for the measurement model. For the full latent variable SEM we are estimating 21 parameters: the 8 unique/error variances, 4 factor loadings, 1 factor variance (for the exogenous variable, Rejection) and 3 variances of disturbances (d1 through d3), and 5 paths. For this model, $36 - 21 = 15 df$. Another way of thinking about df is to apportion them to the measurement versus structural models (Figures 17.3 versus 17.4). As already calculated, the measurement model accounted for 14 df . In the structural model, the 6 factor correlations are replaced by 5 paths, resulting in 1 additional df .

18

Latent Variable Models II

Multigroup Models, Panel Models, Dangers and Assumptions

Single Indicators and Correlated Errors	409
<i>A Latent Variable Homework Model</i>	409
<i>Competing Models</i>	419
<i>Model Modifications</i>	421
Latent Variable Panel Models	424
MultiGroup Models	426
<i>A MultiGroup Homework Model Across Ethnic Groups</i>	426
Dangers, Revisited	435
<i>Omitted Common Causes</i>	435
<i>Path in the Wrong Direction</i>	437
Summary	438
Exercises	439
<i>Notes</i>	443

In the previous chapter we introduced and explored latent variable structural equation models. This chapter will review and consolidate that learning by reviewing another example. We will continue our exploration with several additional topics and an assessment of where we stand in our efforts to conduct meaningful nonexperimental research. The chapter will begin with a model that incorporates two complexities that we have touched on previously: single-indicator variables and correlated errors.

SINGLE INDICATORS AND CORRELATED ERRORS

A Latent Variable Homework Model

Figure 18.1 shows a latent variable version of our earlier Homework model from Chapter 14. The primary variables in the model are Homework, indexed by student reports of average time spent on homework in 8th (Homework 8th) and 10th (Homework 10th) grades, and students' overall Grades in high school, a latent variable estimated by students' high school GPAs in English, Math, Science, and History–Social Studies

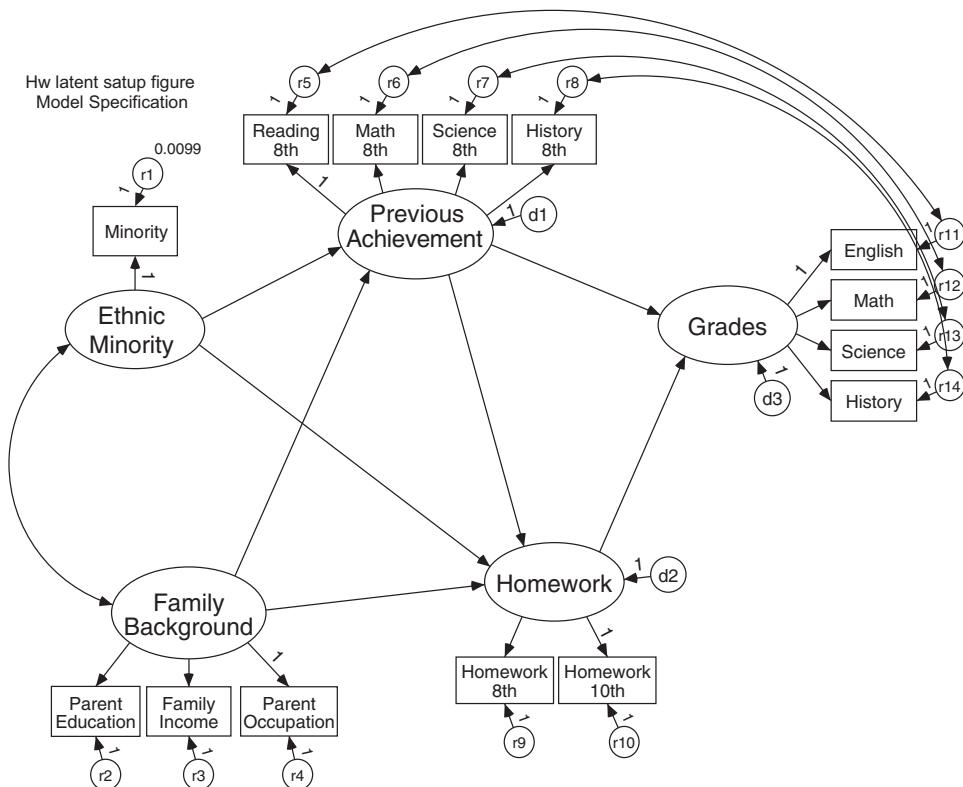


Figure 18.1 Latent variable model of the effects of Homework on High School GPA.

(from students' transcripts at graduation). Other measured variables in the model were as follows:

1. Achievement test scores from 8th grade in Reading, Math, Science, and History–Social Studies (Previous Achievement)
2. Parent Educational attainment, Family Income, and Parent Occupational status (Family Background). These variables were generally taken from the parent file; Parent Occupation and Parent Education were each based on the higher value reported for either the father or the mother.
3. Ethnic background (Minority), coded 1 for ethnic minority group members and 0 for white.

Recall that with the earlier homework model most of the variables were composites of some sort; Previous Achievement, for example, was a composite of the four 8th-grade achievement tests. In the current model, these components were not added together as composites but appear in the model as measured indicators of latent variables. Instead of adding the four tests together to create a Previous Achievement *composite* variable, for example, the four 8th-grade tests are used as indicators of a Previous Achievement *latent* variable.

The model will be estimated from the covariance matrix of the measured variables. The covariance matrix is recovered from the correlation matrix and standard deviations, shown in Table 18.1, and on the accompanying Web site (www.tzkeith.com) under the label "hw latent matrix.sav" or "hw latent matrix.xls". The variable names in the file are the variable names from the Amos model (rather than the variable labels as shown in the Figure); these should either be familiar to you or self-explanatory. The data are 1000 cases chosen at random from

Table 18.1 Correlations, Means, and Standard Deviations Among Measured Variables in the Latent Variable Homework Model

Variable	Minority	bypared	byfaminc	parocc	bytxrstd	bytxnsid	bytxhstd	hw_8	hw10	eng_12	math_12	sci_12	ss_12
Minority	1												
bypared	-.169	1.000											
byfaminc	-.278	.526	1.000										
parocc	-.242	.629	.524	1.000									
bytxrstd	-.204	.386	.288	.339	1.000								
bytxmstd	-.161	.430	.335	.362	.714	1.000							
bytxsstd	-.231	.384	.293	.322	.717	.719	1.000						
bytxhstd	-.210	.396	.308	.346	.731	.675	.728	1.000					
hw_8	-.003	.168	.075	.105	.226	.271	.221	.168	1.000				
hw10	-.056	.208	.155	.173	.219	.286	.206	.207	.271	1.000			
eng_12	-.098	.334	.243	.260	.524	.565	.450	.491	.204	.313	1.000		
math_12	-.071	.285	.220	.218	.418	.587	.415	.409	.173	.289	.761	1.000	
sci_12	-.083	.294	.209	.231	.484	.576	.493	.476	.192	.282	.803	.759	1.000
ss_12	-.111	.328	.253	.265	.519	.567	.485	.519	.181	.284	.851	.745	.795 1.000
SD	.445	1.284	2.523	21.599	10.290	10.380	10.318	10.182	1.131	1.903	2.674	2.747	2.682 2.873
Mean	.272	3.203	9.917	51.694	51.984	52.545	51.883	51.653	1.731	3.381	6.250	5.703	5.952 6.418

the 8th- through 12th-grade NELS data, including information from students' transcripts. For the current model, I have symbolized the unique–error variances of the measured variables as r1 through r14 and the disturbances of the latent variables as d1 through d3.¹

Note that each latent variable has its scale set by a single factor loading (path from the latent to measured variable) set to 1 (ULI). Each error–unique (residual) variance has its scale set by constraining the path from it to its corresponding measured variable to 1.

The model examines the effect of time spent on homework on subsequent GPA while controlling for students' previous school performance. Two background variables, Ethnic Minority background and Family Background, are also controlled, although the model specifies that both background variables affect Grades only indirectly through Previous Achievement and Homework. Note that the model is simply a latent variable version of the path model from Chapter 14 and, like that model, is supported by theory and previous research.

Single-Indicator Latent Variable

The model included several less common characteristics, as well. First, notice the value associated with the residual (r1) of the Minority variable (.0099). The latent variable Ethnic Minority is indexed by a single measured variable (Minority), and this portion of the measurement model would be underidentified without further constraints. As discussed in Chapter 16, a common method for dealing with single-indicator factors is to constrain the error–unique variance of that measured variable to some value, often a value of 1 minus the estimated reliability of the measured variable. Why, you may ask, would a variable as clear-cut as ethnic background be unreliable? Students' reports of their ethnic identity should be very, but not completely, reliable. Students may misread the questionnaire item or might decide on a whim to mark it incorrectly. Students of mixed racial/ethnic backgrounds can only chose one group when they belong to more than one. Those who enter the data into the computer may make transcription errors. All these possibilities add small amounts of error. For these reasons, I estimated the reliability of the Minority variable at approximately .95. Thus 5% of the variability of the Minority measured variable is due to unreliability, or error. The variance of Minority is .198 (from Table 18.1, $SD^2 = .445^2 = .198$), and 5% of this variance is .0099; the error variance of the Minority variable was constrained to this value. (Note that I am using the term reliability here quite loosely. Strictly speaking, unreliability refers to random error, whereas my examples include random and systematic error. Our estimates of error variance in SEM often include both.)

Correlated Errors

The model also includes correlations between the error and unique variances of the Achievement test scores and later Grades. The model, for example, specifies that the unique and error variance of the 8th-grade Math achievement test score is correlated with the unique and error variance of the 12th-grade Math GPA. Conceptually, this correlated error means that we believe that the Math test score and Math Grades share something in common above and beyond the effect of general Previous Achievement on overall Grades. If you think about it, this makes sense, and we can even label that "something" that Math test scores and Math grades share in common: specific Math achievement. The model also includes correlated errors between Reading–English, Science test and grades, and History–Social studies test and grades. Such correlated errors are common in longitudinal models in which a single measure is administered more than once or when closely related measures are administered at two different times (as in the present model). Indeed, the ability to

take the possibility of correlated unique and error variances into account is an important advantage of SEM.

The full SEM model (Figure 18.1) has 66 degrees of freedom. Sixty-four degrees of freedom are from the measurement portion of the model. Simply note all the factor loadings that could be included in the model that are not included (e.g., a path from Homework to Reading 8th or Parent Occupation); these constraints are the source of this $64\ df$. The structural model produces the other $2\ df$, resulting from the paths from Ethnic Minority and Family Background to Grades that are constrained to zero. A little later in the chapter we will estimate the measurement model separately and then add in the structural model.

Results

The model (Figure 18.1 and in the file “hw latent 1.amw”) and the data (Table 18.1 and the file “hw latent matrix.sav”) were analyzed via Amos. Figure 18.2 shows relevant fit indexes, along with the standardized output. The model showed a good fit to the data. The RMSEA was below .05 (.046, 90% confidence interval = .039–.053), and the TLI was above .95. The SRMR for this model was .029, meaning that the matrix implied by the model differed from the actual correlation matrix, on average, by only .029. The full array of fit indexes is shown in Figure 18.3. (Now might be a good time to review the suggestions for fit indexes and their evaluation in Chapter 14.) Because the model generally fits well, we’ll first interpret the results. Later in the chapter we will take a look at the more detailed fit information to see how the model might be modified.

Figure 18.4 shows more detail concerning the paths and factor loadings, including the unstandardized coefficients, their standard errors, and critical ratios (z statistics). All the

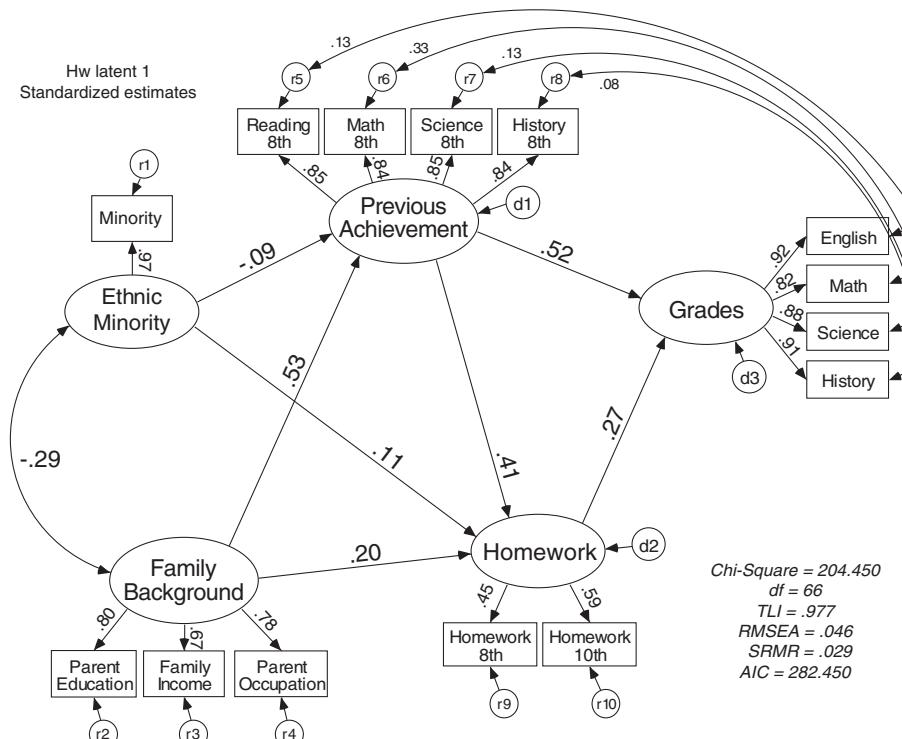


Figure 18.2 Standardized output for the latent variable homework model.

Model Fit Summary***CMIN***

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	39	204.450	66	.000	3.098
Saturated model	105	.000	0		
Independence model	14	8383.652	91	.000	92.128

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	1.389	.972	.955	.611
Saturated model	.000	1.000		
Independence model	24.569	.310	.204	.269

Baseline Comparisons

Model	NFI	RFI	IFI	TLI	CFI
	Delta1	rho1	Delta2	rho2	
Default model	.976	.966	.983	.977	.983
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

Arsimony Adjusted Measures

Model	PRATI	PNFI	PCFI
Default model	.725	.708	.713
Saturated model	.000	.000	.000
Independence model	1.000	.000	.000

FMIN

Model	FMIN	F0	L ₉₀	I ₉₀
Default model	.205	.139	.099	.186
Saturated model	.000	.000	.000	.000
Independence model	8.392	8.301	8.003	8.605

RMS₀₀

Model	RMS _{0A}	L ₉₀	I ₉₀	PCL _S
Default model	.046	.039	.053	.825
Independence model	.302	.297	.308	.000

IC

Model	AIC	CC	IC	CAIC
Default model	282.450	283.639	473.852	512.852
Saturated model	210.000	213.201	725.314	830.314
Independence model	8411.652	8412.079	8480.361	8494.361

Figure 18.3 Fit indexes for the initial homework model.

parameters that were estimated were statistically significant (z greater than approximately 2). Figure 18.5 shows the covariances, correlations, and variances. Note that covariances were also statistically significant, with the exception of the covariance between r8 and r14. The correlated error between 8th-grade History test scores and 12th-grade History grades was not statistically significant; we could, if desired, remove this parameter in subsequent models, presumably without any noticeable loss of fit.

Regression Weights

			Estimate	S.E.	C.R.	P
Previous_Achievement	<---	Family_Background	.278	.020	13.649	***
Previous_Achievement	<---	Ethnic_Minority	-1.774	.646	-2.748	.006
Homework	<---	Family_Background	.013	.004	3.120	.002
Homework	<---	Previous_Achievement	.053	.008	6.640	***
Homework	<---	Ethnic_Minority	.281	.123	2.292	.022
Grades	<---	Previous_Achievement	.145	.012	12.574	***
Grades	<---	Homework	.601	.132	4.566	***
Minority	<---	Ethnic_Minority	1.000			
parocc	<---	Family_Background	1.000			
byfaminc	<---	Family_Background	.100	.005	19.214	***
bypared	<---	Family_Background	.062	.003	21.607	***
bytxrstd	<---	Previous_Achievement	1.000			
bytxmstd	<---	Previous_Achievement	.997	.030	33.737	***
bytxsstd	<---	Previous_Achievement	.990	.030	33.520	***
bytxhstd	<---	Previous_Achievement	.967	.029	32.909	***
eng_12	<---	Grades	1.000			
Math_12	<---	Grades	.896	.024	37.813	***
Sci_12	<---	Grades	.957	.022	43.683	***
ss_12	<---	Grades	1.062	.022	48.194	***
hw10	<---	Homework	1.000			
hw_8	<---	Homework	.453	.060	7.549	***

Standardized Regression Weights

			Estimate
Previous_Achievement	<---	Family_Background	.529
Previous_Achievement	<---	Ethnic_Minority	-.087
Homework	<---	Family_Background	.198
Homework	<---	Previous_Achievement	.413
Homework	<---	Ethnic_Minority	.108
Grades	<---	Previous_Achievement	.518
Grades	<---	Homework	.274
Minority	<---	Ethnic_Minority	.975
parocc	<---	Family_Background	.776
byfaminc	<---	Family_Background	.667
bypared	<---	Family_Background	.805
bytxrstd	<---	Previous_Achievement	.855
bytxmstd	<---	Previous_Achievement	.844
bytxsstd	<---	Previous_Achievement	.846
bytxhstd	<---	Previous_Achievement	.837
eng_12	<---	Grades	.924
Math_12	<---	Grades	.820
Sci_12	<---	Grades	.878
ss_12	<---	Grades	.914
hw10	<---	Homework	.592
hw_8	<---	Homework	.451

Figure 18.4 Unstandardized and standardized factor loadings and paths for the initial latent variable homework model.

Covariances

			Estimate	S.E.	C.R.	P
Family_Background	<-->	Ethnic_Minority	-2.136	.277	-7.705	***
r5	<-->	r11	.704	.248	2.842	.004
r6	<-->	r12	2.856	.342	8.346	***
r7	<-->	r13	.920	.285	3.225	.001
r8	<-->	r14	.533	.277	1.926	.054

Correlations

			Estimate
Family_Background	<-->	Ethnic_Minority	-.294
r5	<-->	r11	.128
r6	<-->	r12	.331
r7	<-->	r13	.130
r8	<-->	r14	.082

Variances

	Estimate	S.E.	C.R.	P
Family_Background	280.913	21.615	12.996	***
Ethnic_Minority	.188	.009	21.232	***
d1	53.163	3.516	15.122	***
d2	.915	.182	5.017	***
d3	3.150	.202	15.596	***
r1	.010			
r4	185.127	13.024	14.215	***
r3	3.528	.193	18.267	***
r2	.580	.046	12.691	***
r5	28.555	1.718	16.617	***
r6	31.233	1.822	17.145	***
r7	30.165	1.768	17.059	***
r8	30.864	1.775	17.390	***
r11	1.052	.075	14.108	***
r12	2.378	.122	19.536	***
r13	1.653	.094	17.617	***
r14	1.364	.090	15.125	***
r10	2.352	.202	11.633	***
r9	1.018	.058	17.553	***

Figure 18.5 Covariances, correlations, and variances for parameters estimated in the initial homework model.

Interpretation

Let's now focus on the meaning of the results. First, our primary interest was in the effects of Homework on GPA. As already noted, this effect was statistically significant (see the Homework --> Grades path in Figure 18.4). The standardized coefficient was .27, meaning that for each SD change in the latent Homework variable Grades should change by .27 of a standard deviation, other things being equal. This finding, in turn, suggests a strong effect of time spent on homework on subsequent GPA (given the adequacy of the model). This effect is larger than in our previous path analyses using only measured variables (even though we are focusing on a longer time span—through 12th grade, rather than 10th grade) and larger than the effect shown in Part 1 when we examined the effect of homework on learning using

multiple regression. As noted in Chapter 15, our measures of variables in research are always error laden. Latent variable SEM removes unreliability and invalidity from the estimates of the effects of one variable on another. The most common effect of removing measurement error from our estimation process is to increase the apparent effect of one variable on another. This effect is illustrated well by comparing the present homework model with previous versions. The current, latent variable model is a more accurate representation of the true effects of homework on learning, because it gets closer to the level of the constructs of true interest.

Unstandardized Coefficients

Focus on the unstandardized coefficients (Figure 18.6). The unstandardized effect of Homework on Grades was .60, meaning that for each 1-unit change in the latent homework variable Grades increase .60 point. To understand the meaning of this statement, we need to understand the scales involved. The Homework latent variable was set to have the same scale as the measured Homework 10th variable, whereas the Grades latent variable was set to the same scale as the English GPA measured variable. If Homework 10th had been measured on a simple hour scale and English GPA on a standard 4.0 scale, interpretation would be relatively straightforward. Unfortunately, the underlying scales for both variables are not that meaningful, which is one reason I am focusing more on the interpretation of standardized as opposed to unstandardized coefficients. The Homework 10th measured variable was a mean of two questions, F1S36A1 and F1S36A2 (average time spent on homework in school and time spent on homework out of school). I changed the scale of each of these items so

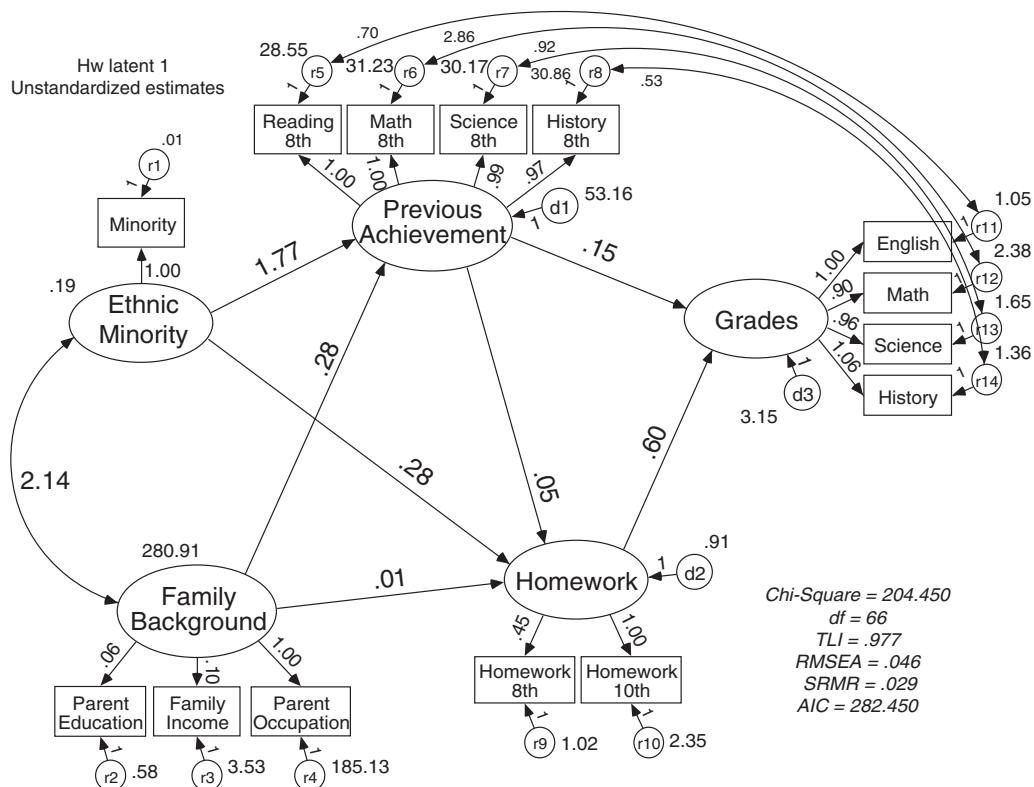


Figure 18.6 Unstandardized output for the initial homework model.

that they ranged from 0 (none) to 9 (over 15 hours a week; I changed the scale from a 0 to 7 scale so that it would be consistent with the scale used for the 8th-grade homework question). The homework scales in NELS are presumably designed to take into account the curvilinear nature of the effect of homework on learning. The English GPA scale ranged from 0 (an F average) to 12 (A+). Again, the unstandardized coefficients are less interpretable than would be ideal.

Effects on Homework, Indirect and Total Effects

Many more interesting findings are contained in the model, as well. The analysis has shown that Homework affects Grades, but this raises another question. Which other variables in the model affect Homework? That is, who spends more time on homework? Previous Achievement had a strong effect (.41, standardized) on Homework. Students who achieve at a higher level spend more time on homework than those who achieve at lower levels; this increase in Homework time subsequently results in higher Grades as well. The coefficients from Family Background suggest that students from more advantaged backgrounds have higher 8th-grade achievement (.53) and complete more homework (.20). Students' Ethnic background had only a small effect on 8th-grade Achievement (-.09). Ethnic Minority had a positive effect on Homework (.11). Given the coding of the Minority variable (1=Minority, 0=White), this means that students from ethnic minority backgrounds report higher levels of homework time than White students. The unstandardized coefficient for the Minority-Homework path (.28) shows that minority students report .28 points higher on the Homework time scale than do White students when the other variables in the model are controlled.

Figure 18.7 shows the standardized indirect and total (as well as the direct) effects of the latent variables on each other. Note that, because there are no paths from Ethnic Minority or Family Background to Grades, there are, of course, no direct effects for these variables on Grades. Family Background, however, had a large indirect effect on Grades (.388), primarily through its effect on Previous Achievement (.529 times the total effect of Previous

Standardized Total Effects

	Ethnic_Minority	Family_Background	Previous_Achievement	Homework	Grades
Previous_Achievement	-.087	.529	.000	.000	.000
Homework	.072	.417	.413	.000	.000
Grades	-.025	.388	.631	.274	.000

Standardized Direct Effects

	Ethnic_Minority	Family_Background	Previous_Achievement	Homework	Grades
Previous_Achievement	-.087	.529	.000	.000	.000
Homework	.108	.198	.413	.000	.000
Grades	.000	.000	.518	.274	.000

Standardized Indirect Effects

	Ethnic_Minority	Family_Background	Previous_Achievement	Homework	Grades
Previous_Achievement	.000	.000	.000	.000	.000
Homework	-.036	.219	.000	.000	.000
Grades	-.025	.388	.113	.000	.000

Figure 18.7 Standardized direct, indirect, and total effects for the initial homework model.

Achievement on Grades, $.631 = .334$). The indirect effect of Family Background on Grades through Homework was smaller but still meaningful ($.198 \times .274 = .054$). Because there are no direct effects of Family Background on Grades, the total effects are the same as the indirect effects. In contrast, the total effect of Ethnic Minority on Grades is very small ($-.025$). It was not statistically significant (when bootstrapping was used to estimate standard errors of indirect and total effects), and I would probably consider it nonmeaningful even if it were statistically significant. The reason this effect is so small is that the negative indirect effect of Ethnic Minority through Previous Achievement is cancelled out by the positive indirect effect through Homework.

With the use of a single indicator accounting for the likely error in the measured variable, the latent variable Ethnic Minority behaves like all the other latent variables in the model. Note that the loading of Minority on Ethnic Minority (standardized) was $.97$ (Figure 18.2). This value is simply a function of the reliability estimate used to fix the error variance (the standardized loading is equal to $\sqrt{r_{tt}}$, with r_{tt} equal to the estimate of the reliability used to constrain the error). Another option is to simply have Ethnic appear as a measured, rather than latent variable, an option that does not recognize the error inherent in the variable. As a result of building error into this variable, the standardized estimates of the effects from Ethnic Minority are slightly larger than they would have been without recognition of this error.

Note that I simply made an educated guess as to the likely reliability of the Minority variable. If reliability estimates are available for a variable, or a similar variable, use them, but sometimes a guess is the best you can do. In such cases, it may be worthwhile to try different values for the reliability (e.g., $.90$ or $.98$ versus $.95$) to make sure you have a good understanding of what happens to the parameters of interest when you make these changes.

If you return to Figure 18.2 and focus on the correlated errors, you will see that the correlated error between the Math test and subsequent Math grades is substantial ($.33$), suggesting that these measures indeed share something in common (specific math achievement) above and beyond the effect of general achievement on overall grades. The other correlated errors are smaller, but all except one (History 8th–History) are statistically significant. The expectation for the existence of correlated errors is probably reasonable.

Competing Models

We may wonder if, indeed, Ethnic Minority and Family Background *really* only affect Grades indirectly, only through Previous Achievement and Homework. We could test this hypothesis by comparing the initial model to one in which paths are estimated from Ethnic Minority and Family Background to Grades. The fit statistics for this model are shown in Table 18.2 under the label Direct Background Effects. The model in which the background variables affect Grades directly results in a smaller χ^2 (202.263), but the change in χ^2 is not statistically significant ($\Delta\chi^2 = 2.187$, $df = 2$, $p = .335$). When $\Delta\chi^2$ is not statistically significant, our rule of thumb is to prefer the more parsimonious model, which is the initial model in Figure 18.2. In other words, yes, it appears that Ethnic Minority and Family Background only affect Grades indirectly, not directly (and the effect of Ethnic Minority was nonsignificant).

Was the assumption that the error–unique variances are correlated across similar tests and grades really necessary? To test the veracity of this assumption, we can delete these correlated errors from the model and compare the fit of this No Correlated Errors model with the initial model. This new model is more parsimonious than the initial model because it includes four fewer parameters to be estimated (the four correlated errors), so if the models fit equally well, we would prefer the more parsimonious No Correlated Errors model. The fit indexes for the model are also shown in Table 18.2. The deletion of the correlated errors resulted in a $\Delta\chi^2$ of 114.962 , with four df . This increase in χ^2 is statistically significant, meaning that the No Correlated Errors model, although more parsimonious, resulted in a statistically significantly worse fit to the data than did the

Table 18.2 Comparison of Fit of Alternative Homework Models

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	TLI	SRMR	RMSEA	AIC
1. Initial	204.450	66				.977	.029	.046	282.450
2. Direct Background Effects	202.263	64	2.187	2	.335	.976	.029	.047	284.263
3. No Correlated Errors	319.412	70	114.962	4	<.001	.961	.031	.060	389.412
4. No Homework Effects	235.867	67	31.417	1	<.001	.972	.039	.050	311.867
5. Measurement Model	202.263	64	2.187	2	.335	.976	.029	.047	284.263
6. Ethnic Minority Measured	204.450	66				.029	.046	.046	282.450

Note: All models compared to the initial model.

initial model. Our rule of thumb is that if the $\Delta\chi^2$ is statistically significant we prefer the less parsimonious model. The increase in parsimony is not worth the cost of additional misfit; the initial model appears a better representation of the effect of these variables on Grades; the correlated errors are needed. [It is also worth noting that the RMSEA for this model is not particularly good, and even its 90% CI (.053–.066) does not include our cut-off (.05) for a good model. We might have rejected this model even based on the RMSEA used as a stand-alone fit index.]

We may also want to test directly the statistical significance of the effect of Homework on Grades by comparing the initial model to one in which the path from Homework to Grades is set to zero. The previous two competing models essentially tested assumptions underlying the initial model, whereas this competing model tests the substantive research question guiding the research: whether homework affects high school grades. The fit of this model is also summarized in Table 18.2. When this No Homework Effect model is compared to the initial model, the $\Delta\chi^2$ is 31.417 ($df = 1$, $p < .001$). Although the No Homework Effect model is more parsimonious, the parsimony (the extra df) resulted in too great a cost in model fit; the $\Delta\chi^2$ increase is statistically significant. Yes, Homework indeed has a strong and statistically significant effect on students' high school GPA. Of course, we would come to this same conclusion through examination of the statistical significance of the Homework to Grades path in the original model (Figure 18.4), but as long as we are testing other aspects of the model using the fit indexes, it makes sense to test this one as well.

Fit indexes for two additional models are shown in Table 18.2. Model 5 shows the fit for a measurement model, that is, a CFA model in which the measured variables are loaded onto the same factors shown in Figure 18.2, but the factors themselves are simply allowed to correlate with one another. As noted earlier, this type of model is often tested prior to the full structural model and the fit compared to such a model. As shown in the Table, the difference in fit between this measurement model and the initial model is trivial and non-significant, and thus we would likely decide that the full SEM model was a reasonable one. Note also that the fit of this model is identical to that of the Direct Background Effects model 2. Make sure you understand why this is the case.

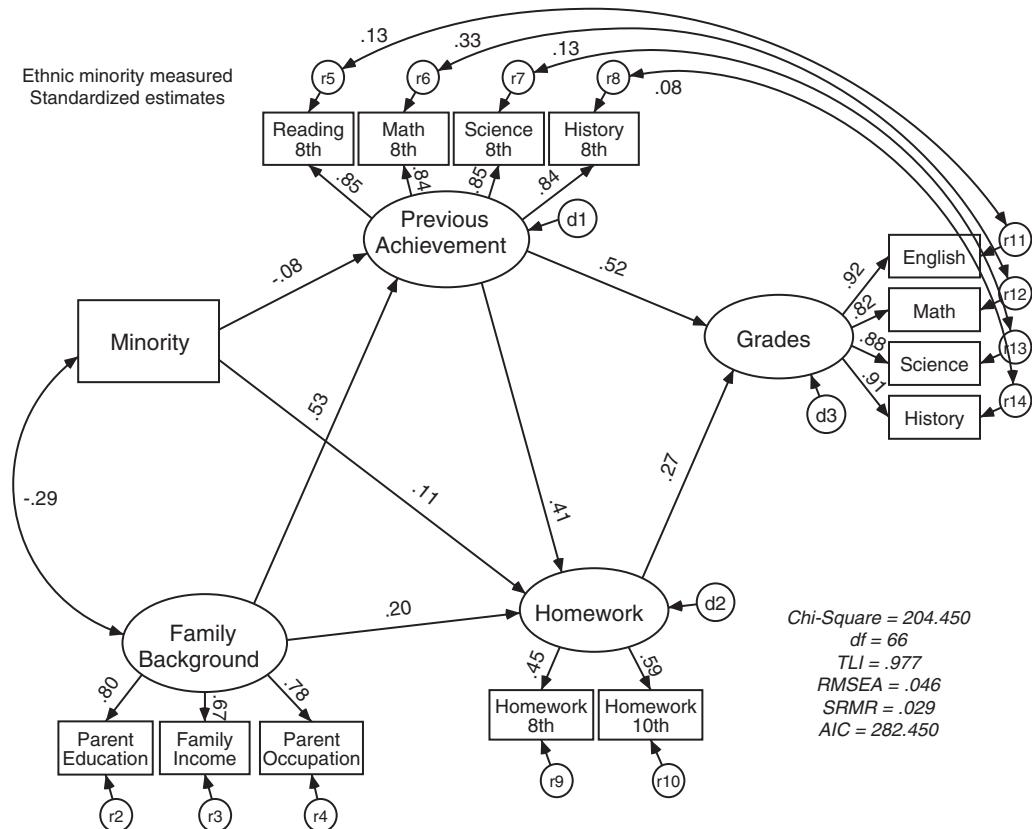


Figure 18.8 Latent variable homework model, with Ethnicity as a measured, rather than a single-indicator latent variable (standardized output).

The final model shown in the Table replaced the single indicator Ethnic Minority latent variable with the measured Minority variable. The standardized results for this model are also shown in Figure 18.8. Note that replacing the latent single-indicator variable with the simple measured variable did not change the fit of the model at all. The purpose of using a single indicator latent is not to improve fit but to obtain more accurate estimates of effects in the model. The fact that standardized path estimates changed very little from one version of this model to the other is because our estimate for the error variance for Minority was quite small (and our estimate of reliability quite large).

Model Modifications

Should we consider other, post hoc model modifications? One possible modification was already discussed: constraining the correlated error between History test scores and History GPA to zero, thus specifying that these error and unique variances are not correlated. This additional constraint results in a worse fit of the model (increase in χ^2), but this change will likely not be statistically significant.

We may wonder if there are model modifications we can make that will *improve* the fit of the model. Because our initial model fit well, this change may have lower priority than it would have if the model did not fit well; but it is still worth exploring if for no other reason than to reinforce the concepts presented in previous chapters. Figure 18.9 shows the modification indexes greater than 4 (not all are listed for space reasons). The largest modification index is for the covariance between r6 and d2 and suggests that χ^2 can be lowered by at least 28.05 by freeing the correlation

Covariances

	M.I.	Par Change
r9 <--> d3	5.461	-.155
r14 <--> d2	4.042	-.133
r12 <--> r13	10.441	.222
r8 <--> d2	6.712	-.769
r8 <--> r9	6.483	-.531
r7 <--> Ethnic_Minority	6.788	-.216
r7 <--> d3	10.663	-1.230
r7 <--> r11	11.075	-.794
r7 <--> r8	6.160	2.866
r6 <--> Ethnic_Minority	9.522	.251
r6 <--> Family_Background	8.285	9.729
r6 <--> d2	28.049	1.526
r6 <--> d3	21.132	1.688
r6 <--> r9	7.955	.571
r6 <--> r13	6.817	.697
r6 <--> r8	13.541	-4.148
r2 <--> Ethnic_Minority	17.664	.053
r2 <--> d1	5.900	.564
r3 <--> Ethnic_Minority	17.316	<u>-.115</u>
r1 <--> r7	5.743	-.197
r1 <--> r6	6.493	.205
r1 <--> r2	18.380	.053
r1 <--> r3	17.347	-.114

Regression Weights

	M.I.	Par Change
hw_8 <--- bytxmstd	4.546	.007
Sci_12 <--- Math_12	5.365	.038
Sci_12 <--- bytxmstd	5.329	.010
eng_12 <--- bytxsstd	6.293	-.010
bytxhstd <--- hw_8	7.752	-.479
bytxsstd <--- Ethnic_Minority	4.822	-1.004
bytxsstd <--- Grades	6.760	-.209
bytxsstd <--- Sci_12	4.924	-.159
bytxsstd <--- eng_12	11.603	-.246
bytxsstd <--- Minority	4.891	-.961
bytxmstd <--- Ethnic_Minority	5.668	1.065
bytxmstd <--- Homework	20.912	1.024
bytxmstd <--- Grades	16.586	.321
bytxmstd <--- hw_8	15.357	.655
bytxmstd <--- hw10	9.557	.307
bytxmstd <--- ss_12	11.112	.219
bytxmstd <--- Sci_12	21.287	.324
bytxmstd <--- Math_12	10.536	.228
bytxmstd <--- eng_12	16.159	.284
bytxmstd <--- bypassed	5.130	.334
bytxmstd <--- Minority	5.737	1.018
bypassed <--- Ethnic_Minority	15.925	.277
bypassed <--- Homework	5.914	.084
bypassed <--- Grades	4.378	.026
bypassed <--- hw_8	4.089	.053
bypassed <--- eng_12	5.345	.025
bypassed <--- bytxmstd	4.416	.006
bypassed <--- Minority	16.129	.265
byfaminc <--- Ethnic_Minority	15.531	-.598
byfaminc <--- Minority	15.698	-.572
Minority <--- bypassed	4.393	.022
Minority <--- byfaminc	8.266	-.015

Figure 18.9 Modification indexes for the initial homework model.

between the error of measurement for the 8th-grade Math test and the disturbance for Homework. Allowing such a change suggests that the Math Test and Homework share something in common, or have a common cause, other than those shown in the model. Although we could probably think up all sorts of reasons why this might be if we tried hard enough, there is no real theoretical or research-based reason to allow such a correlation. Another possibility is to free the covariance between the Math test unique-error variance and the disturbance for Grades (modification index = 21.132), but this change also makes little sense. In a related fashion, the modification indexes for regression weights suggest that we consider allowing a path from Homework to Math Achievement test scores (20.912) or from Science grades to Math test scores (21.287). Again, these modification indexes make little sense, other than to suggest that the 8th-grade Math test score seems to be a general source of misfit in the model.

Figure 18.10 shows the standardized residual covariances, one of our other methods of isolating sources of misfit in the model. There are no especially large standardized residuals, which is consistent with our overall satisfaction with the model fit. If we arbitrarily pick a value of +2 as representing a larger standardized residual, there are three large values in the matrix. Consistent with the speculation that the Math test score is something of a source of misfit, two of these

Standardized Residual Covariances

	hw_8	hw10	ss_12	Sci_12	Math_12	eng_12	bytxhstd	bytxmstd	bytxsstd	bytxrstd	bypared	byfaminc	parocc	Minority
hw_8	.000													
hw10	.136	.000												
ss_12	-1.180	-.105	-.027											
Sci_12	-.580	.136	-.284	-.176										
Math_12	-.623	1.108	.212	1.243	.813									
eng_12	-.531	.681	.141	-.303	.399	-.045								
byrxhstd	-.618	-1.204	.014	-.180	-.919	-.424	.051							
bytxsstd	.969	-1.310	-.582	-.781	-.878	-1.714	.537	.038						
bytxmstd	2.508	1.122	1.704	2.482	1.129	1.509	-.846	.095	-.149					
bytxrstd	1.047	-.997	.216	-.280	-.959	-.530	.373	-.162	-.265	-.067				
bypared	.768	.619	1.130	.391	.894	1.206	.685	.190	1.551	.105	.000			
byfaminc	-1.382	-.031	.346	-.721	.237	-.038	-.055	-.606	.639	-.854	-.318	.000		
parocc	-1.037	-.263	-.490	-1.202	-.915	-.735	-.427	-1.259	-.064	-.891	.105	.165	.000	
Minority	.598	-.851	.421	1.147	1.233	.871	-.378	-.960	1.223	-.041	1.908	-2.684	-.590	.000

Figure 18.10 Standardized residuals for the homework model.

large values are with the 8th-grade Math test (bytxmstd). They suggest that the model does not adequately account for the correlation between the math test and 8th-grade homework time (hw_8) or 12th-grade science GPA. They were also the largest correlation residuals (not shown here), but they were still small, with the largest between bytxmstd and 12-grade science GPA. The third large standardized residual (and correlation residual) suggests that the model does not completely account for the correlation between the Minority measured variable and Family Income. Again, although we can likely think of reasons why this may be so, we are not slapping ourselves on the head, thinking “I can’t believe I did not think of that before!” There are no really compelling reasons to relax any of the constraints in the model to improve the fit.

Latent Variable Panel Models

In Chapter 14 we examined panel models as one type of model appropriate for longitudinal data. Now that we have developed an understanding of latent variables and correlated errors, we can examine this topic in a little more depth.

Figure 18.11 shows a latent variable panel model designed to determine the longitudinal effects of achievement on locus of control and of locus of control on achievement. Recall from Part 1 what locus of control means: this is the degree to which people believe that they control what happens to them (an internal locus and a high score) versus they believe their lives are controlled by outside forces (an external locus and a low score). It makes sense that students with an internal locus of control would achieve at a higher level, perhaps as the

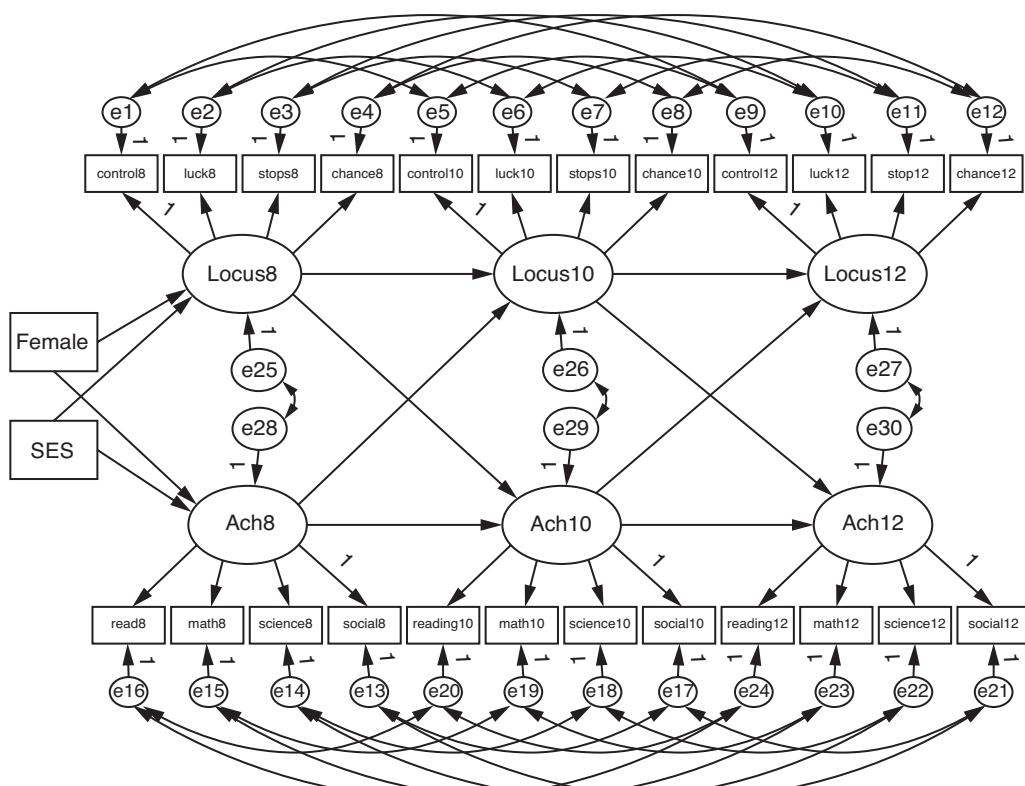


Figure 18.11 Latent variable panel model designed to compare the effect of achievement on subsequent locus of control with the effect of locus of control on achievement.

result of hard work or additional study (we just posited some potential mediators!). But then it also makes sense that high achieving students would, as a result of that success, develop a more internal locus of control. These two possibilities are embodied in the figure, and can be tested in this panel model. Spend a few minutes considering which of these you consider more likely. Or are they both likely?

I estimated the model using the NELS 8th through 12th grade data. The data (a correlation matrix and *SDs*) are in the file “sc locus ach matrix n12k.xls.” The Achievement test scores, SES, and Female (Sex, male=0, female=1) variables are familiar. The locus of control latent variable is indexed by four items in grade 8:

- BYS44B I don't have enough control over the direction my life is taking (called control8 in the figure),
- BYS44C In my life, good luck is more important than hard work for success (luck8),
- BYS44F Every time I try to get ahead, something or somebody stops me (stops8), and
- BYS44M Chance and luck are very important for what happens in my life (chance8).

The same items were administered in 10th and 12th grade. The website also contains a word file providing more information about each of the variables “Codebook for sc locus ach data.docx.”

Note that each of the measured locus and achievement variables has a correlated error with the same variable measured at each of the other two time points. It makes sense that the control item at time 8, for example, should share something in common with control10 and control12, beyond general locus of control. Disturbances of locus of control and achievement are also correlated at each time point. One of the reasons for doing a panel model is to determine the primary direction of influence, and so it makes sense that no causal influence is specified at each time point but that a correlation, an agnostic causal relation, is allowed.

The figure shows a fairly strict version of a panel model, however, in that the background variables of sex and SES only influence the 10th-grade variable through the eighth grade variables. Likewise, the 8th grade constructs of interest (locus and achievement) only influence 12th grade locus and achievement indirectly, through the 10th grade variables. These are model variations that could be tested as alternative models. You will have a chance to do so in the exercises!

The standardized results are shown in Figure 18.12. The model fit the data well using our normal criteria: RMSEA = .050, CFI = .964, SRMR = .044. The χ^2 was very large and statistically significant ($\chi^2 [262] = 790167$, $p < .001$), but then that is expected with a sample size of over 12,000. Given that large sample size, every path and correlation in the model is statistically significant, even the ones that we would consider too small to be meaningful. Given an acceptable fit, the results suggest that the primary direction of effect is from achievement to locus of control, rather than the reverse. Students who achieve at higher levels show higher, more internal, locus of control as a result. The effect of locus of control on subsequent achievement is negligible. Note that achievement in 8th grade also has substantial indirect effects on 12th-grade achievement, via both 10th-grade achievement (.93 * .14 = .13) and via 10th-grade locus of control (.10 * .54 = .05).

Panel models, as illustrated, can be useful for helping to understand the primary direction of effect. The model is also useful for illustrating the plausible unfolding of a developmental process, in this case how locus of control and achievement are related over middle to high school. We have again just scratched the surface of the topic of panel models. Little's text is an excellent source for more information concerning panel models and longitudinal SEM (Little, 2013). For an example of a panel model using social skills and achievement from Kindergarten to Eighth grade, see Caemmerer and Keith (2015; indeed, this research experience was what made me interested in illustrating panel models in this text). In Chapter 21 we will study another method for answering questions about how attributes develop over time, latent growth modeling.

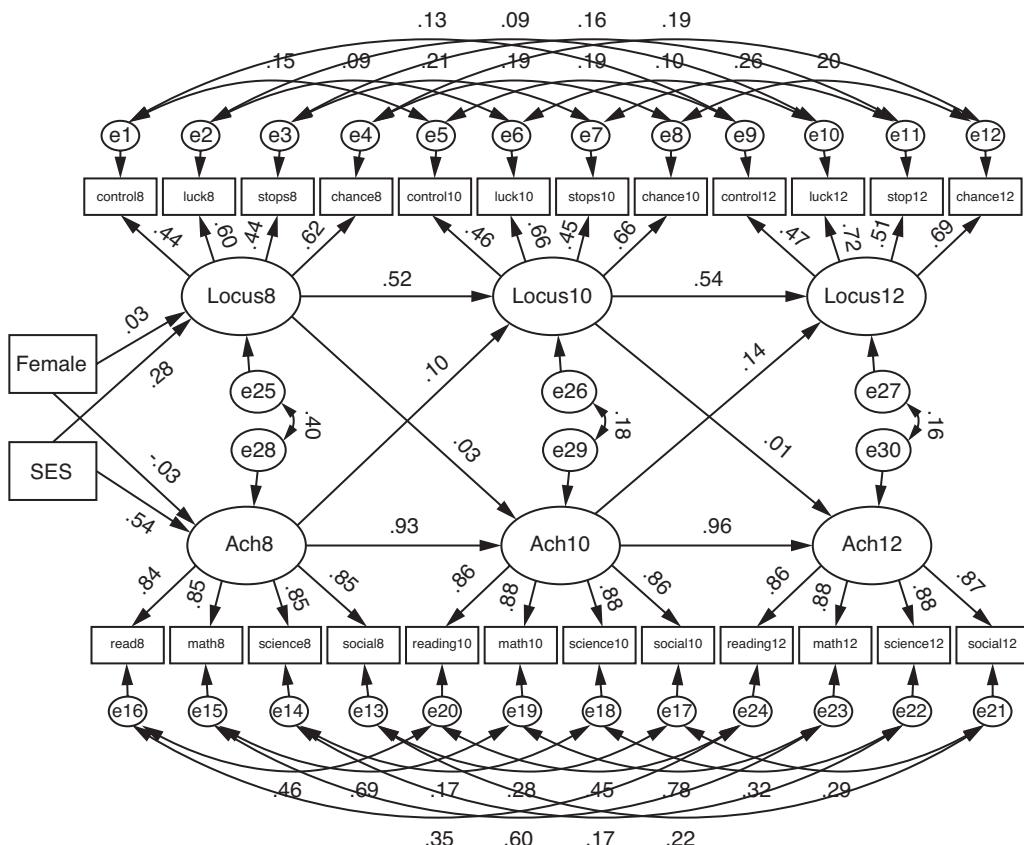


Figure 18.12 Standardized results for the panel model.

MULTIGROUP MODELS

Our previous homework models included the variable Ethnic Minority background. This variable was included for the simple reason that it is often included as a background variable in such models, although our analyses suggested that Ethnic Minority did not need to be considered to make the model valid. The results of the Ethnic Minority variable are interesting, however, in that our analyses suggested that Ethnic Minority has no effect on high school GPA and that minority students spend more time on homework than do White students, other things being equal. Perhaps more importantly for our purpose, its inclusion allowed the illustration of the use of a single-indicator latent variable.

A MultiGroup Homework Model Across Ethnic Groups

Now I want you to consider another possibility. Our explorations so far have suggested that Ethnic Minority background has no effect on GPA. It could be the case, however, that Homework has different effects on GPA depending on students' ethnic group membership. Previous research, for example, has suggested that homework may have larger effects on learning outcomes for minority, as opposed to majority, students (Keith, 1993; Keith & Benson, 1992). If this is the case, it means that a teacher or school that increased homework demands can expect this homework to pay off in increased learning for all students, but to result in an even larger increase in learning for minority students. If this sort of speculation sounds

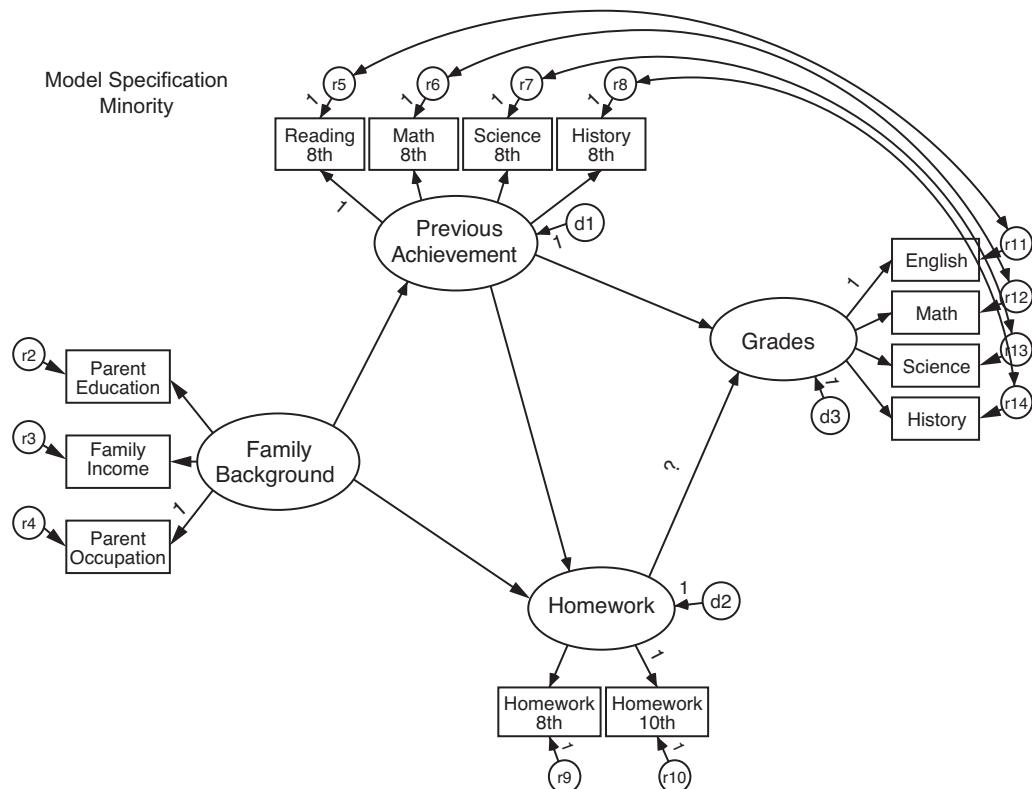


Figure 18.13 Initial multigroup analysis of the effects of Homework on Grades for majority and minority students. This model makes no constraints across groups.

methodologically familiar, it should. What we are talking about is the possibility of testing an *interaction* between Ethnic background and Homework in their effects on Grades. Another way of stating this is that we are interested in whether Ethnic background *moderates* the effect of Homework on Grades.

Conceptually, we can test this hypothesis by analyzing a homework model separately for Ethnic minority and White students and then comparing the effect of homework on grades for the two groups.² Such a model is illustrated in Figure 18.13; we can analyze the model for Ethnic minority students and find the unstandardized value for the path from Homework to Grades (denoted with a question mark in the figure). We can then analyze the model for White students and examine the same path. We might even put a 95% confidence interval around one of the coefficients and see if the other value was within this interval, as we did with regression coefficients in Chapter 2.

Before moving on, make sure you understand why Ethnic Minority does not appear in the model (because it is the variable on which the sample is divided into subsamples). Also, make sure you understand why I said to use the unstandardized paths (review the reasons in Chapter 2).

Constraining Parameters Across Groups

Although this method will work, there is a better method for testing the equivalence of this path, and other parts of the model, across groups. Within Amos and other SEM programs, it is possible to test *multigroup* (MG) or *multisample* models, which generally means the same model tested across two or more groups. With such MG models, it is possible to constrain

parameters to be equal across groups and compare the fit of these constrained models to models without constraints. An example will illustrate.

Figure 18.13 also illustrates the basic, or initial, model for the multigroup analysis. The identical model is specified for each group (Ethnic minority and White), and each group's model is estimated from its own data matrix. Both models are estimated within a single analysis. Thus, Figure 18.13 represents the input model for one group; the model for the other group is identical. In Amos, this is accomplished using the Manage Groups option under the Analyze menu. The manuals of other SEM programs will detail their method for conducting multigroup analyses. The file "initial multi group model.amw" shows this initial model, and the files "minority matrix.xls" "white matrix.xls" and contain Excel versions of the correlation matrices, means, and standard deviations necessary to estimate the models.¹

This initial model has no constraints across the two groups; the path from Homework to Grades is not constrained to be equal for Ethnic minority and White students, nor are there any other constraints. The reason is that this represents the baseline model to which we will compare models with such constraints. The fit statistics for the MG analysis represent the fit of "all models in all groups" (Jöreskog & Sörbom, 1993, p. 54). With no constraints across groups, the χ^2 and degrees of freedom for the multigroup analysis are the same as if we had analyzed the majority model and minority model separately and added together the values (the χ^2 is not always identical but should be quite close).

Figures 18.14 and 18.15 show the unstandardized output for the unconstrained multigroup model for Ethnic Minority and White students, respectively. First note the fit indexes. The χ^2 for the initial multigroup analysis was 219.576, with 112 degrees of freedom. In contrast, separate analyses (not shown) resulted in χ^2 equal to 92.097 (56 df) for Ethnic Minority

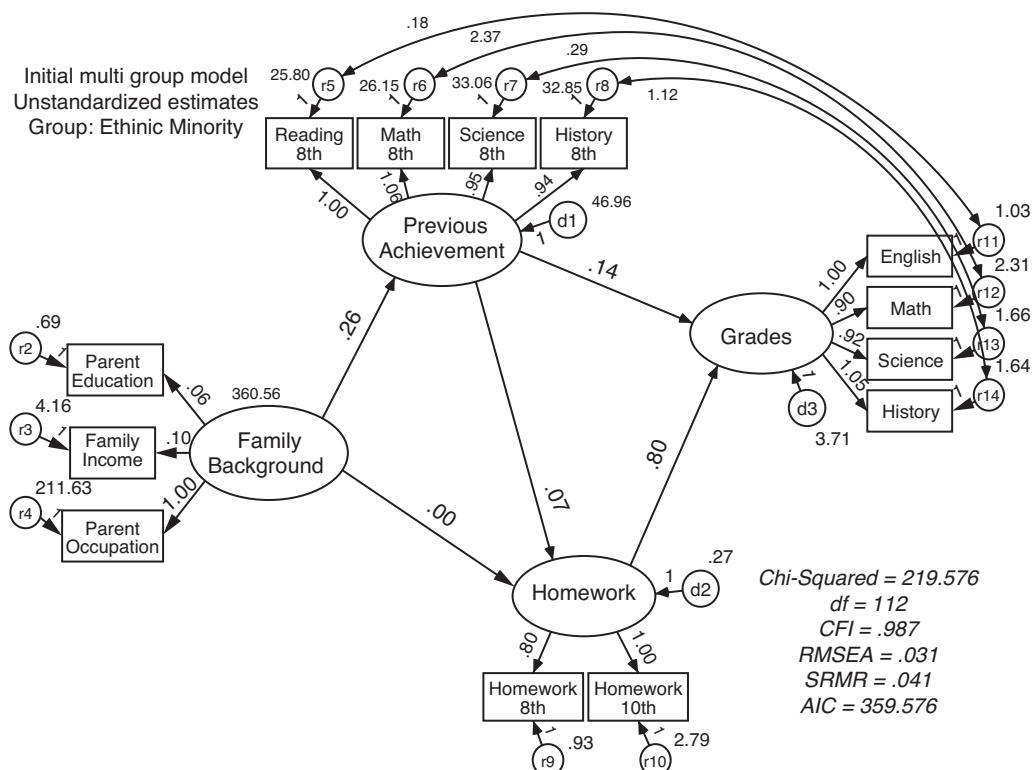


Figure 18.14 Unstandardized output for the unconstrained multigroup homework model. These results are for Ethnic Minority students.

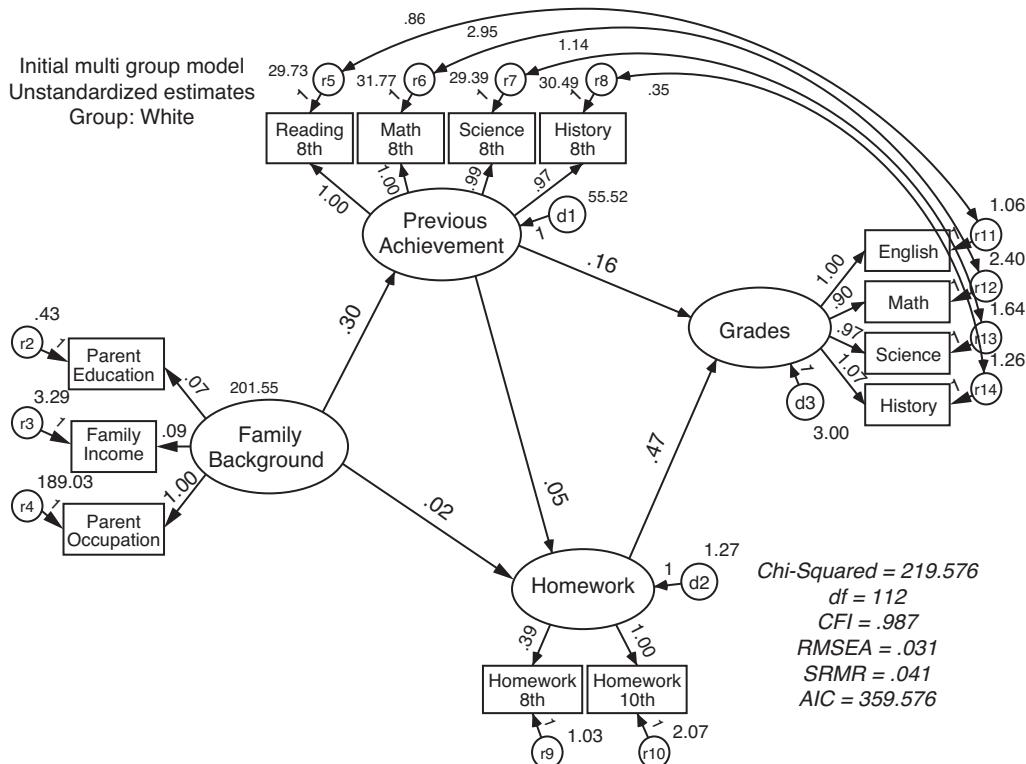


Figure 18.15 Unstandardized output for the unconstrained multigroup homework model. These results are for Ethnic Minority students.

students and 127.401 (56 *df*) for White students, which sums to 219.498 (112 *df*); the initial model, with no constraints across groups, has essentially the same fit (χ^2) as the two groups analyzed separately. The RMSEA for the multigroup analysis was .031, suggesting a good fit. Steiger (1998) has argued that the RMSEA should be adjusted in multigroup analyses, however, by multiplying it by the square root of the number of groups analyzed:

$$RMSEA_{adjusted} = RMSEA \times \sqrt{number\ groups}$$

$$= .031 \times \sqrt{2}$$

$$= .044$$

This adjusted value, .044, is closer to the average of the RMSEAs when the two groups are analyzed separately (.041 and .049) but also suggests a good fit of the models to the data across the two groups.² The corrected value has been used in the table of fits (Table 18.3), although the figures show the uncorrected values. The other stand-alone fit indexes (CFI, SRMR) also suggest a good fit of the model to the data across groups. The initial model appears to fit well and should serve as a good baseline for comparing subsequent models.

Our primary interest, of course, is whether the path from Homework to Grades is the same across groups. For minority students, the unstandardized effect was .80, versus .47 for majority students. Perhaps homework does have different effects for the two groups! Although if we use standard errors (.637 minority and .114 majority), we will be tempted to say that the two parameters are not different from one another, we will make such tests more directly in just a minute. Interestingly, the standardized paths for minority and majority students are nearly identical (.24 and .25), illustrating again the fact that the standardized and

Table 18.3 Comparison of MultiGroup Homework Models

Model	χ^2	df	$\Delta\chi^2$	df	p	CFI	RMSEA	SRMR	AIC	aBIC
1 All Free	219.576	112			.987 .044		.041	359.576	480.795	
2 Compare Loadings	235.591	121	16.015	9	.067 .986 .042		.044	357.591	521.436	
3 Compare Homework Effects	236.522	122	.931	1	.335 .986 .042		.044	356.522	517.681	
4 Compare All Effects	241.323	126	4.801	4	.308 .986 .042		.048	353.323	503.738	
5 All Parameters Invariant	288.262	147	46.939	21	.001 .983 .044		.054	358.262	452.271	

Note: Each model is compared to the previous model.

unstandardized coefficients may produce different answers to the question of equivalence across groups (again, the reasons for focusing on unstandardized as opposed to standardized coefficients for such comparisons are spelled out in Chapter 2).

Our baseline model fits well; let's now compare it to several models in which we add constraints across groups. In Amos, the way to add constraints across groups is to fix the relevant parameters to some alphabetic (not numeric) label. We can, for example, set the path from Homework to Grades to a value of *a* for both groups (or "path1" or some other label). This constraint will allow the parameter to be estimated but will constrain the (unstandardized) estimate to be identical across the two groups. In Mplus, like numbers or labels in parentheses are used to make the equality constraints [e.g., (1) in both groups for the Homework to grades path].

Measurement Constraints

Although our primary interest is in comparing the Homework to Grades path across groups, the first model to be compared actually involves a different set of constraints. The model shown in Figure 18.16 sets the factor loadings from all latent to measured variables to be the same across groups. The model shown is the setup for Ethnic Minority students. As in all previous models, note that one factor loading from each latent variable is set to 1.0. In addition, however, the other factor loadings are set to specific labels (fl2, for factor loading, through fl14). The model for White students, if displayed, would show the same constraints for the factor loadings, thus constraining these loadings to be equal for Ethnic Minority and White students.

Why start with constraints on factor loadings? Basically, this constraint specifies that the latent variables (Homework, Grades, etc.) are the same across the two groups. This specification means that we are measuring the same thing across groups, that our variables of interest mean the same thing for Ethnic Minority students as for White students. Consider for a minute what it would mean if Homework meant something different for one group compared to the other. If Homework has one meaning for one group and a different meaning for another, then it really doesn't make much sense to ask whether Homework has the same effect across the two groups, does it? Differences in the measurement model (factor structures) across groups suggest a difference in the constructs being measured. You will also hear this step of comparisons referred to as testing the *invariance* of the factor or measurement model across groups (we will cover this topic in depth in Chapter 20).

Table 18.3 shows the fit statistics for this Compare Loadings model in comparison to the initial model. There are 9 additional degrees of freedom for this model, representing the 9 factor loadings that were constrained to be equal across groups (one factor loading per latent variable was already set to 1 for both groups). The model is more parsimonious than the initial model, and thus χ^2 is larger. However, the $\Delta\chi^2$ was not statistically significant, meaning that the

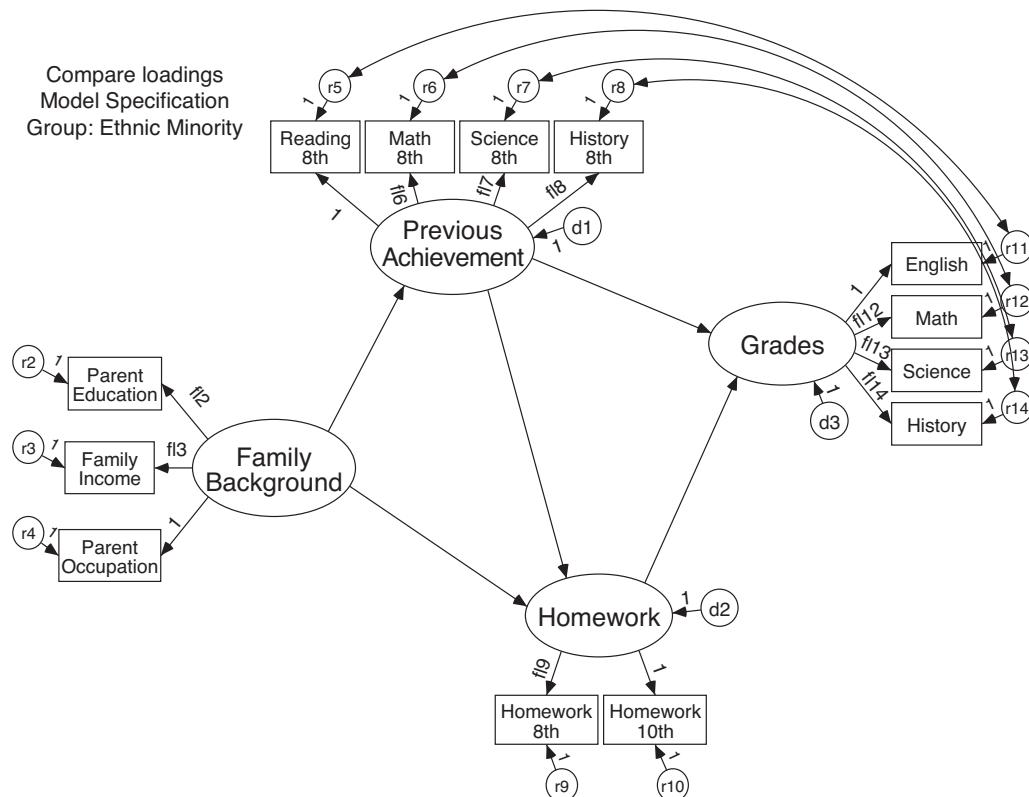


Figure 18.16 Multigroup homework model with factor loadings constrained to be equal across groups. Model specification for minority students; factor loadings are constrained to the same values (e.g., fl2, fl3) for majority students.

additional constraints are justified. The specification that the factor loadings of the latent variables be identical across groups cannot be rejected; the measured variables represent the same constructs for minority and majority youth; the latent variables have the same meaning across groups. Given that we are measuring the same constructs across groups, we can now determine whether these latent constructs have the same effects on each other across groups.

Does Homework Have the Same Effect Across Groups?

The next step in model comparison answers the question in which we are most interested: whether Homework has the same effect on Grades across groups. The model specification for Ethnic Minority students is shown in Figure 18.17. For this model, all the constraints from the last model (Compare Loadings) are retained, and one new constraint is added. For both groups, the path from Homework to Grades was set to a value of a , meaning that the path will be freely estimated but that the unstandardized path will be constrained to be equal across groups.

The $\Delta\chi^2$ and other fit indexes for this Compare Homework Effect model are also shown in Table 18.3. As you can see, the $\Delta\chi^2$ was not statistically significant. The additional constraint specifying that the effect of Homework on Grades be identical for minority and majority students did not lead to a statistically significant degradation in the fit of the model. It appears, then, that homework has about the same effect on high school students' grades whether they come from Ethnic Minority backgrounds or not. When students spend time on homework it will have the same effect on grades whatever their ethnic background.

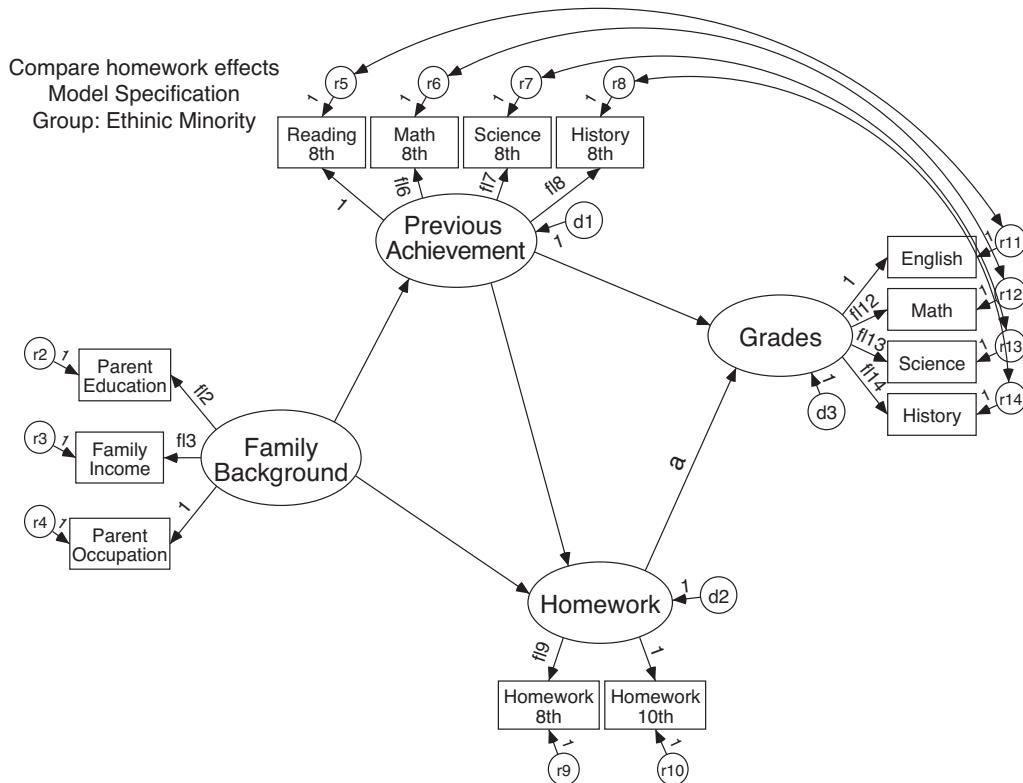


Figure 18.17 Multigroup homework model testing the equivalence of the effects of Homework on high school GPA across groups.

(at least for the gross division of Ethnic Minority/White). Another way of saying this is that there is *no interaction* between Ethnic background and Homework in their effects on high school Grades, or that Ethnic background does not appear to *moderate* the effect of Homework on Grades. Thus you now know how to test for interactions (moderation) between a categorical and a continuous variable in SEM. We will learn how to test for interactions among latent continuous variables in Chapter 22.

Other Effects

There may be several other comparisons of interest to pursue in these multigroup analyses. Although it appears that Homework has the same effect on Grades for both groups, we may wonder if the other variables in the model have the same effects on each other across groups. In essence, we are asking if *any* of the variables interact with Ethnic background in their effects on other variables in the model. To test this possibility, we can simply set all other paths (Family Background to Previous Achievement, and so on) in the model to be the same across groups. For this model, four additional constraints are required beyond those for the Compare Homework Effect model. The results of this Compare All Effects model are also shown in Table 18.3. Again, this more constrained model did not lead to a statistically significant $\Delta\chi^2$. It appears that all the effects of one latent variable on another in the model are consistent across groups; the variables in the model have the same effects on each other for minority as for majority youth.

None of the models so far has made constraints on the errors of measurement or the disturbances. Both of these types of parameters represent errors of some sort, either errors of measurement (r2 through r14) or the variance left unexplained by the other variables in the model (d1

through d3). These parameters do not really represent substantive portions of the model, and thus it is probably not reasonable to expect them to be invariant across groups (Marsh, 1993). Likewise, I can think of no substantive reason why the variance of the exogenous variable (Family Background) or the correlated errors (between the test scores and corresponding grades) should be expected to be equal across groups. For these reasons, these errors, variances, and covariances were not constrained to be equal across groups in any of the models, and we could reasonably stop our model testing without such constraints. For our present purposes, however, it will be instructive to see if these nonsubstantive parameters are indeed equivalent across groups.

The results for the All Parameters Invariant model are shown in the bottom row of Table 18.3. For this model, 13 measurement errors (r2 through r14), three disturbances, the variance of the Family Background latent variable, and the four correlated errors are constrained to be the same for the two groups. These 21 additional constraints resulted in a statistically significant increase in $\Delta\chi^2$. Taken together, these equality constraints across the two groups resulted in a statistically significant degradation in model fit. The errors and other nonsubstantive aspects of the models, as expected, are not identical across groups. If desired, we can fix or free these parameters in smaller blocks to see exactly where the differences are (less formally, we can compare the unstandardized parameters for the models shown in Figures 18.14 and 18.15 to look for differences). Note, however, that if we were to use the aBIC as our primary criteria for choosing among competing models, we would have concluded that the All Parameters model was the best fitting model. As we will see in Chapter 20, simulation research suggests the possible use of change in CFI for such tests of invariance.

Figures 18.18 and 18.19 show the standardized estimates for the Compare All Effects model for Ethnic Minority and White students, respectively. (Note this is the next to last model in

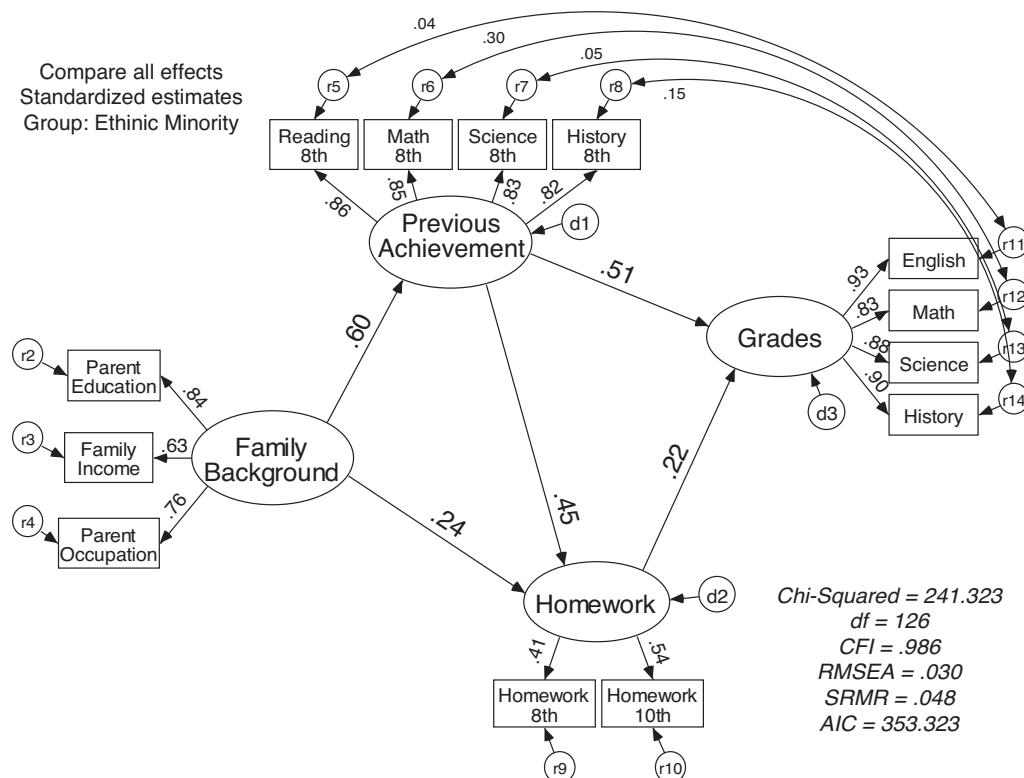


Figure 18.18 Standardized estimates of the effects of Homework and other influences on GPA for Ethnic Minority students. These results pertain to the model in which all influences were constrained to be equal across groups (Model 5: Compare All Effects).

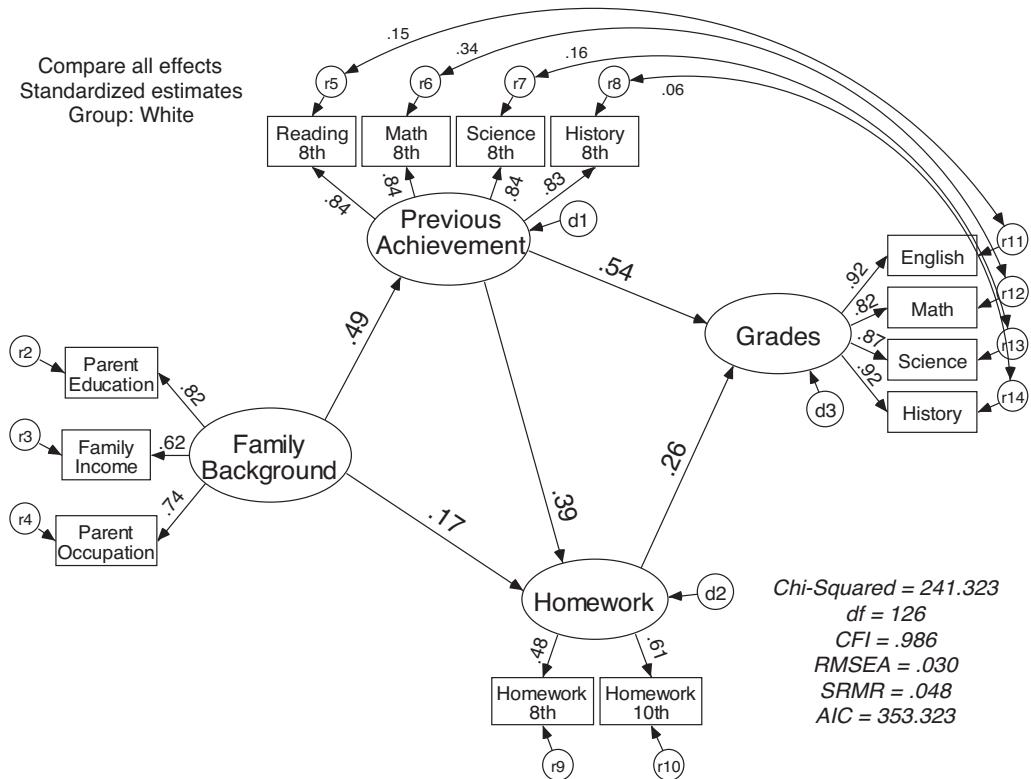


Figure 18.19 Standardized estimates for the compare all effects model for White students. The standardized estimates differ across groups because constraints were made for *unstandardized* parameters.

the table, model 5.) Of course, for this model the factor loadings and paths are set to be equal across groups, so in the *unstandardized* estimates they will be the same for majority and minority youth. Note the minor differences, however, for the standardized estimates across groups. Again, the unstandardized coefficients should be used to compare across groups; standardized estimates should only be used for interpretations within each group. Nevertheless, our interpretation of effects will be similar within each group and also consistent with the estimates for the overall model given earlier in the chapter.

Summary: MultiGroup Models

This series of analyses has illustrated a method for conducting tests of interactions in SEM via the comparison of nested, MG models. The method can be used to test for an interaction between a single categorical variable and a single continuous variable or, more broadly, between one categorical variable and *all* other variables in the model. This broader orientation (e.g., the Compare All Effects model) essentially asks if entire models are comparable across groups and may be of interest when you have questions such as “Are the variables that influence the learning of White students also important for Ethnic Minority students?” Such questions are common. For example, in the 1980s, one outcome of the controversial report *A Nation at Risk* (National Commission on Excellence in Education, 1983) was a proposal for an ideal, academic high school curriculum (Bennett, 1987). Columnist William Rasberry (1987) agreed that such a curriculum should work well for White and middle-class youth but wondered if it would work equally well for youth from Ethnic Minority backgrounds. One

way to test this question would be to test a multigroup school learning model across various ethnic groups (e.g., Keith & Benson, 1992).

There is nothing sacred about the order in which I tested successive models. It is just as defensible to begin with the most constrained model and gradually free parameters. Different groupings of constraints will also work well. For example, I could have constrained parameters one at a time, rather than as a block, in going from the Compare Homework Effect model to the Compare All Effects model. The main considerations in this process of model comparison should be that you do them in a logical, systematic fashion, that you understand exactly what is and what is not being tested at each step, and that your model comparisons answer research questions of interest. In the upcoming chapter on more advanced aspects of CFA (Chapter 20), we will spend more time discussing the meaning of each step in invariance testing.

There is also one additional model that is sometimes tested in such analyses. We could also compare the overall covariance matrix for Ethnic Minority youth with that of White youth. Consider that all the models that we have estimated are derived from the covariance matrices. Thus, if we specify that the two covariance matrices be identical across groups, this means that we are specifying that *all* aspects of the model be identical across groups, but without specifying a model. In essence, this comparison says, “I don’t know (or don’t care) what the model is, but whatever it is, it’s the same across groups.” This model is thus nested with, but less constrained than the All Parameters Invariant model, and the difference between the two represents the cost of specifying a *particular* model.

This example has illustrated MG models as a method of testing for an interaction between a categorical variable and other variables in a SEM. There are also methods of testing for interactions between continuous variables in SEMs, to be introduced in Chapter 22 of this book. Schumacker and Marcoulides (1998) is a good resource for more information about this method. Recent versions of the Mplus program have made such tests considerably easier than in the past. In tests of invariance of structures across groups, it is also possible to test for invariance in means and intercepts; this topic will be covered in subsequent chapters.

DANGERS, REVISITED

Recall that in Chapter 13 we discussed the dangers of path analysis in particular and nonexperimental research in general. Given that I have argued for the advantages of overidentified models (in Chapter 14) and latent variable models in the last few chapters, you may wonder if the fit statistics that result from using SEM programs to analyze overidentified models or the advantages of latent variables somehow obviate these dangers. Let’s find out.

Omitted Common Causes

Throughout this book, I have argued that the biggest danger in nonexperimental research is the possibility of neglecting to include in the analysis an important common cause of the presumed cause and the presumed effect. As shown previously, a neglected common cause will result in inaccurate estimates of the effects of one variable on another. Do fit indexes and latent variables control this danger; do SEM programs alert you when you have neglected a common cause? Unfortunately, no, generally they do not.

Figure 18.20 shows one of the homework models analyzed in the beginning section of this chapter. It is obvious that Previous Achievement is an important common cause of Grades and Homework. Previous Achievement had a large effect on both Homework ($\beta = .41$) and Grades (.52). What will happen if we delete it from the model? Will the fit statistics or some other aspect of feedback alert us to the deletion?

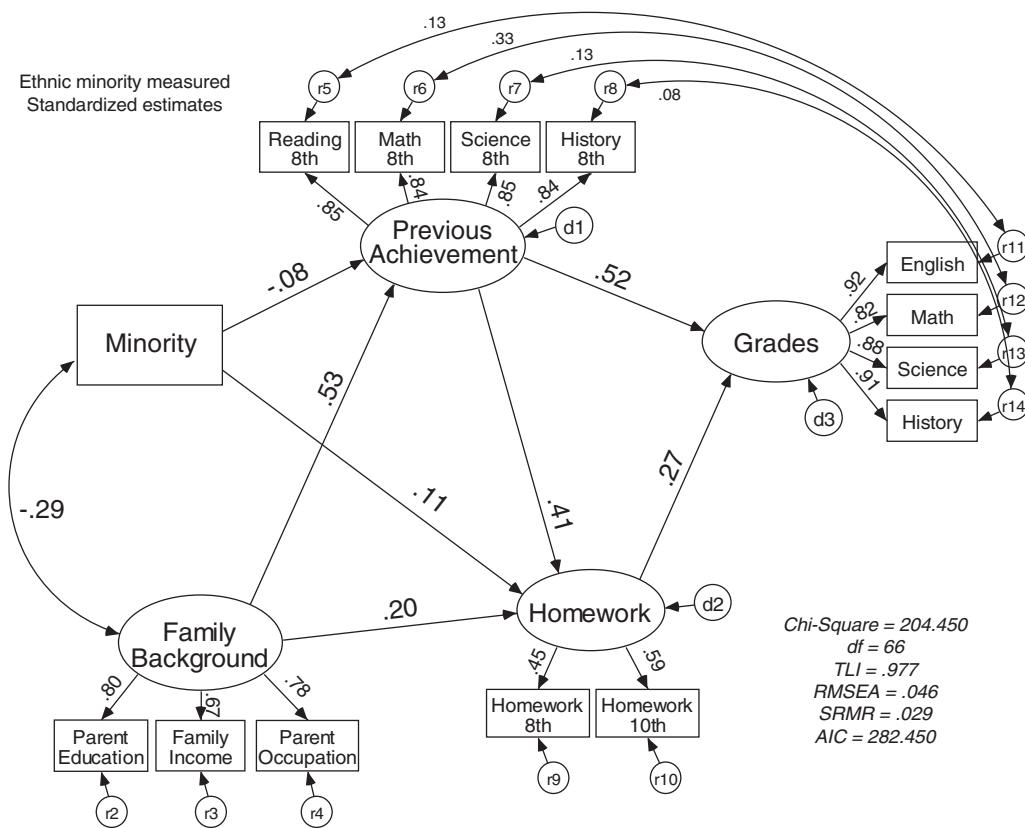


Figure 18.20 The latent variable homework model from the beginning of the chapter.

Figure 18.21 shows the results of an analysis without Previous Achievement in the model. As expected and consistent with previous analyses, the apparent effect of Homework on Grades changed dramatically from the previous analysis. Also consistent with previous discussions, the omission of this important common cause led to an inflated estimate of the effect of Homework on Grades. In this model, the standardized effect of homework was .66, much inflated from the .27 value in earlier figures.

Notice the fit indexes accompanying the model. Obviously, it is not the case that the omission of an important common cause resulted in a worse fit; in fact, the model fits better without Previous Achievement in the model! (Of course, with different variables in the model, you can't compare the chi-square values formally, but at an informal level the model without the common cause shows a better fit. The AIC for the model without the common cause is also lower, suggesting a better fit.) Likewise, there is nothing in the more detailed fit information that suggests to you, the researcher, that you have done something wrong, such as neglecting an important variable. As this example illustrates, the fit statistics of latent variable SEM do not protect against the danger of omitted common causes; they do not alert us to any errors. If you think about it, this makes sense. The fit statistics can only tell us about the fit of the variables *in* the model; they don't inform us about things that are *not* included in the model.

In contrast, if we want to find out if Previous Achievement is indeed a common cause of Homework and Grades, we can put the variable in the model and set the path from Previous Achievement to either Homework or Grades (or both) to zero. In this case, the fit statistics will show a statistically significant degradation. But the common cause must generally be in the model to test it.³

Missing common cause
Standardized estimates

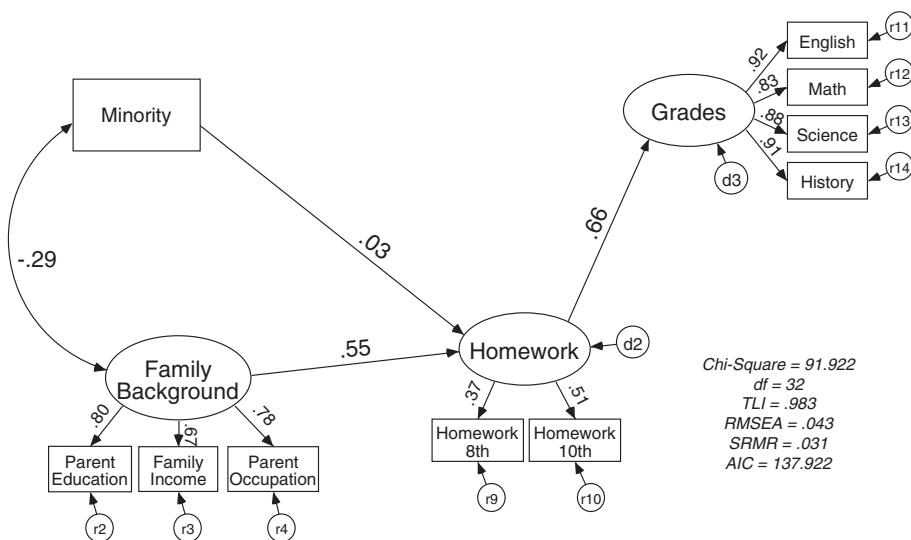


Figure 18.21 Homework model with the Previous Achievement variable omitted. The fit statistics do not alert us to the fact that we have not included an important common cause of Homework and GPA.

Path in the Wrong Direction

What about the other major danger in nonexperimental analysis: assuming a variable is an effect when it is really a cause (or vice versa)? Figure 18.22 shows the results of a model that is misspecified, with a path drawn in the wrong direction. In fact, unlike many models, we can be certain that this model is incorrect, because the Grades variable occurs in time mostly after the Homework variable. More importantly, using the rules for generating equivalent models from Chapter 14, this model is *not* equivalent with the original model from Figure 18.2 (and Figure 18.20). The Homework and Grades variables do not have the same variables pointing to them, and the path cannot be reversed and still have an equivalent model. The two models are not nested, however, because they have the same degrees of freedom. Thus, the two models should be comparable via fit statistics (e.g., the AIC) that do not require nested models. It is gratifying to see that the model with the path drawn in the correct direction (Figures 18.2 and 18.20) indeed had the lower AIC and thus the better fit. In this example, we would have chosen the model with the path drawn in the correct direction even without prior knowledge of the correct order of the variables. Yes, under the right conditions, with overidentified models that are also nonequivalent we *may* be able to guard against the danger of drawing paths in the wrong direction.

You might also wonder whether latent variable SEM has led to improvements in our tentative causal statements over those we made using multiple regression. Recall the assumptions needed to interpret regression coefficients as effects, from Chapter 13:

1. There is no reverse causation;
2. The exogenous variables are perfectly measured;
3. The causal process has had a chance to work (equilibrium);
4. No neglected common causes.

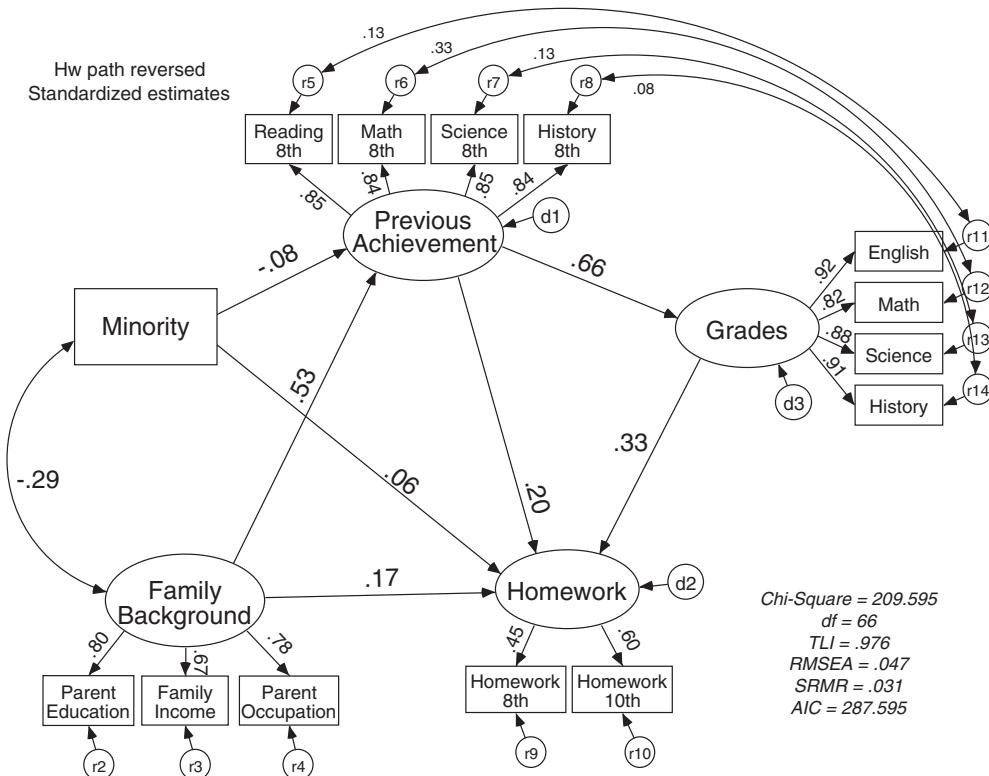


Figure 18.22 Path from Homework to Grades is incorrectly reversed in this model. The model has a worse fit than the initial model.

As we have seen, the addition of latent variables specifically addresses the issue of imperfect measurement (assumption/danger 2). SEM models can test for reverse causation and, if carefully developed (as we have just demonstrated), test for effects in one direction versus the other (assumption/danger 1). As already noted, the issue of common causes is probably more commonly and more properly addressed through careful consideration of previous research and theory. The issue of equilibrium is also likely best addressed non-statistically, at least for non-recursive SEM models (Kline, 2016, chap. 6). The important lesson is that those who wish to use SEM to answer research questions need to know their area of research, not just the ins and outs of the statistical method.

SUMMARY

This chapter reviewed and built on the previous chapter and covered several more complex topics in latent variable SEM. We estimated a latent variable version of our earlier homework model. This model included two interesting features: a latent variable estimated from a single measured variable and correlations among the unique and error variances.

The latent variable homework model included one latent variable (Ethnic Minority) that was indexed by a single measured variable (Minority). The primary reason for doing this, rather than simply using only a measured variable, is to build into the model estimates of the error inherent in the measured variable and to take this error into account in the analysis. In the example, we estimated the Minority variable to have a reliability of .95, and thus 5% of the variance in Ethnic Minority is attributable to error. To use this information in the model, we constrained the error variance of the measured Minority variable to be 5% of its total variance.

The homework model also included the specification of correlations between the unique and error variances of the 8th-grade Achievement test scores and high school Grades in related areas. The reason for specifying these correlated errors was to recognize that Grades and tests in a particular area (e.g., Mathematics) may share more than simply the effect of general achievement on general Grades. Our initial and subsequent comparative analyses showed that these correlated errors are indeed important and that removing them from the model resulted in a statistically significantly worse fit to the model.

A latent variable panel model of locus of control and achievement made even more use of correlated errors and disturbances. Such longitudinal models are often used to understand a developmental process and to test questions of causal ordering.

We explored multigroup models as a method of testing interactions between categorical and other variables in SEMs. To illustrate the method, we analyzed the homework model separately for Ethnic Minority and White youth in an effort to determine whether the effects of Homework on Grades are the same for both groups. The example gradually constrained parameters to be the same across groups and used $\Delta\chi^2$ to test the viability of these constraints. In the example, we showed that the constructs (latent variables) are equivalent across groups and that the variables in the model have the same effects on each other across groups. Homework, it appears, has the same effect on the learning of Ethnic Minority and White youth. Ethnic background (at least using the coarse categorization used here) does not moderate the effect of Homework on Grades.

The final section of the chapter revisited some of the dangers we discussed previously in connection with structural equation modeling, path analysis, multiple regression, and nonexperimental research. Do the fit statistics and other advantages of SEM obviate these dangers? We showed that there is nothing in the fit statistics or other aspects of a latent variable SEM that will alert us when we neglect an important common cause in our models. In contrast, the measures of fit did alert us when we estimated a model with a path drawn in the wrong direction. This was only the case because we were working with an overidentified model and comparing the fit of two nonequivalent models. Of course we also explicitly tested this question of direction; if we had not done so we would not have known that we made a mistake in the model shown in Figure 18.22. Latent variable SEM methodology does not protect against the danger of an omitted common cause, but if you plan carefully to construct nonequivalent models, you may be able to guard against the danger of a path drawn in the wrong direction.

EXERCISES

1. Analyze the series of full homework models starting with the model shown in Figure 18.1. Make sure your results match those presented here (minor differences may occur in different programs). If you are using a student version of a program that places a limit on the number of variables you can analyze, try eliminating the Ethnic variables, Family Income, and the Science and History Test and Grades variables. Estimate this smaller model; compare your results to those presented in this chapter. Are the results similar?
 - a. Study the parameter estimates and standard errors, the fit statistics, modification indexes, and standardized residuals. Are there changes that you might make to the model? Are they theoretically justifiable?
 - b. Interpret the model. Be sure to interpret the indirect and total effects in addition to the direct effects.
 - c. Compare the model with the two competing models discussed in this chapter (the Direct Background Effects model and the No Homework Effect model).

Eisenberg et al.
Model Specification

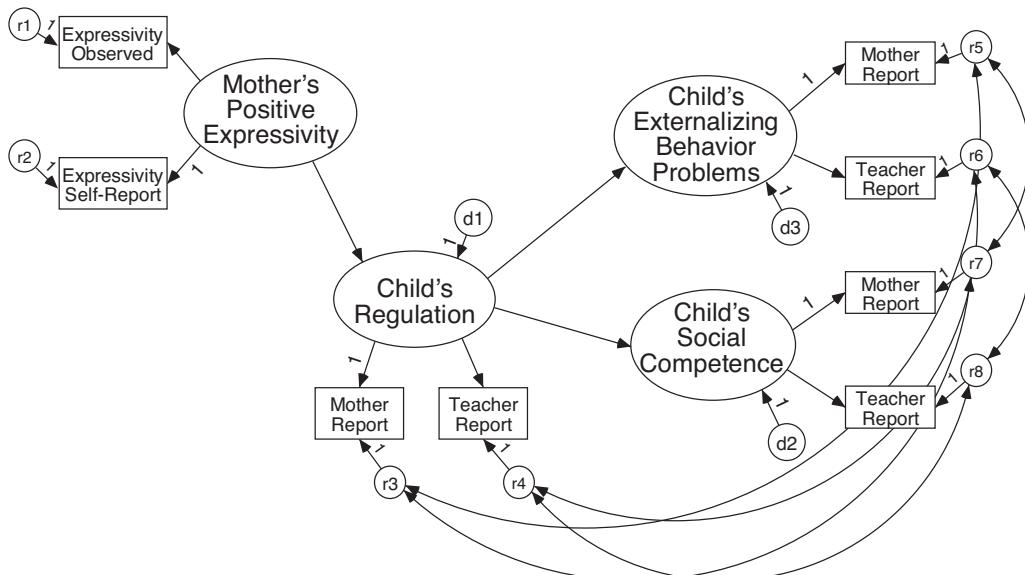


Figure 18.23 Model testing the effects of Mother's emotional expression on child outcomes. The model is drawn from Eisenberg et al., 2001.

2. Nancy Eisenberg and colleagues (2001) conducted research to determine the effects of mothers' emotions on their young children's behavior problems and social competence. One interest in the research was whether these effects are mediated by children's own emotional regulation. Figure 18.23 shows a model patterned after those in the article. It includes fewer variables but still includes many interesting aspects of the original research. Mother's Positive Expressivity represents mothers' expression of positive emotions with their children, both as rated by the mothers and as observed during their work on a task. Child's Regulation, Externalizing, and Social Competence are latent variables representing these child characteristics, and each was rated both by the mothers and by teachers. The model includes correlated errors between the child ratings by the mother and the child ratings by the teacher. The model is contained in the Amos file "Eisenberg et al 1.amw." Simulated data designed to mimic the relevant portions of the correlation matrix presented in the article are contained within the Excel file "Eisenberg et al 2001.xls" and the SPSS file "Eisenberg et al 2001.sav" on the accompanying Web site. The variable names in the data and the corresponding labels from the model are shown in Table 18.4.
 - a. Estimate the model as shown. Focus on the fit indexes and, if you judge them to be adequate, interpret the model.
 - b. Estimate a model without the correlated errors. What happens to the fit of the model? Were these parameters justified?
 - c. Compare the fit of this model with one in which Mother's Expressivity is also allowed to have direct effects on the two child outcomes. Based on the change in fit, would you say that children's Regulation completely or partially mediates the effect of Mother's Expressivity on Behavior Problems and Social Competence?
 - d. Calculate and interpret direct, indirect, and total effects for your accepted model.
 - e. Test any additional alternative models that are of interest.

Table 18.4 Variable Names and Variable Labels for the Eisenberg and Colleagues Model

Variable name (Eisenberg et al 2001.xls)	Variable label (from figure 16.22)
Exp_mo	Expressivity Observed
Exp_msr	Expressivity Self-Report
Reg_mr	Mother Report (Child's Regulation latent variable)
Reg_tr	Teacher Report (Child's Regulation)
Exter_mr	Mother Report (Externalizing Problems)
Ext_tr	Teacher Report (Externalizing Problems)
Soc_mr	Mother Report (Social Competence)
Soc_tr	Teacher Report (Social Competence)

3. Analyze the latent variable panel model of locus of control and achievement. The data are in the file “sc locus ach matrix n12k.xls,” and more information about the variables is shown in the related codebook file. Estimate the model as shown in the figure; make sure your results match those presented in the chapter.
- Given the nature of the measured variables, are there any correlated errors you might add a priori to improve model fit? Consider, in particular, if some of the Locus items measure something in common other than general locus of control.
 - Examine the modification indexes and standardized residuals to determine if there are other model modifications you might make to improve fit. Do you have a theoretical justification for doing so? How can the modification indexes be so large in a model that generally fits well?
 - Using the concepts of setting parameters to be the same from the multigroup section of the chapter, constrain the factor loadings of the Locus factor to be the same over time. That is, constrain the loadings for luck8, luck10, and luck12 to be the same, constrain the loadings for stops8, stops10, and stops12 to be the same, and so on. Do the same for the Ach factor loadings. Using a criterion of ΔCFI larger than an absolute value of .01, decide whether these model modifications are reasonable. This is a type of factor invariance; it and the ΔCFI criterion will be discussed in more detail in Chapter 20.
 - Specify two alternative models to determine, based on fit, whether Ach affects Locus, and whether Locus affects Ach. Why is $\Delta\chi^2$ not a good choice for these comparisons? What might you examine instead?
 - Specify an alternative model to answer the question of whether SES has any direct effects on 10th-grade locus of control and achievement. Next test for direct effects on 12th-grade Locus and Achievement. Are the effects statistically significant? Are they meaningful? How did you decide?
 - Specify an alternative model in which Locus8 is allowed to affect Locus12 and Ach8 is allowed to affect Ach12. Are the effects statistically significant? Are they meaningful? How did you decide?
- 4 Does Achievement have different effects on Locus of Control for girls versus boys? Conduct a multigroup analysis using the model shown in Figure 18.24 for girls and boys. The data (a subset of the NELS data) are in the raw data file “ach locus sex data 2.sav”. The variables should be fairly self-explanatory. The SES variable is a composite of parent education, occupational status, and family income. The Ach8 latent variable is indexed by 8th grade test scores in Reading, Math, Science, and Social Studies. The Locus10 variable is a latent variable estimated by items asking students their agreement

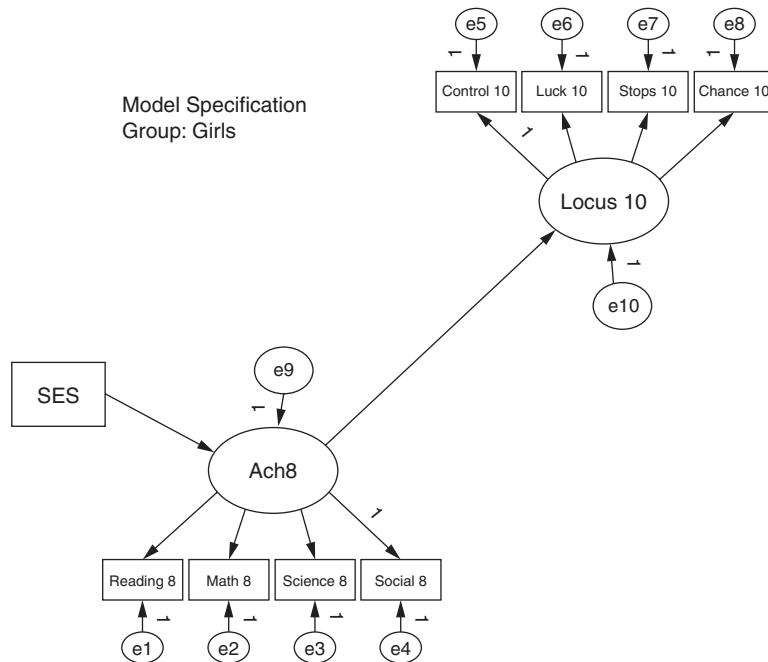


Figure 18.24

Table 18.5 Locus of Control items in the file “ach locus sex data 2”

Variable	Wording
F1S62b	I don't have enough control over the direction my life is taking
F1S62c	In my life, good luck is more important than hard work for success
F1S62f	Every time I try to get ahead, something or somebody stops me
F1S62m	Chance and luck are very important for what happens in my life

with the statements shown in Table 18.4. The statements espouse an external locus of control, but “strongly agree” has a value of 1 and “strongly disagree” has a value of 4, so high scores represent an internal (more healthy) locus of control.

Estimate three multi-group models:

1. A model that makes no constraints across the groups.
2. A model with all factor loadings constrained to be equal across the two groups.
3. A model in which the path from Ach8 to Locus8 is constrained to be equal across groups.

Prepare a table of fits comparing these three models. Include χ^2 and df; $\Delta\chi^2$, Δdf , and probability; RMSEA; CFI; SRMR; and AIC, or whatever indexes are recommended by your instructor. Evaluate the fit of the initial model. Explain what you would conclude by

comparing model 1 with 2 and by comparing model 2 and model 3. Which model fit the best? What does that mean? Interpret the results from the model that you believe is most supported. What do the results mean? Explain as you might to your Uncle Gus (a real world interpretation without jargon).

Extra credit: The model shown is a subset of the panel model shown in Figure 18.11. Yet the estimate of the effect of Ach 8 on Locus10 is quite different from the effect shown in Figure 18.12. Why might that be?

Notes

- 1 Most SEM programs, including Amos, also allow the analysis of a single, raw data file, with some selection or grouping variable used to separate the two groups. We will not delve that deeply into program specifics here, but it is good to know that this option is available.
- 2 Many programs, including Mplus, Steiger's SPATH, and LISREL, have the correction already built in.
- 3 I think it is *possible*, through the use of carefully planned overidentified models, to test whether there are unmeasured common causes not included in a model. Such models might include both a path between the two variables of interest and a correlated disturbance between those variables. This is not common, however, so in normal usage I don't see latent variable SEMs protecting against the danger of unmeasured common causes.

19

Latent Means in SEM

Preparatory Work	445
<i>Displaying Means and Intercepts in SEM</i>	445
<i>Estimation of Means and Intercepts in Single Group SEM Models</i>	448
Overview: Two Methods to Test for Differences in Latent Means	454
Example: Hypnosis for Hot Flashes	456
<i>Single Group/Dummy Variable Approach</i>	456
<i>MG-MACS Approach</i>	460
<i>Comparing the Two Methods</i>	466
<i>Other Technical Issues</i>	470
Summary	471
Exercises	473
<i>Notes</i>	474

Most of the models we have considered so far have only focused on the analysis of variances and covariances. This should not be surprising, given that one name for SEM is the analysis of *covariance* structures. Yet it is also possible to analyze means in SEM, and there are definite advantages in doing so.

In fact, we have already analyzed several models that included analysis of means. In Chapter 17, for example, one of the exercises analyzed data from Head Start, with the Head Start variable coded 0 for those in the control group and 1 for those in the experimental group. As we will see in this chapter, the resulting unstandardized path from Head Start to Cognitive Ability represented the mean difference between the experimental and control groups on the cognitive ability latent variable (controlling for SES and Education). Including dummy variables in an SEM is thus one way of analyzing latent means. We will explore this method in more depth, and then delve into a more complete method as well. Latent means are also of interest in CFA.

Why focus on latent means? As suggested in the Head Start example, including means in our analyses can help us understand whether an experimental treatment results in differences in some outcome. This may not seem like a big deal; analysis of variance (ANOVA), after all, can answer that question. Why go to all the trouble? Why use an electron microscope when a magnifying glass will do? The advantage is that SEM can focus on latent, rather than measured, variables. As a result, SEM with latent means can help us determine whether the treatment resulted in change in the construct of true interest (e.g., Cognitive Ability) rather

than an error-laden measured variable (scores on a single measure of cognitive ability). Of course, SEM can also be used to test for differences in latent means for other categorical variables, such as differences across the sexes, ethnic groups, or family structures. SEM can also test assumptions taken for granted in ANOVA.

In Chapter 18 we explored how to test for interactions (moderation) in SEM when we examined possible differential effects of homework on achievement for different ethnic groups. In prior analyses, with the variable Ethnic Minority (vs. White) in the model, we examined the main effect of ethnic background on Achievement (latent mean differences). When we conduct multi-group analyses using latent means it will be possible to examine both questions—main effect and interaction—in a single analysis.

As noted, latent means are also of interest in confirmatory factor analyses. In Chapter 18 we also discussed the importance of testing for invariance of constructs prior to testing for differences in effects across groups: in the example, we tested whether the constructs in the homework model were equivalent prior to testing for differences in the effects of homework across groups. This need for invariance extends to analysis of latent means. There are also substantive questions that can be answered via the inclusion of latent means in CFA. We may be interested, for example, whether there are *true* differences across groups on some latent variable, as opposed to differences on an error-laden measured variable.

We will start with some preparatory work: reviewing slopes, intercepts, and means from regression, and then seeing how to integrate means and intercepts into SEM. An example will illustrate the explicit estimation of measured means and intercepts in a SEM model. This work will set the stage for the estimation of latent means and intercepts in SEM, and two methods will be shown for accomplishing this purpose. One, via the inclusion of one or more dummy variables in a single-group SEM, has already been illustrated, but the latent means aspect has not been emphasized. The second, via multigroup analysis, will allow the testing of both main effects and interactions in a single latent variable analysis.

PREPARATORY WORK

Displaying Means and Intercepts in SEM

Although I have been talking about latent means in SEM, we are in fact interested in both means and intercepts (I will sometimes refer to these in combination as “mean structures”). We haven’t talked about intercepts since discussing multiple regression, so let’s review, and then we’ll see what these look like in SEM.

Figure 19.1 shows the results from the regression of a 10th-grade Math test on 10th-grade Homework (time spent on homework out of school). The data are in the file labeled “math & hwork means.sav,” which is a subset of the National Education Longitudinal Study (NELS) data set. The figure shows the descriptive statistics for the two variables, along with the table of coefficients for the regression (intercept and regression coefficient). The figure also shows the scatterplot with regression line.

The table and graph should serve as a reminder as to what the regression coefficients represent: the *intercept* (47.74) is the predicted value on the *dependent* variable for students who have a value of zero on the homework variable. Our best bet for the math achievement score for a student who does no homework is 47.74. The unstandardized regression coefficient (1.57), in turn, represents the slope of the regression line. Other things being equal, each additional unit of time spent on homework should result in a 1.57-point increase in math achievement.

Figure 19.2 shows the unstandardized results of the same regression in SEM (Amos) format, with the added specification that mean and intercepts are analyzed.¹ The value for the

Descriptive Statistics

	Mean	Std. Deviation	N
Math MATH STANDARDIZED SCORE	51.5465	9.56419	500
Homework TIME SPENT ON HOMEWORK OUT OF SCHOOL	2.42	1.592	500

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 Constant (Intercept)	47.743	.754		63.359	.000	46.262	49.223
Homework TIME SPENT ON HOMEWORK OUT OF SCHOOL	1.569	.260	.261	6.037	.000	1.058	2.080

a. Dependent Variable: Math MATH STANDARDIZED SCORE

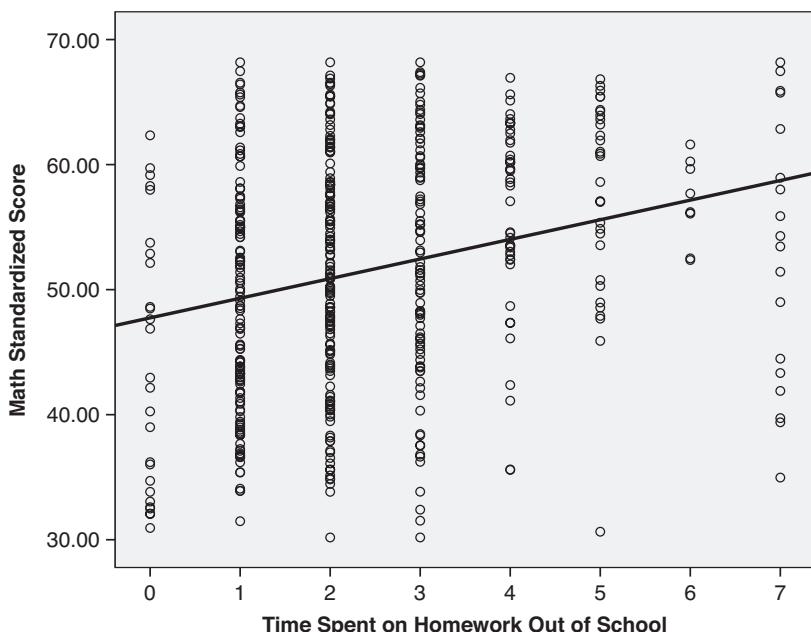


Figure 19.1 Regression results for Math Achievement on Homework.

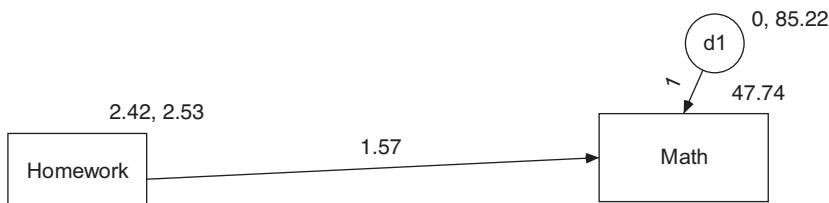


Figure 19.2 Regression results in SEM (Amos) format.

path (1.57) is, as in previous models, the unstandardized coefficient, the slope of the regression of Math achievement on Homework time. The values above the Homework rectangle show the mean (2.42) and variance (2.53) of the Homework variable, and match the regression results (Homework SD squared, $1.592^2 =$ Homework variance, 2.53). The value above the Math rectangle (47.74) shows the intercept for the regression. The values beside the disturbance (d1) show the mean and variance of the disturbance. In most cases we will assume that latent variables (including error terms) have means of zero. Just to be clear, means are estimated for exogenous variables, whereas intercepts are estimated for endogenous variables. Figure 19.3 labels these components as they will be used in subsequent SEM figures.

As you read articles and other books describing SEM findings when means are analyzed you are likely to encounter another graphic method of displaying means and intercepts: McArdle and McDonald's reticular action modeling (RAM) format (McArdle & McDonald, 1984). Figure 19.4 displays the current regression using RAM symbols (top) and lists the

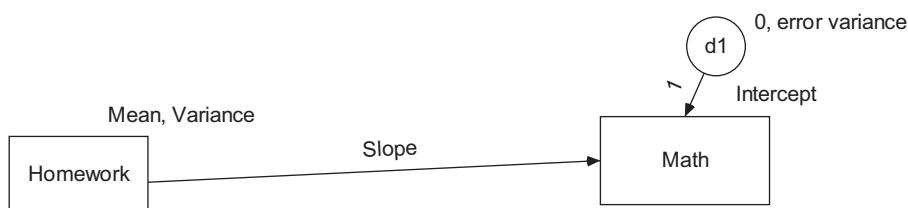


Figure 19.3 Components of SEM results (unstandardized solution) with the analysis of means and intercepts. The mean of the disturbance is set to zero; this is a common assumption for latent variables.

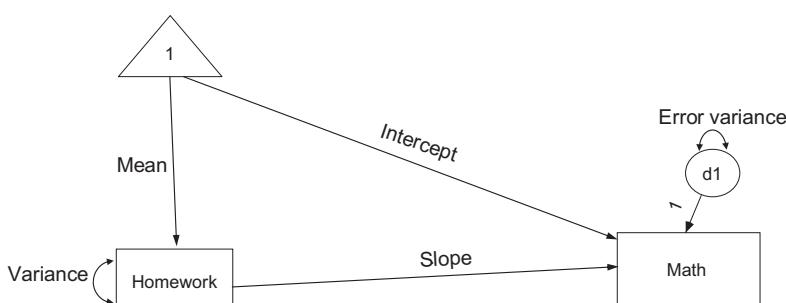
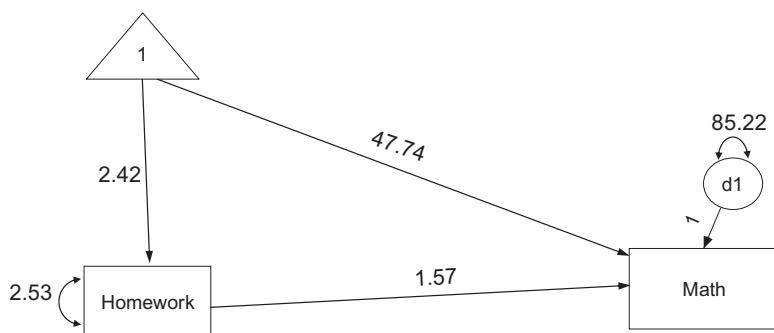


Figure 19.4 Regression results and format for display using the RAM format when means and intercepts are estimated.

components of the format (bottom). The biggest departure from the previous display is the inclusion of a triangle, with paths pointing to Homework and Math achievement. The presence of the triangle with “1” inside tells you that means and intercepts were analyzed (you can actually get the same results in your general statistics program by regressing Math on Homework and a variable with a constant value of 1, which is what this figure symbolizes). The path from the triangle to the exogenous variable (Homework) shows its mean; the path from the triangle to the endogenous variable (Math) shows its intercept. The curved double-headed arrows pointing to Homework and the disturbance are the variances (these represent the covariance of the variable with itself, i.e., the variance).

Of course the example used here includes measured means and intercepts rather than latent means and intercepts; the figural display will easily generalize, however. I prefer the format shown in the initial display (Figures 19.2 and 19.3) and will use it in this and subsequent presentations of SEM with mean structures. You should be familiar with the display using a triangle to signify the estimation of means, however, because it is common (see, for example, Hancock & Mueller, 2013).

Estimation of Means and Intercepts in Single Group SEM Models

Figure 19.5 shows a variation of the latent variable homework model first analyzed in Chapter 18. The model has been simplified by the exclusion of the Ethnic Minority variable. The data used are different as well. The data (homework means.sav) are a subsample of the NELS data, but a different subset than the primary data set we have been using, including data from 8th through 12th grades. Extraneous variables have been deleted from the data set to simplify it, as well.

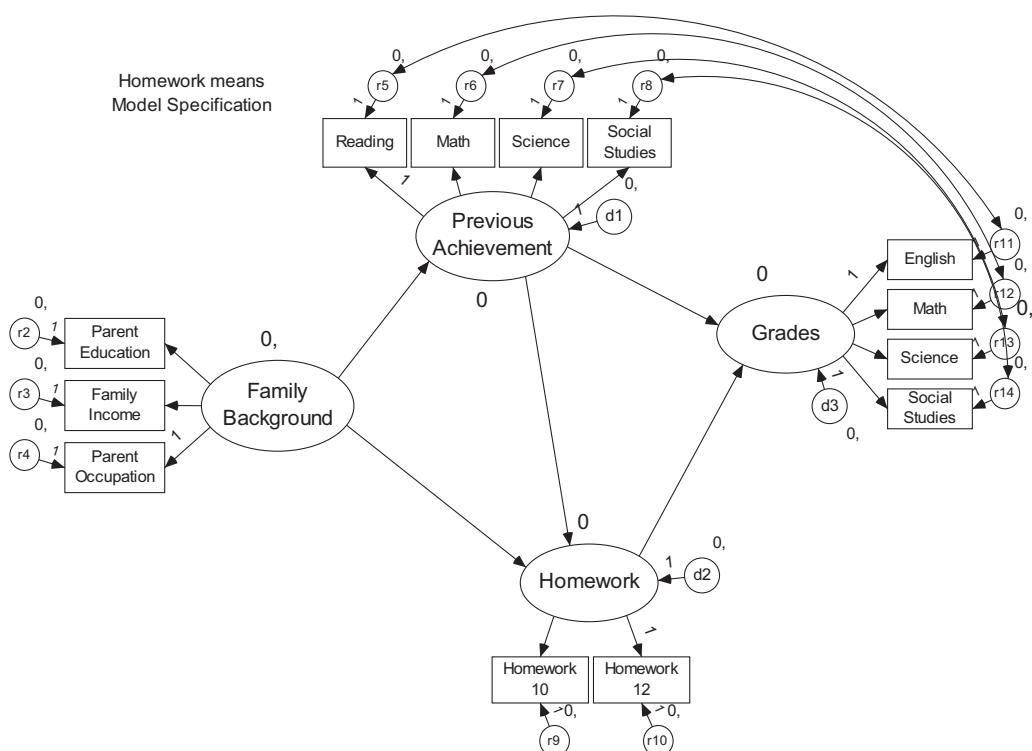


Figure 19.5 Setup for the latent variable homework model; means and intercepts included.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
parocc PARENT OCC STATUS COMPOSITE	991	7.32	81.87	51.5851	21.35184
byfaminc Family Income	958	1.00	15.00	9.8518	2.58663
bypared PARENTS' HIGHEST EDUCATION LEVEL	999	1	6	3.13	1.259
bytxrstd READING STANDARDIZED SCORE	971	23.098	67.499	51.29714	9.996038
bytxmstd MATHEMATICS STANDARDIZED SCORE	970	30.282	71.222	51.54440	9.891007
bytxsstd SCIENCE STANDARDIZED SCORE	969	26.505	75.973	51.20633	10.003315
bytxhstd HISTORY/CIT/GEOG STANDARDIZED SCORE	968	24.183	69.508	51.41394	9.687334
eng92 average grade in english	980	.00	11.25	6.1230	2.64740
math92 average grade in math	981	.00	11.50	5.5223	2.62251
sci92 average grade in science	986	.00	11.50	5.7884	2.63468
soc92 average grade in social studies	989	.00	11.67	6.2334	2.80404
f1s36a2 TIME SPENT ON HOMEWORK OUT OF SCHOOL	954	0	7	2.53	1.683
f2s25f2 TOTAL TIME SPENT ON HMWRK OUT SCHL	904	0	8	3.37	1.965
Valid N (listwise)	798				

Figure 19.6 Descriptive statistics for the variables in the homework latent variable model.

Descriptive statistics for the variables in the model are shown in Figure 19.6. Most of the variables in this model have been described previously. The two homework variables (f1s36a2 and f2s25f2) are student self reports of time spent on homework out of school in 10th and 12th grades. The measured grades variables are recorded from students' transcripts.

The model setup in Figure 19.5 looks quite similar to previous models. The big difference between these models and those we have analyzed previously is the presence of the values of zero beside all of the latent variables. These values represent the latent means of exogenous variables (Family Background, r1, d1, and others) and the latent intercept for the endogenous variable (Homework, Grades). When we first began discussing latent variables (in the chapter on CFA), I noted that latent variables have no natural scale, and therefore we have to set the scale of latent variables either by setting a single factor loading to one (ULI) or by setting the latent variable variance to one (UVI). Likewise, latent variables have no natural mean, and we generally set the *means* of all latent variables to zero.

These model changes are accomplished by specifying that means and intercepts are to be analyzed. This specification will be program-specific. For example, in Amos, this is accomplished by selecting "Estimate means and intercepts" under "Estimation" under "Analysis Properties" (Figure 19.7). Amos automatically sets the means and intercepts for latent variables to 0 when this option is chosen. In Mplus, in contrast, the estimation of mean structures is the default. The estimation of means and intercepts is turned off via a MODEL=NOMEANSTRUCTURE option as a part of the ANALYSIS command.

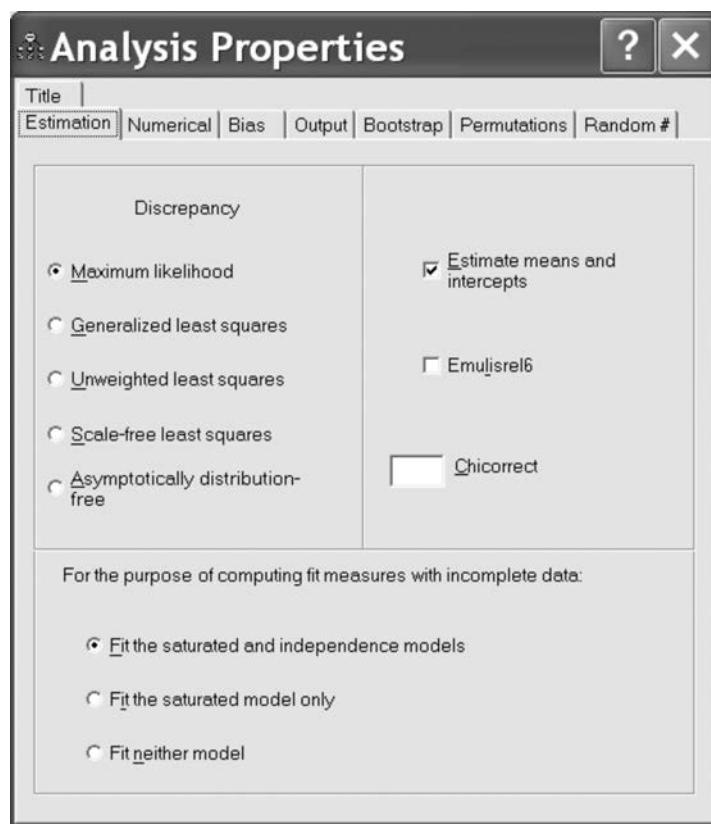


Figure 19.7 Amos setup for estimating mean structures.

The standardized output from the analysis in which means and intercepts are explicitly estimated looks identical to an analysis without that estimation (Figure 19.8). Indeed, without other constraints, the fit of the model and even the degrees of freedom are the same as they would be if means and intercepts were not analyzed. As shown in the Figure, the model fits well; the CFI of .993 and the RMSEA of .032 are better than our normal rule-of-thumb values. Although not shown in the Figure, the standardized root mean square residual (SRMR) was also good (.023; more on this later). Given a good fit, we interpret the parameters. The results suggest that out-of-school Homework has a strong direct effect on high school Grades (.32), and that Previous Achievement and Family Background characteristics have strong and moderate effects on time spent on Homework, respectively.

With the explicit estimation of means and intercepts, the results for the unstandardized model are considerably more complex than in previous models. Figure 19.9 shows a portion of the unstandardized output with the new parameters labeled. Each of the latent variables, including the residuals of the measured variables and the disturbance of the latent variable, has a mean (or intercept) of zero. These are means for exogenous variables, and intercepts for endogenous variables. What's the difference? One way of thinking about this is to say that if a variable is not influenced by other variables, we then estimate its mean. In contrast, we estimate intercepts for variables that are influenced by other variables. Said differently, any latent variable that has an arrow pointing to it will have an intercept. This includes the Homework latent variable in the Figure. Any latent variable with no arrows pointing to it will have a mean. This includes Family Background, r9, r10, and d2. Again, all of these means and intercepts of the *latent* variables are set to zero.

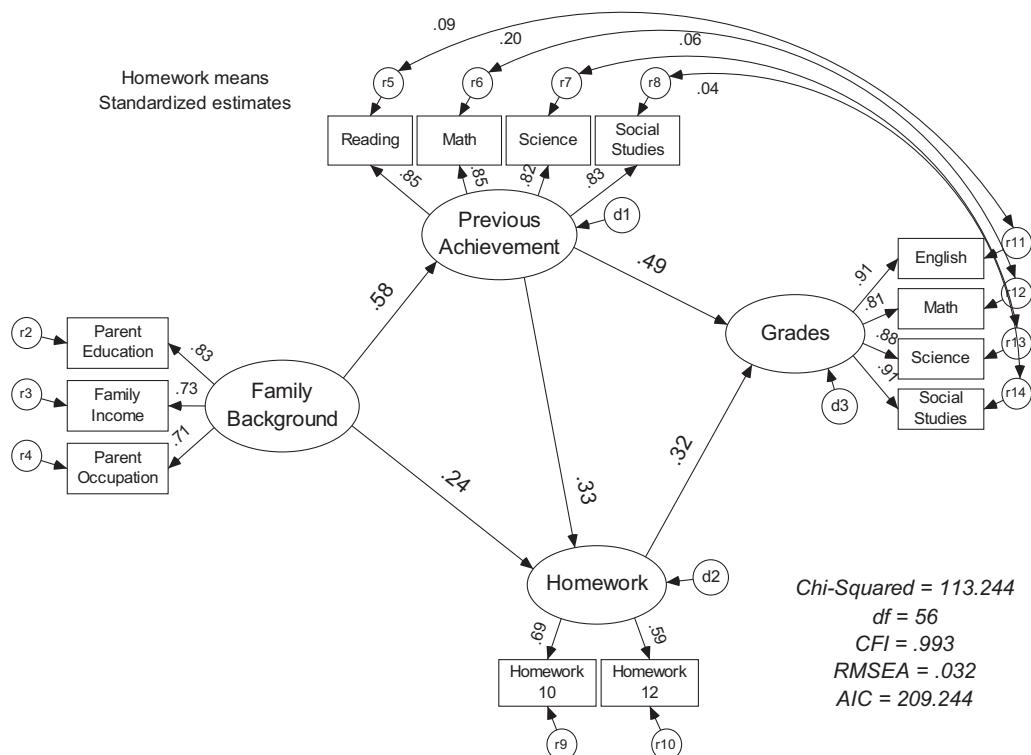


Figure 19.8 Standardized estimates for the homework latent variable model.

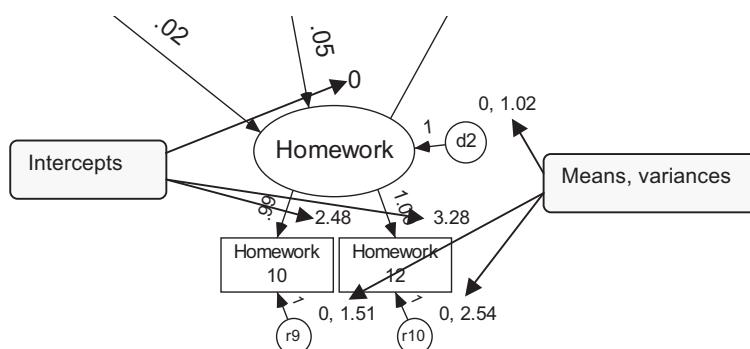


Figure 19.9 Detail from the unstandardized model results showing the location of means and intercepts in the homework latent variable model.

In contrast, these values (means and intercepts) are freely estimated for the measured variables, that is, they are not constrained to zero. Thus, the numbers (2.48 and 3.28) above and to the right of the two Homework measured variables are the measured intercepts. Why intercepts rather than means? Because the measured variables are endogenous; Homework 10 is caused, in part, by the latent Homework variable.

This distinction between means and intercepts may be confusing at first, but just think of intercepts as being related to the means, but controlling for the other influences in the model. Recall from multiple regression that intercepts are where the regression line crosses the Y axis; they are the predicted score on the outcome for those with a value of zero on

Intercepts:

Variable	Estimate	S.E.	C.R.	P
parocc	51.388	.679	75.639	***
byfaminc	9.841	.083	118.357	***
bypared	3.128	.040	78.528	***
bytxrstd	51.257	.320	160.308	***
bytxmstd	51.493	.316	162.994	***
bytxsstd	51.179	.320	160.071	***
bytxhstd	51.373	.310	165.851	***
eng92	6.074	.085	71.860	***
math92	5.482	.084	65.648	***
sci92	5.770	.084	68.619	***
soc92	6.207	.089	69.581	***
f2s25f2	3.280	.065	50.293	***
f1s36a2	2.481	.055	45.432	***

Implied Means

f1s36a2	f2s25f2	soc92	sci92	math92	eng92	bytxhstd	bytxsstd	bytxmstd	bytxrstd	bypared	byfaminc	parocc
2.481	3.280	6.207	5.770	5.482	6.074	51.373	51.179	51.493	51.257	3.128	9.841	51.388

Figure 19.10 Intercepts versus implied means, homework model.

the influence(s). So, for those with a value of zero on the latent Homework variable, their predicted, or average, score will be 2.48 on Homework 10 and 3.28 on Homework 12. If you want to know the model-predicted score for any other value of the influence, just substitute that value for X in the simple regression equation: $Y' = a + bX$. Because the mean of the latent Homework variable is zero, these values (2.48 and 3.28) also represent the model-implied means for Homework 10 and Homework 12, respectively. Figure 19.10 shows some of the more detailed text output from this same analysis. Note in Figure 19.10 that the intercepts of all of the measured variables are equal to the model-implied means for those variables. This is because the only influences on these measured variables are the latent variables, and all of these latent variables have means of zero, a function of the assumption that latent variables have means of zero.

This equality between the measured intercepts and the model-implied means will not hold when we start estimating means across multiple groups, because with multi-group analyses we will be able to actually estimate means and intercepts for the latent variables for some of the groups (technically, we will estimate mean and intercept *differences* from one group to another). As we will see, this is a major reason for adding the estimation of means and intercepts to SEM: to examine differences across groups in *latent* means and intercepts.

Note that the values listed as implied means are also close to the actual means calculated in SPSS (Figure 19.6). In fact, the values would be identical, except that SPSS and Amos (and most other SEM programs) calculate means differently when there are missing data in a data set. Most SEM programs use maximum likelihood methods whereas SPSS uses listwise deletion or simply calculates each mean separately. More on this later.

If you still find the distinction between means and intercepts confusing, take heart that you are not alone. And most of the time the distinction is not that important, either. Perhaps for this reason, many writers simply refer to this process of including means and intercepts in the model as the analysis of “mean structures.” I will do so, as well.

Related Points

I mentioned earlier that the fit of this model would be the same whether we estimated mean structures or not. In fact, all parameter values—unstandardized and standardized paths and factor loadings—would be the same whether we estimated mean structures or not. You may

be wondering why we are bothering with the estimation of mean structures in SEM if the fit and parameter values are the same either way. Why go through all the hassle for no additional payoff?

Next Steps

There are several valid reasons for estimating mean structures. First, this discussion has really been preparatory work for estimating means and intercepts in multi-group SEM and CFA. As you will soon find out, analysis of mean structures adds some very interesting information to multi-group analyses. With multiple groups, it is possible to estimate latent means and intercept differences for all groups except the first. Just as we think of latent variables as getting closer to the construct level—true Homework rather than measured Homework, true Happiness rather than a simple survey-reported measured-variable version of happiness—we can think of latent means as getting closer to the *true mean differences* across groups. Estimating differences in latent means and intercepts in SEM will be the focus of most of the remainder of this chapter. Estimating such differences in CFA will be one focus of the next chapter.

Missing Values

The second reason for estimating means and intercepts in SEM when using Amos is that Amos *requires* this addition in analyses of raw data in which there are missing values. Recall that for all analyses preceding this one we have either analyzed matrix data or raw data sets in which there are no missing data. The data set for this example includes missing data for each of the measured variables in the model (note the *Ns* in Figure 19.6). If you were to try to estimate the homework model without clicking on the “Estimate means and intercepts” option, the analysis would not run and Amos would return the error message: “In order to analyse data with missing observations, you must explicitly estimate means and intercepts.” You now know how to do that.

Unfortunately, Amos does not provide all of the output we like when there are missing observations. In particular, when there are missing data the detailed output will no longer include standardized residuals or modification indices, the information we found useful for figuring out possible modifications to models. Bootstrapping is also not allowed, and Amos will no longer calculate the SRMR. These are disadvantages to the analysis of raw data with missing values in Amos. One possibility is to analyze matrix data in the preliminary stages of data analysis but then to double-check all analyses via the analysis of raw data. Another option is to use a program without these limitations. Although Mplus requires the estimation of mean structures when data are missing, it will compute SRMR and conduct bootstrapping with missing data. Indeed, the SRMR listed for the latent homework model earlier in this chapter (.023) was calculated using Mplus.

You may wonder why use raw data in an SEM program when there are missing data? Why not generate a matrix in SPSS using one of its missing data methods, listwise or pairwise deletion, and then analyze the matrix in the SEM program? Or why not just get rid of all missing data, a strategy that is equivalent to listwise deletion? The reason is that Amos and other SEM programs use a more sophisticated method for dealing with missing data, generally referred to as full information maximum likelihood (FIML) estimation. Missing data are ubiquitous in research. Traditional methods of dealing with missing data can distort estimates of means, covariances, and variances (Wothke, 2000). Modern methods, including maximum likelihood methods, generally come closer to estimating model parameters accurately, and are recommended by methodologists (Enders, 2010; Enders & Bandalos, 2001; Graham, 2009; Muthén, Kaplan, & Hollis, 1987; Schafer & Graham, 2002). The issue of missing values will be discussed further in Chapter 22.

Calculating Degrees of Freedom

Because additional parameters are analyzed, calculating degrees of freedom is slightly different when means are analyzed. For the current example (Figure 19.5) there are 13 measured variables, so there are $\frac{p \times (p+1)}{2}$, or $\frac{13 \times 14}{2} = 91$ items in the variance covariance matrix, plus the means of the 13 measured variables, for a total of 104. An alternative formula when estimating means structures is $\frac{p \times (p+3)}{2} = \frac{13 \times 16}{2} = 104$ pieces of information in what we will now refer to as the “moment” matrix.

How many parameters are freely estimated in the model? There are:

1. 5 paths
2. 9 factor loadings
3. 4 error covariances
4. 13 error variances, 3 disturbance variances, and the variance of the 1 exogenous latent variable
5. 13 measured variable intercepts for a total of 48 freely estimated parameters. The $df =$ the number of moments (means, variances, covariances) minus the number of estimated parameters = $104 - 48 = 56$. As shown in Figure 19.8, the df for the model are indeed 56.

OVERVIEW: TWO METHODS TO TEST FOR DIFFERENCES IN LATENT MEANS

As noted above, this introductory work is really prep work to lay the groundwork for estimating *latent* means and intercepts. Before beginning this topic in earnest, it is worth examining a quick overview to get a general sense of what will be covered. Figure 19.11 shows

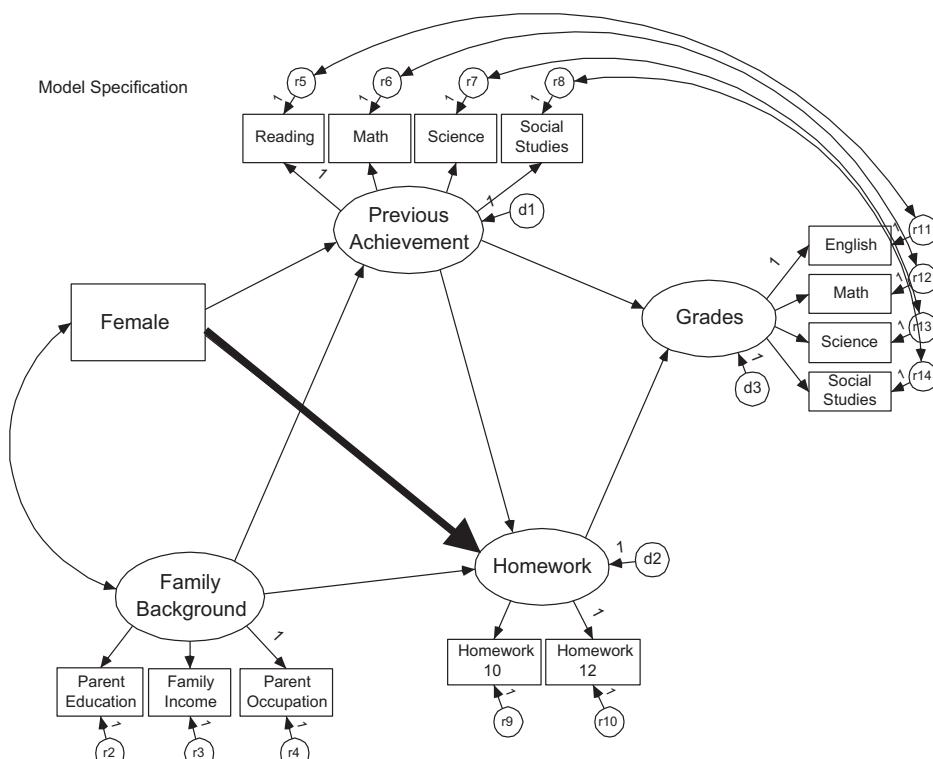


Figure 19.11 One method of testing for mean and intercept differences in SEM.

a variation of the latent variable homework model from Chapter 18. This model includes the exogenous dummy variable Female, coded 1 for girls and 0 for boys. Note in the model the bolded path from Sex to Homework: what does it represent? If this path were statistically significant and positive, it would suggest that girls, on average, do more homework than boys (taking into account Family Background and Previous Achievement). The finding would suggest a higher mean on Homework for girls versus boys (strictly speaking, it would suggest a higher intercept for girls versus boys. Imagine two parallel regression lines [like those shown in Chapter 7], one for boys and one for girls, with the girl line higher than the boy line). Said differently, this finding would show a *main effect* for Sex on Homework. This is one method of estimating latent means and intercepts in an SEM model. In contrast, if the path were statistically significant and negative, it would suggest that, other things being equal, boys report more homework than girls.

As shown in Chapter 18, it is also possible to test for interactions (aka moderation) between categorical and continuous variables in SEM through the use of multi-sample, or multigroup (MG), models. With the current example, this would involve removing Female from the model and conducting a multi-group analysis with one model for boys and one for girls (Figure 19.12). For this MG model, a difference in the magnitude of the (unstandardized) path from Homework to Grades for boys versus girls would suggest that Homework had differential effects on Grades for the sexes. If, for example, the path were larger for girls than boys, the finding would suggest that homework has a larger effect on grades for girls and that each additional hour spent on homework has a bigger effect on the grades of girls than boys. We would likely test the statistical significance of the difference by constraining the boy path and the girl path to be the same and examining the change in fit of the model (as is done in Figure 19.12). Again, this multi-group approach tests whether Sex and Homework interact in their effect on Grades. Or, to use the “it depends” method of describing interactions (see Chapters 7 and 18), if someone were to ask you the extent of the effect of Homework on Grades, you would need to answer, “it depends on whether you are a boy or a girl.” Alternatively, we might say that such findings suggest that sex moderates the effect of homework on grades.

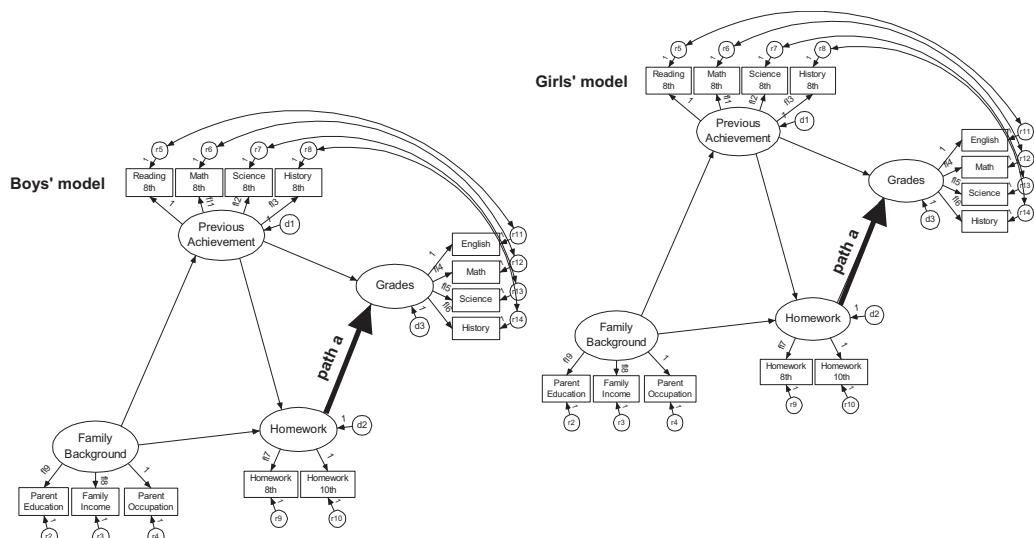


Figure 19.12 Multi-group analysis to test for differential effects of Homework on Grades, by Sex. The boys' model is on the left, the girls' model on the right.

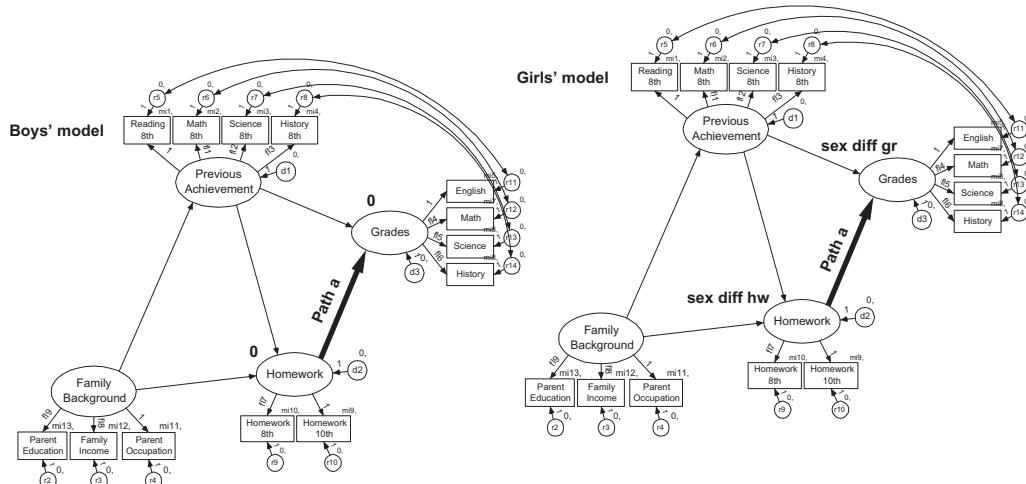


Figure 19.13 Multi-group model with means and intercepts. This model estimates both mean and intercept differences and differential effects (interactions) in the same analysis.

The second method for testing mean structures in SEM builds on the multi-group approach but tests for main effects and interactions in a single analysis (see Figure 19.13). This approach is often referred to as the “multi-group mean and covariance structures,” or MG-MACS, approach. As in the previous multi-group approach, separate models are specified for boys and girls; the path from Homework to Grades can be constrained versus freed across groups in order to test the statistical significance of the interaction. In addition, however, means and intercepts are estimated in the multi-group model. The statistical significance of the main effect of sex on homework is tested by comparing a model with the boy and girl Homework intercepts both constrained to zero versus a model in which the Homework intercept is freed for one group (in this case, girls). Given certain other constraints, the value for the girl intercept for Homework (sex diff hw in the right half of the figure) would equal the unstandardized path from Sex to Homework in the model in Figure 19.11; in both cases, this value represents the *difference* in intercepts for girls as compared to boys. Think of this as the true mean difference in homework time for boys versus girls.

EXAMPLE: HYPNOSIS FOR HOT FLASHES

Single Group/Dummy Variable Approach

Elkins and colleagues used hypnosis as an experimental treatment to control hot flashes among women who were breast cancer survivors; menopause and hot flashes are a common side effect of chemotherapy (Elkins et al., 2008). Sixty women with hot flashes were randomly assigned to a hypnosis intervention (five weeks) or a no-treatment control group. A variety of outcomes, including hot flash frequency and severity, were assessed in a pretest–posttest control group design. Results were analyzed using multivariate analysis of covariance (MANCOVA). The hypnosis group showed a large and statistically significant reduction in hot flash frequency and severity in comparison with the control group.

Here, we will analyze a simulated version of the Elkins data using SEM, and using the two methods described to analyze means and intercepts. The simulated raw data (“hot flash

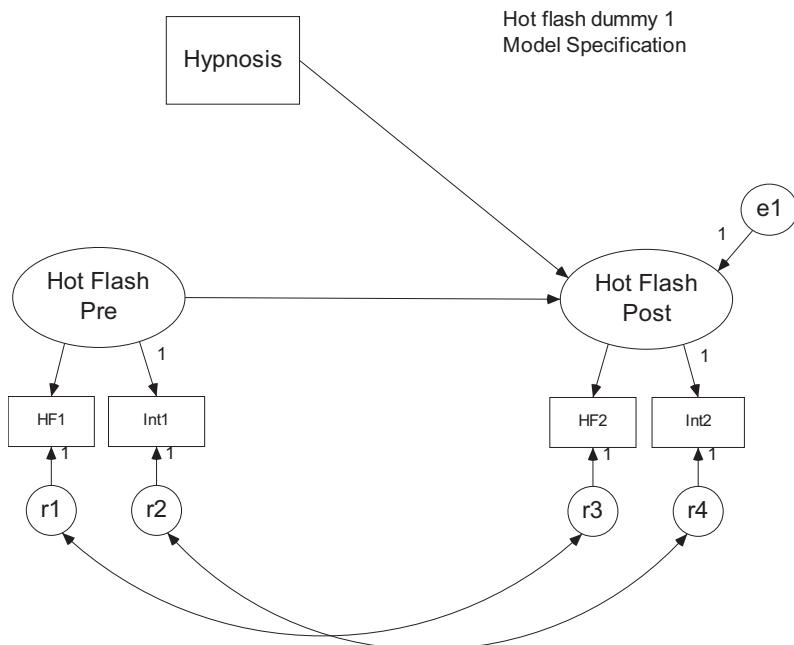


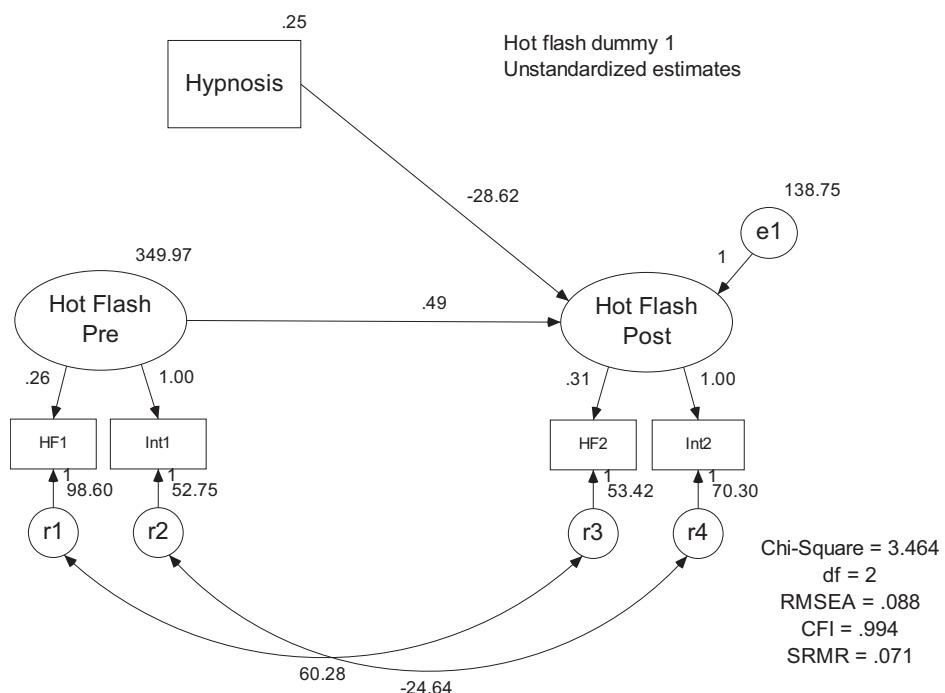
Figure 19.14 Dummy variable model designed to test the effect of hypnosis on hot flash severity and interference.

simulated.sav") include a larger sample, with 48 women each in the control and hypnosis experimental groups.

The initial model is shown in Figure 19.14. Five measured variables are included in the model. Four are measures related to hot flashes: hot flash scores (a combination of ratings of frequency and severity of hot flashes, from daily diaries) from pretest and posttest (HF1 and HF2) and Interference scores from pre- and posttest (ratings of the degree that the hot flashes interfere with daily life, Int1 and Int2). The Hypnosis variable is a dummy variable coded 0 for women in the control group and 1 for those in the experimental (hypnosis) group (the variable is labeled Group in the raw data). The pretest hot flash measures are used as indicators of a latent Hot Flash Pretest score, with the posttest measures used as indicators of a latent Hot Flash Posttest score. For both the measured and latent hot flash variables higher scores represent worse outcomes, that is, more frequent, severe, and interfering hot flashes. The model allows cross-time correlations among the errors of measurement (residuals) because these are the same measures administered on two occasions. Note also that there is no correlation (covariance) between the Hypnosis variable and Hot Flash Pretest scores. Because assignment to groups was random, group membership (Hypnosis vs. Control) should be unrelated to the initial severity of hot flashes. This is an assumption that could be tested in the analysis.

The model is similar to those analyzed in previous chapters. There is no explicit estimation of means and intercepts in this model; instead, the intercept for the latent Hot Flash Posttest will be shown by the path from the dummy Hypnosis variable to Hot Flash Posttest. We could add the explicit estimation of means and intercepts, but it would add little to the example. In addition, I want to show similarities and difference between this type of analysis, done in previous chapters, and the MG-MACS approach. The descriptive statistics for the data are shown, by group, in Figure 19.15.

Report					
Group		HF1	int1	HF2	int2
0 Control	Mean	17.0769	46.3125	15.5078	42.2500
	Std. Deviation	10.82255	21.39413	11.20556	21.84228
	Minimum	1.50	2.00	.64	10.00
	Maximum	38.01	82.00	39.07	93.00
	N	48	48	48	48
1 Hypnosis Intervention	Mean	14.3963	39.0000	5.0361	10.8750
	Std. Deviation	11.34721	18.37320	5.08552	10.73595
	Minimum	1.56	9.00	.00	.00
	Maximum	46.35	69.00	22.31	34.00
	N	48	48	48	48
Total	Mean	15.7366	42.6562	10.2719	26.5625
	Std. Deviation	11.11146	20.17337	10.13013	23.27538
	Minimum	1.50	2.00	.00	.00
	Maximum	46.35	82.00	39.07	93.00
	N	96	96	96	96

Figure 19.15 Descriptive statistics for hot flash data.**Figure 19.16** Unstandardized results, hot flash model. Women in the hot flash group scored 28 points lower on the hot flash latent variables, on average, compared to women in the control group.

The unstandardized results are shown in Figure 19.16. The model shows an adequate fit to the data. CFI is good ($> .95$), and χ^2 is statistically non-significant. The RMSEA of .088 is not as low as we would like, although RMSEA often behaves this way in small samples (Hu & Bentler, 1998), with small models with few degrees of freedom (Kenny,

Kaniskan, & McCoach, 2011). The average difference in correlations for the actual versus predicted matrix was .071. Again, taken together, these fit indices suggest an adequate fit of the model to the data.

The most interesting finding for the model is the path from Hypnosis to the Hot Flash Posttest latent variable (-28.62). Because the Hypnosis dummy variable was coded 0 for those in the control group and 1 for those in the hypnosis (experimental) group, this coefficient means that those in the experimental group scored 28 points lower, on average, than those in the control group on the latent Hot Flash Posttest, controlling for pretest scores. What does that mean? There are several ways to understand this finding. The latent variable has the same scale as the Int2 measured variable because the path from the latent variable to Int2 was set to one. In other words, the Int2 variable was used to set the scale for the latent Posttest variable. The Int2 measured variable, in turn, has an overall standard deviation of around 23 points (total sample, Figure 19.15). This means that the hypnosis intervention resulted in a huge decrease in (latent) hot flashes, with the experimental group women scoring more than a *SD* below the control group women. As shown in the text output (Figure 19.17), this value is, not surprisingly, statistically significant. The standardized estimates in Figure 19.17

Regression Weights

	Estimate	S.E.	C.R.	P	Label
HFPost <--- Group	-28.615	3.165	-9.040	***	
HFPost <--- HFPre	.487	.196	2.487	.013	
Int1 <--- HFPre	1.000				
HF1 <--- HFPre	.260	.112	2.310	.021	
Int2 <--- HFPost	1.000				
HF2 <--- HFPost	.310	.037	8.436	***	

Standardized Regression Weights

	Estimate
HFPost <--- Group	-.693
HFPost <--- HFPre	.441
Int1 <--- HFPre	.932
HF1 <--- HFPre	.439
Int2 <--- HFPost	.927
HF2 <--- HFPost	.659

Covariances

	Estimate	S.E.	C.R.	P	Label
r1 <--> r3	60.277	10.877	5.542	***	
r2 <--> r4	-24.643	55.335	-.445	.656	

Correlations

	Estimate
r1 <--> r3	.831
r2 <--> r4	-.405

Figure 19.17 Standardized and unstandardized parameters estimates for the hot flash dummy variable model.

Table 19.1 Fit of Alternative Dummy Variable Hot Flash Models

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	RMSEA	SRMR	CFI
Initial	3.464	2				.088	.071	.994
Pretests Vary	.087	1	3.377	1	.066	.000	.007	1.000
No Effect	16.079	2	15.992	1	<.001	.272	.113	.939

Note: Each model compared to the previous model.

also show the importance of the hypnosis intervention; the path from Hypnosis (Group) to Hot Flash Post was $-.693$. For each SD change in Hypnosis group, Hot Flashes decreased by $.693 SD$ units. Because half of the women were in each group (control and hypnosis), the SD of the Hypnosis measured variable was $.5$. Therefore, the Hypnosis group scored $1.386 SDs$ ($2 \times .693$) lower on the Hot Flash latent variable than did the control group. I will not discuss the rest of the findings contained in Figures 19.16 and 19.17, but you should review the rest of the coefficients to make sure you understand them and can interpret them.

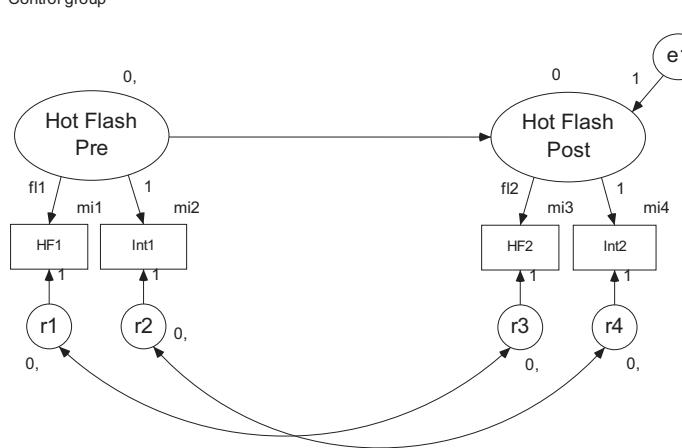
As noted earlier, it is possible to test the success of the random assignment to treatment groups. In the initial model there was no correlation allowed between Hypnosis group and pretest score. As shown in Table 19.1, allowing the Hypnosis group dummy variable and the Hot Flash Pretest to covary resulted in a reduction in χ^2 and improvement in RMSEA and SRMR, but the reduction χ^2 was not statistically significant. I would likely conclude that the randomization was successful and that the two groups were statistically equivalent on the latent pretest. In contrast, constraining the effect from Hypnosis group to Hot Flash Posttest to zero resulted in a statistically significant increase in χ^2 , further demonstrating that the hypnosis treatment had a statistically significant effect.²

Also noted earlier, Elkins et al (2008) analyzed their hot flash data using MANCOVA. Those familiar with MANOVA may wonder how the current analysis corresponds to MANOVA and MANCOVA. As noted in Chapter 1, MANOVA (and MANCOVA) are subsumed under SEM. MANOVA essentially combines the multiple dependent variables into a single latent dependent variable. An advantage of the SEM approach is that any covariates (the pretests) can also be modeled as one or more latent variables, thus reducing the effects of unreliability and invalidity. In MANCOVA, each measured pretest is considered separately, and the analysis assumes that the measures are completely reliable (cf. Arbuckle, 2017, examples 9 and 13). Recall from the chapter on error (Chapter 15) that unreliable exogenous variables are a particular danger in path analysis, regression, and other analyses based on the general linear model. Covariates in ANCOVA and MANCOVA are exogenous variables, and unreliability in them can affect the estimates of effects. For more information on the correspondence between SEM and MANOVA, see Cole, Maxwell, Avery, and Salas (1993) or Green and Thompson (2006).

MG-MACS Approach

Figure 19.18 shows the model setup to analyze the hot flash data via a multigroup approach that explicitly estimates means and intercepts, the MG-MACS approach. The upper model shows the setup for the control group and the lower model the experimental (hypnosis) group. As in other multi-group models, the categorical group variable is removed from the model but is used to differentiate the two models. That is, the upper model analyzes data only from the control group and the lower model analyzes data only for the experimental group.

Hot flash mgmacs 1
Model Specification
Control group



Hot flash mgmacs 1
Model Specification
Hypnosis group

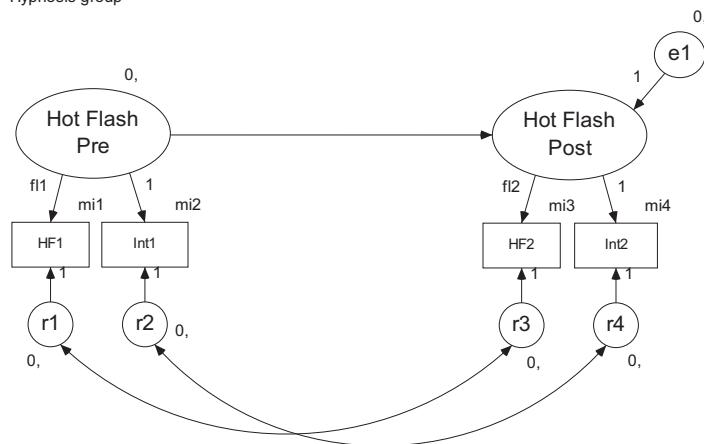


Figure 19.18 Analyzing the hot flash experiment via a MG-MACS model. Notice the model constraints across groups.

(To analyze two groups using a single raw dataset in Amos, you will need to tell the program what the grouping variable is [the variable Group in these data] and which value corresponds to which group. This is done in the same window used to specify the data set to use.)

As in our previous discussion of multi-group models, note the path from the latent Hot Flash Pretest to HF1 is set to fl1 (for factor loading 1) for both groups, meaning that the loading is estimated, but it is constrained to be equal across groups. Similarly, the loading of HF2 on the latent Hot Flash Posttest is constrained to fl2 for both groups. These constraints ensure that the latent variables reflect the same underlying constructs across groups. (This topic, invariance, will be further explored in the next chapter.)

We take these constraints one step further for the MG-MACS model. The values of mi1 through mi4 (for measured intercepts) constrain the intercepts for each of the measured

variables to be the same for one group as for the other. Setting the factor loadings to be equal across the groups puts the scales in the same metrics. Setting the intercepts of the measured variables to be equal gives those scale the same starting, or zero, point. Equality of measured variable intercepts is another aspect of invariance that we need to consider as we add the estimates of mean structures to our SEM models.

As in the previous means model illustration, most of the latent means (Hot Flash Pretest, r1, e1, and others) and latent intercept for the endogenous variable (Hot Flash Post) are set to zero. As noted earlier, we generally set the means of all latent variables to zero (this is done automatically for those using Amos by choosing to estimate means and intercepts). The exception is for the latent intercept for Hot Flash Posttest for the Hypnosis group. For this group, the value was not constrained and will be freely estimated. The result of this difference—constraining the latent intercept to 0 for one group and no constraint in the other group—is that the value for the second group is the *difference* in latent intercepts for the control versus the experimental group. This, then, is the test of the main effect of the hypnosis treatment.

Here is one way of thinking about what we are doing by constraining the intercepts and means. If you examine the data in Figure 19.15, it is obvious that the Hypnosis group scores lower than the control group on both of the hot flash post measures, HF2 and Int2. By constraining the measured intercepts to be equal (mi3, mi4) and allowing the latent intercepts to differ, we are saying that the only reason for this difference on HF2 and Int2 is because the *true* (latent, underlying) level of Hot Flashes differs across groups. In other words, this formulation says that the true mean level of hot flashes at posttest differs for women in the control and experimental groups, and this difference in the true (i.e., latent) variable is what causes the measured variables to differ. The difference in the measured variable means is fully explained by the difference in latent means. Another, mechanistic way of thinking of this is that by constraining the measured intercept to be equal we have forced any differences to show up at the latent variable level.

There are several other points worth mentioning:

1. As already noted, it is not possible to use this method to estimate latent means and intercepts for a single group (at least not without some other constraints). Of course, as we have already seen, it is possible to have the program estimate means and intercepts for a single group, but the *latent* means and intercepts must be set to zero for at least one group. Thus to estimate latent means and intercepts using this method, a multi-group approach must be used and the latent means and intercepts constrained to zero in one of the groups; it is then possible to free this constraint in the other group or groups to determine the difference in means and intercepts. It is also certainly possible to estimate means and intercepts across groups using the dummy variable approach we used earlier. We will soon see that there are some advantages to the MG-MACS approach, however.
2. When conducting experimental research, the group constrained to zero will likely be the control group in order to estimate the degree to which the other groups deviate from the control group.
3. Note that the pretest means (Hot Flash Pretest) were constrained to zero for both groups. Random assignment was used to assign women to groups and we have assumed that the groups are equal on the latent pretest. As in the single group dummy variable analysis, we could test the validity of this assumption, in this case by freeing the pretest mean for the Hypnosis group and examining the change in fit for the model. Likewise, we could constrain the latent intercept (Hot Flash Post) of the Hypnosis group to 0 to further test the hypothesis of no posttest difference for the Hypnosis group.

The unstandardized results for the MG-MACS models are shown in Figure 19.19 for the figural output and Figure 19.20 for the text output. The top portion of Figure 19.19 shows

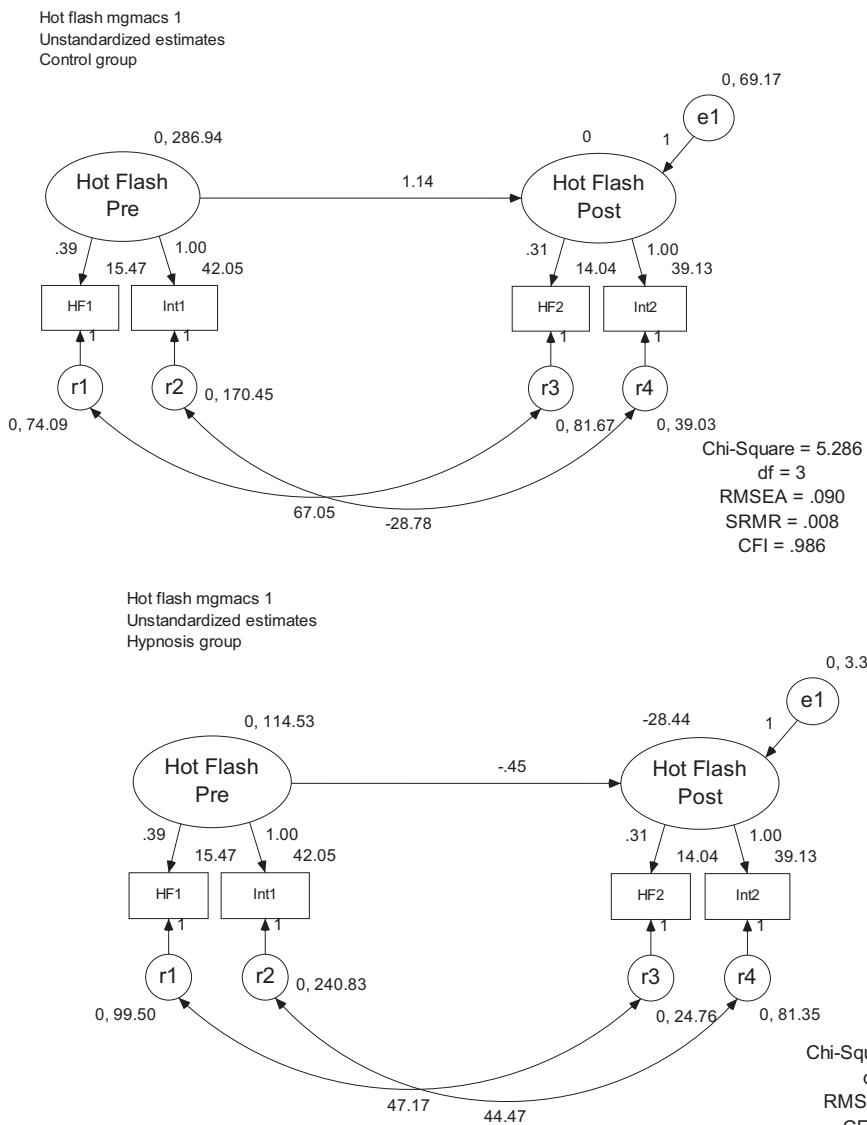


Figure 19.19 Hot flash results for the MG-MACS model. The intercept difference for the hypnosis group (shown above the Hot Flash Post latent variable) shows that women in the Hypnosis group scored 28 points lower, on average, than did women in the control group.

the findings for the control group. As shown in the Figure, this initial version of the Hot Flash model fit the data fairly well. The χ^2 was not statistically significant and the CFI was .986. The RMSEA of .09 was larger than our rule of thumb (corrected for two groups it would be even larger, .127), but the average difference between the actual and predicted correlation matrices was only .008.

Much of the output, such as the factor loadings and the path from Pretest to Posttest, are already familiar, and we will not spend time discussing them. Focus on the upper portion of the figure. The values above the latent Hot Flash Pretest represent its mean and variances; the mean was set to zero, and the variance is 286.94 (for the control group). For the Posttest,

Regression Weights: (Control group)

		Estimate	S.E.	C.R.	P	Label
Hot Flash_Post	<---	Hot Flash_Pre	1.139	.277	4.110	***
Int1	<---	Hot Flash_Pre	1.000			
HF1	<---	Hot Flash_Pre	.388	.116	3.350	*** f11
Int2	<---	Hot Flash_Post	1.000			
HF2	<---	Hot Flash_Post	.312	.037	8.346	*** f12

Regression Weights: (Hypnosis group)

		Estimate	S.E.	C.R.	P	Label
Hot Flash_Post	<---	Hot Flash_Pre	-.452	.460	-.982	.326
Int1	<---	Hot Flash_Pre	1.000			
HF1	<---	Hot Flash_Pre	.388	.116	3.350	*** f11
Int2	<---	Hot Flash_Post	1.000			
HF2	<---	Hot Flash_Post	.312	.037	8.346	*** f12

Standardized Regression Weights: (Control group)

		Estimate	
Hot Flash_Post	<---	Hot Flash_Pre	.918
Int1	<---	Hot Flash_Pre	.792
HF1	<---	Hot Flash_Pre	.607
Int2	<---	Hot Flash_Post	.959
HF2	<---	Hot Flash_Post	.587

Standardized Regression Weights: (Hypnosis group)

		Estimate	
Hot Flash_Post	<---	Hot Flash_Pre	-.936
Int1	<---	Hot Flash_Pre	.568
HF1	<---	Hot Flash_Pre	.384
Int2	<---	Hot Flash_Post	.498
HF2	<---	Hot Flash_Post	.309

Intercepts: (Control group)

		Estimate	S.E.	C.R.	P	Label
Int1		42.049	2.049	20.526	***	mi2
HF1		15.474	1.101	14.057	***	mi1
Int2		39.132	2.719	14.394	***	mi4
HF2		14.044	1.261	11.138	***	mi3

Intercepts: (Hypnosis group)

		Estimate	S.E.	C.R.	P	Label
Hot Flash_Post		-28.440	3.168	-8.976	***	
Int1		42.049	2.049	20.526	***	mi2
HF1		15.474	1.101	14.057	***	mi1
Int2		39.132	2.719	14.394	***	mi4
HF2		14.044	1.261	11.138	***	mi3

Covariances: (Control group)

		Estimate	S.E.	C.R.	P	Label
r1	<->	r3	67.054	18.192	3.686	***
r2	<->	r4	-28.775	99.659	-.289	.773

Covariances: (Hypnosis group)

		Estimate	S.E.	C.R.	P	Label
r1	<->	r3	47.166	10.684	4.415	***
r2	<->	r4	44.469	37.891	1.174	.241

Correlations: (Control group)

		Estimate	
r1	<->	r3	.862
r2	<->	r4	-.353

Correlations: (Hypnosis group)

		Estimate	
r1	<->	r3	.950
r2	<->	r4	.318

Variances: (Control group)

		Estimate	S.E.	C.R.	P	Label
Hot Flash_Pre		286.937	150.460	1.907	.057	
e1		69.165	50.508	1.369	.171	
r2		170.452	126.116	1.352	.177	
r1		74.086	19.745	3.752	***	
r4		39.028	108.603	.359	.719	
r3		81.670	19.874	4.109	***	

Variances: (Hypnosis group)

		Estimate	S.E.	C.R.	P	Label
Hot Flash_Pre		114.534	70.836	1.617	.106	
e1		3.336	52.506	.064	.949	
r2		240.833	75.359	3.196	.001	
r1		99.498	23.842	4.173	***	
r4		81.354	25.702	3.165	.002	
r3		24.764	5.666	4.370	***	

Figure 19.20 Detailed results for hypnosis on hot flashes, MG-MACS model.

the value shown (0) is the intercept, which was set to zero for the control group. As in the homework model, the values above and to the right of the measured variables are the intercepts for each of the measured variables. That is, these are the model-predicted values for the measured variables for those with a value of zero for the latent variable. (The intercepts are the predicted values for the outcome for those with a value of zero on the independent variable. For the measured variables, the independent, or exogenous, variable is the latent variable). Because the latent variables have means of zero, the values shown for the measured variable intercepts are the predicted means for the measured variables for the control group.

The lower portion of Figure 19.19 shows the values for the hypnosis experimental group. The primary finding of interest is the value for the intercept for Hot Flash Posttest: -28.44. This finding means that the Hypnosis group scored 28 points lower, on average, on the Hot Flash Posttest than did the Control group. This value is similar to but not identical with the value found in the dummy variable version of this research (-28.62). Note that there are several other differences between the two groups. The slope for the experimental group is negative (-.45), for example, whereas the value is positive for the control group (1.14).

Figure 19.20 shows the estimates for the Control and Hypnosis groups in tabular form, with the values for the Control group on the left and those for the Hypnosis group on the right. As shown in the table of intercepts, the value for the intercept for the Hypnosis group was indeed statistically significantly different from that of the Control group. The model assumes that the two groups' hot flash scores were statistically equivalent at the beginning of the experiment (and the generally good fit of the model would not suggest otherwise), but by the end of the five-week period the hypnosis intervention had led to statistically fewer, less severe, and less interfering hot flashes for the experimental group.

Figure 19.21 shows the values for the implied matrices for both groups. I noted previously that we estimate means for exogenous variables, but that we estimate intercepts for endogenous variables. Still, you may be interested in the values for the predicted means for the endogenous variables, including the measured variables. These are shown in the tables of implied means, and they represent the predicted values of the means given the model. If the computer program you are using uses a RAM-type notation (Figure 19.4), the predicted means might be shown as a part of the total effects of the constant on various outcomes. If you want to understand how these estimates come about, recall the general form of a regression equation: $Y' = a + bX$. The predicted value for a dependent (endogenous) variable is equal to the intercept plus the regression coefficient (path, slope, or factor loading) times the value for the independent (exogenous) variable. If we substitute the mean of X in this equation the outcome will be the predicted mean for Y . Thus the predicted mean value for the HF1 variable for the Control group is 15.47. This value is derived as follows: The mean for the latent Hot Flash Pretest = 0, slope = .39, intercept = 15.47, and

$$\begin{aligned} Y' &= a + bX \\ Y' &= 15.47 + .39 \times 0 \\ Y' &= 15.47 \end{aligned}$$

This compares to the actual value of 17.08. Note that if we had allowed the latent Pretest means for the Control and Experimental groups to differ, the value for the Pretest mean for the Experimental group would likely be some value other than zero.

Table 19.2 compares the fit of the initial model to one in which the Pretest scores were allowed to differ across groups. As shown in the Table, this model fit better than did the initial model. The difference in $\Delta\chi^2$ was not statistically significant, however, and given our

Implied (for all variables) Covariances (Control group)							Implied (for all variables) Covariances (Hypnosis group)						
	Hot Flash_Pre	Hot Flash_Post	HF2	Int2	HF1	Int1		Hot Flash_Pre	Hot Flash_Post	HF2	Int2	HF1	Int1
Hot Flash_Pre	286.937						Hot Flash_Pre	114.534					
Hot Flash_Post	326.818	441.407					Hot Flash_Post	-51.799	26.762				
HF2	102.025	137.797	124.686				HF2	-16.170	8.355	27.372			
Int2	326.818	441.407	137.797	480.435			Int2	-51.799	26.762	8.355	108.116		
HF1	111.352	126.828	106.647	126.828	117.298		HF1	44.447	-20.102	40.891	-20.102	116.747	
Int1	286.937	326.818	102.025	298.042	111.352	457.389	Int1	114.534	-51.799	-16.170	-7.330	44.447	355.367

Implied (for all variables) Correlations (Control group)							Implied (for all variables) Correlations (Hypnosis group)						
	Hot Flash_Pre	Hot Flash_Post	HF2	Int2	HF1	Int1		Hot Flash_Pre	Hot Flash_Post	HF2	Int2	HF1	Int1
Hot Flash_Pre	1.000						Hot Flash_Pre	1.000					
Hot Flash_Post	.918	1.000					Hot Flash_Post	-.936	1.000				
HF2	.539	.587	1.000				HF2	-.289	.309	1.000			
Int2	.880	.959	.563	1.000			Int2	-.465	.498	.154	1.000		
HF1	.607	.557	.882	.534	1.000		HF1	.384	-.360	.723	-.179	1.000	
Int1	.792	.727	.427	.636	.481	1.000	Int1	.568	-.531	-.164	-.037	.218	1.000

Implied (for all variables) Means (Control group)							Implied (for all variables) Means (Hypnosis group)						
	Hot Flash_Pre	Hot Flash_Post	HF2	Int2	HF1	Int1		Hot Flash_Pre	Hot Flash_Post	HF2	Int2	HF1	Int1
Hot Flash_Pre	0	0	14.044	39.132	15.474	42.049	Hot Flash_Pre	0	-28.44	5.166	10.692	15.474	42.049

Figure 19.21 Implied matrices and means, MG-MACS results.

Table 19.2 Comparison of MG-MACS Hot Flash models to the Initial Model and the revised Initial Model (Initial 2)

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	RMSEA	SRMR	CFI
1. Initial	5.286	3	—	—	.090	.008	.986	
2. Pretests Differ	1.789	2	3.497 ^a	1	.061	.000	.007	1.000
3. Test Assumptions	84.615	12	79.329 ^b	9	<.001	.254	.128	.566
4. Slopes Vary	44.252	11	40.363 ^a	1	<.001	.179	.047	.801
5. Initial 2	6.927	5	1.641 ^b	2	.440	.064	.012	.988
6. No Main Effect	67.312	6	60.385 ^c	1	<.001	.330	.120	.633
7. No Slope Difference	22.496	6	15.569 ^c	1	<.001	.171	.011	.901

^aModel compared to previous model

^bModel compared to Initial model

^cModel compared to Initial 2 model

rule of thumb (if $\Delta\chi^2$ is not statistically significant, stick with the more constrained model), I would continue to focus on the initial model for interpretation.

Comparing the Two Methods

Why bother with the MG-MACS approach? It seems like a lot of work and we can get the same information treating the categorical variable as a dummy variable and using it in the analysis. There are several reasons, but the primary one is that the dummy variable approach requires several assumptions, but those assumptions go untested in that approach. MANCOVA required but did not test the assumption of perfectly reliable covariates. That assumption was not required (but could be tested) using the dummy variable SEM approach. Similarly, using the MG-MACS approach, we can evaluate the assumptions made with, but not tested by, the dummy variable approach.

What are those assumptions? Most have to do with equality constraints across the two groups (Control and Hypnosis). With the dummy variable approach, as in Figures 19.14 and 19.16, almost all parameters were constrained to be equal for the Control and the Hypnosis groups, because the two groups were analyzed in a single model. Factor loadings and measured intercepts were constrained to be equal in both models (we consciously constrained them for the MG-MACS model), and this level of invariance is needed in order to compare latent means and intercepts. In addition, however, the error variances (r_1 , r_2 , e_1 , etc.) and the error covariances were also constrained to be equal across groups. Again, these constraints are a function of the fact that there are not separate groups in the analysis in the dummy variable approach, and thus it was not possible to allow different estimates across groups. Finally, the path from Hot Flash Pretest to Hot Flash Posttest (the slope of the regression of the latent posttest on the latent pretest) was constrained to be equal across groups. We saw in the section on multiple regression that the requirement for equal slopes in ANCOVA is not always reasonable, and the similar requirement for equal slopes in the dummy variable approach to estimating means in SEM may also not be reasonable.

It is possible to test the validity of these constraints using a MG-MACS model. Figure 19.22 shows the constraints needed for such a model, with the Control group model above and the Hypnosis group model below. Note that all the parameters, except one, are constrained to be equal across the two groups: factor loadings, intercepts, variances, error

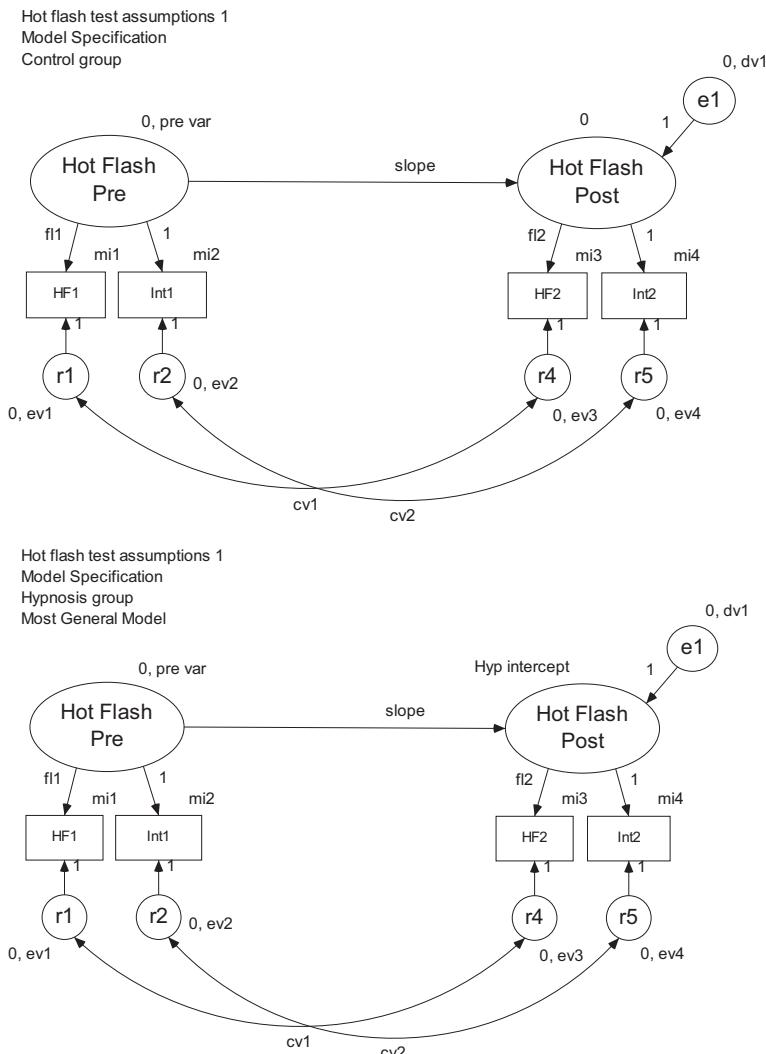


Figure 19.22 Model constraints needed to obtain the same results the MG-MACS model as for the dummy-variable version of the hot flash model. These represent assumptions made but not tested in the dummy-variable model.

variances, covariances, pretest means, and slopes. The one parameter allowed to vary across the groups was the intercept for the Hot Flash Posttest. This parameter was set to zero for the Control group but was named Hyp_intercept and freely estimated for the Hypnosis group. If we have set this model up correctly, with the right mix of constraints and free parameters, the parameter estimates, such as the differences in intercepts, should be identical to the findings for the initial dummy variable model because it has the same restrictions as that model.

Figure 19.23 shows the unstandardized results for the hypnosis group. Note first the value for the intercept for the Hypnosis group: -28.62, a value identical to that shown for the path from the group dummy variable to the Hot Flash Posttest in Figure 19.16. Note that all other values—factor loadings, intercepts, and so on—are the same as those in the initial dummy variable model. Now note how poorly this model fits the data, with a RMSEA of

Hot flash test assumptions 1
Unstandardized estimates
Hypnosis group
dummy variable assumptions

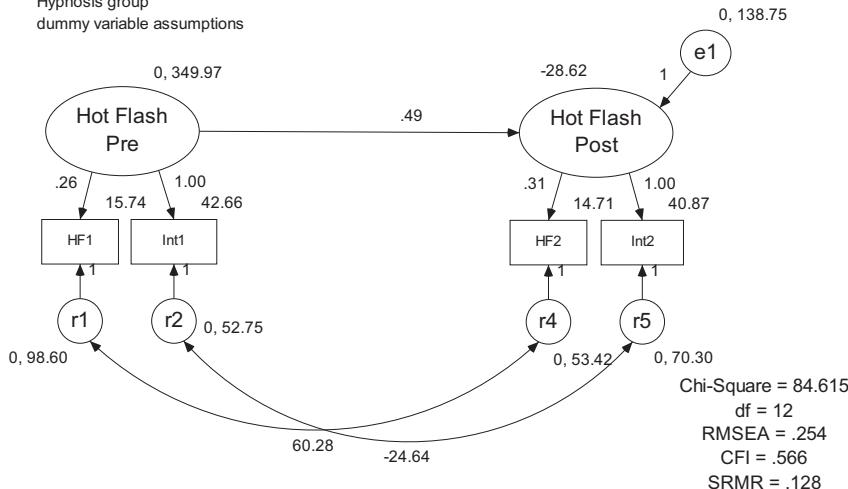


Figure 19.23 Model results when testing the assumptions underlying the dummy variable version of the hot flash model. Results are for the hypnosis group.

.254 (.359 corrected) and a CFI of .566, for example. What does this finding mean? The model fit well when we analyzed it using the dummy variable approach, but now when we analyze it using the MG-MACS approach, the model shows a poor fit. The difference between the two approaches is that with the MG-MACS we are now testing assumptions that were hidden using the dummy variable approach. And when tested, those assumptions are not supported. Said differently, we made some implicit assumptions about the equality of various parameters in the dummy variable approach, but when we tested those equality constraints using the MG-MACS approach they were not supported. This model and the initial model are nested, so we can also compare the fit of the two. This comparison is done in Table 19.2, and, as shown, this model, termed the Test Assumptions model, fit statistically significantly worse than did the initial model.

As noted earlier, the assumption that the slopes are equal across groups may be especially suspect, and an inspection of the models in which the slopes were freely estimated (Figure 19.19) shows that the values for the Control and the Hypnosis groups are quite different. Model 4 in Table 19.2 freed this constraint. Thus the “Slopes Vary” model kept all of the restrictions of the Test Assumptions model but allowed the slopes to vary. As shown in the Table, this relaxation of invariant slopes resulted in a statistically significant improvement in model fit over the Test Assumptions model. Nevertheless, this model also fit statistically significantly worse than did the initial model ($\Delta\chi^2 = 38.966$ [8], $p < .001$). We would likely stick with the initial model, with fewer equality constraints. We could also try other model relaxations, perhaps based on the modification indices.

Testing Main Effects and Interactions

Earlier in this chapter I noted that one of the advantages of the MG-MACS approach is that it allowed the testing of main effects and interactions in a single analysis. The testing of slope differences above tested for the presence of an interaction. By allowing the paths from pretest to posttest (the slopes) to vary across groups (and comparing that with a model that required equal slopes), we tested whether the pretest interacts with the treatment in its effect on the posttest. The results suggested that these two variables do interact, and that the latent pretest

had differential effects on the latent posttest, depending whether women were in the control or the experimental groups.

The comparison is less than ideal, however, because even though the Slopes Vary model fit statistically significantly better than did the Test Assumptions model, it still had a horrible fit. Let's go back, then, and ask the main effect and interaction question again but using a better fitting baseline model. Refer back to the output shown in Figure 19.20. Note that the covariances allowed between r2 and r4 were not statistically significant for either group. The "Initial 2" model shown in Table 19.2 removed the covariance between r2 and r4. This model is reasonable, given that this covariance is unnecessary, and removing it will provide two extra degrees of freedom for the model. As shown in the Table, the model fit the data well. The χ^2 increased slightly with this constraint, but the $\Delta\chi^2$ was not statistically significant. This model thus provides a good baseline for additional comparisons. Parameter estimates for the Control and Hypnosis groups are shown in Figure 19.24 on the left and right, respectively.

For the "No Main Effects" model, the latent posttest intercept for the Hypnosis group was constrained to zero (as was the intercept for the control group in all analyses). Thus this model specified no difference in means (intercepts) for those in the control versus hypnosis groups and no effect for the treatment on true hot flash severity and interference. As shown in the Table, this constraint resulted in a poor overall fit and a statistically significant increase in $\Delta\chi^2$ compared to the Initial 2 model. Given a statistically significant change in $\Delta\chi^2$, we would favor the less constrained, or Initial 2 model. Said differently, we should reject the hypothesis that the Hypnosis treatment had no effect on the severity, frequency, and interference of women's hot flashes. This finding is consistent with the large and statistically significant effect shown for the latent intercept for the Hypnosis group in the Initial 2 model in Figure 19.24.

The "No Slope Difference" model had the same specifications as the Initial 2 model, except that the path from Pretest to Posttest was constrained to be equal for the Control and the Hypnosis groups. As shown in Table 19.2, this constraint also resulted in a large and statistically significant $\Delta\chi^2$ in comparison with the Initial 2 model. We should reject the hypothesis that the pretest had the same effect on the posttest for both the control and the Hypnosis groups. As shown in Figure 19.24, the Control group path from the Pretest to the Posttest was positive, large (standardized path = .912), and statistically significant. For the Hypnosis

Regression Weights: (Control group - Initial model 2)

		Estimate	S.E.	C.R.	P
Hot Flash_Post	<---	Hot Flash_Pre	1.169	.274	4.27 ***
Int1	<---	Hot Flash_Pre	1.000		
HF1	<---	Hot Flash_Pre	.416	.098	4.244 ***
Int2	<---	Hot Flash_Post	1.000		
HF2	<---	Hot Flash_Post	.320	.036	9.009 ***

**Standardized Regression Weights:
(Control group - Initial model 2)**

		Estimate	
Hot Flash_Post	<---	Hot Flash_Pre	.912
Int1	<---	Hot Flash_Pre	.748
HF1	<---	Hot Flash_Pre	.620
Int2	<---	Hot Flash_Post	.933
HF2	<---	Hot Flash_Post	.593

Intercepts: (Control group - Initial model 2)

		Estimate	S.E.	C.R.	P
Int1		42.076	2.048	20.541	***
HF1		15.623	1.109	14.092	***
Int2		38.97	2.708	14.389	***
HF2		14.276	1.245	11.469	***

Regression Weights: (Hypnosis group - Initial model 2)

		Estimate	S.E.	C.R.	P
Hot Flash_Post	<---	Hot Flash_Pre	-.191	.199	-.962 .336
Int1	<---	Hot Flash_Pre	1.000		
HF1	<---	Hot Flash_Pre	.416	.098	4.244 ***
Int2	<---	Hot Flash_Post	1.000		
HF2	<---	Hot Flash_Post	.320	.036	9.009 ***

**Standardized Regression Weights:
(Hypnosis group - Initial model 2)**

		Estimate	
Hot Flash_Post	<---	Hot Flash_Pre	-.345
Int1	<---	Hot Flash_Pre	.641
HF1	<---	Hot Flash_Pre	.452
Int2	<---	Hot Flash_Post	.627
HF2	<---	Hot Flash_Post	.401

Intercepts: (Hypnosis group - Initial model 2)

		Estimate	S.E.	C.R.	P
Hot Flash_Post		-28.308	3.176	-8.912	***
Int1		42.076	2.048	20.541	***
HF1		15.623	1.109	14.092	***
Int2		38.97	2.708	14.389	***
HF2		14.276	1.245	11.469	***

Figure 19.24 Parameter estimates for the control and hypnosis groups for the Initial 2 MG-MACS hypnosis model.

(experimental) group, the effect was negative and not statistically significant. For women in the Hypnosis group, the pretest level of hot flashes had no effect on the Posttest level of those hot flashes. What is the effect of initial level of hot flashes on posttest level? Interaction lingo alert: It depends. It depends on whether women are in the control group or the hypnosis experimental group. In this research both the main effect and the interaction were statistically significant.

Other Technical Issues

Analyzing Matrices Versus Raw Data

We analyzed the model in this chapter using raw data to help convey the continuity between the dummy variable and the MG-MACS approaches. The same raw data file was used for both analyses (hot flash simulated.sav). As in other SEM analyses, it would also be possible to conduct these analyses using matrix data. Different matrices are required for the two approaches (dummy variable vs. MG-MACS), however, because the dummy variable method analyzes the data as a single group, and thus requires a single matrix. The MG-MACS approach, in contrast, analyzes two (or more) groups, and thus requires two (or more) matrices. Table 19.3 shows the matrix for the dummy variable approach, and Table 19.4 shows the two matrices

Table 19.3 Correlation Matrix and SDs for the Hot Flash Example Analyzed via the Dummy Variable Approach

Variable	Group	HF1	HF2	Int1	Int2
Group	1.000				
HF1	-.121	1.000			
HF2	-.520	.718	1.000		
Int1	-.182	.409	.342	1.000	
Int2	-.678	.248	.642	.426	1.000
SD	.503	11.111	10.130	20.173	23.275

N = 96

Table 19.4 Correlations, Means, and SDs for the Hot Flash Example Analyzed via the MG-MACS Approach

Variable	HF1	HF2	Int1	Int2
Control Group n = 48				
HF1	1.000			
HF2	.880	1.000		
Int1	.481	.436	1.000	
Int2	.516	.552	.632	1.000
Mean	17.077	15.508	46.313	42.250
SD	10.823	11.206	21.394	21.842
Hypnosis Group n = 48				
HF1	1.000			
HF2	.732	1.000		
Int1	.307	-.027	1.000	
Int2	-.290	.057	-.042	1.000
Mean	14.396	5.036	39.000	10.875
SD	11.347	5.086	18.373	10.736

for the MG-MACS approach (with the control matrix on top and the Hypnosis matrix on the bottom). Note that because we explicitly analyzed means and intercepts in the MG-MACS analysis these matrices also each include a row of means. In contrast, note that there are no means in Table 19.3, but that the Group variable appears in the correlation matrix. (Of course we could include a row of means in the matrix, but it is not needed because we did not explicitly analyze means and intercepts for this model.)

Calculating df

How do the 3 *df* for the MG-MACS Initial model come about? There are 4 measured variables for each of the 2 groups, and thus there are 14 Control + 14 Hypnosis = 28 moments (means, variances, covariances) to be analyzed. For the control group, 15 parameters are estimated in the model (2 factor loadings, 1 path, 4 measured intercepts, 6 variances, and 2 covariances). Ten parameters are estimated for the Hypnosis group: 2 factor loadings and 4 intercepts are constrained to be equal to the values for the Control group, but the intercept for the Hot Flash Posttest is freely estimated. Moments minus parameters estimated = $28 - (15 + 10) = 3$.

SUMMARY

Up until now in our exploration of SEM we have mostly been concerned with using covariances to estimate paths and correlations (covariances) among measured variables (initially) and latent variables (more recently). It is also possible to estimate mean structures (i.e., means and intercepts) in SEM, and that has been the focus of the current chapter. Early in the chapter we conducted a simple regression in path form as a reminder about means versus intercepts. Briefly, for exogenous variables we estimate the means, whereas for endogenous variables we estimate intercepts, which are the estimated means for those with a value of zero on the corresponding exogenous variable. This becomes slightly more complex with latent variable models, because measured indicators of latent variables are endogenous (influenced by the latent variables). If this seems confusing, just think of intercepts as estimated means adjusted for the variables that have paths pointing to the variable under consideration. We generally assume that latent variables have means of zero.

With a single-group analysis not much changes when the estimation of mean structures is added. The output of the analysis becomes a little more complex, but the results are the same as those when means and intercepts are excluded from the analysis. Many SEM programs require the analysis of means and intercepts when there are missing data, however. In the summary chapter for Part Two we will focus a little more on missing data; for now, simply understand that there are advantages for using missing data handling features (maximum likelihood estimation) in most SEM programs.

Although our concern up until this point has been with the estimations of paths, it turns out that several of our examples in previous chapters have indeed focused on mean structures. When dummy exogenous variables were included in the model (the Head Start exercise from Chapter 17 and the Homework example with ethnic group membership in the model in Chapter 18), it turns out that the paths we were estimating from these dummy exogenous variables were the intercepts on the latent outcome variables across the groups. Said differently, these paths estimated the differences across groups on the latent dependent variable, or the main effect for group membership on the latent (true) outcome variable.

The estimation of means and intercepts becomes even more interesting when we conduct multi-group analyses. Multi-group mean and covariance structures (MG-MACS) analysis allows us to test both main effects and interactions in one analysis. In previous chapters we estimated the main effect by including a dummy variable in one analysis and we tested for interactions (moderation) in a separate MG analysis across groups. Review Chapter 18 if you

are unclear how we did this. MG-MACS also allows us to test such models more completely than we did using dummy variable models.

An example was used to illustrate the similarities and differences between the two approaches—the dummy variable versus the MG-MACS approach—for estimating mean structures. Simulated data designed to be consistent with the findings of a true experiment using hypnosis to treat hot flashes in postmenopausal breast cancer survivors were analyzed. In the first analysis, a dummy variable was used to represent membership in the control group versus the experimental (hypnosis) group. The path from this dummy variable to a latent hot flash outcome variable (indexed by hot flash scores and hot flash interference in daily life) showed the effect of treatment on this latent outcome variable. Pretest hot flash scores and interference were also controlled. The use of a latent outcome variable had the advantage of coming closer to the true variable of interest (hot flash frequency, severity, and interference) than would an approach that relied only on measured outcome variables. This approach is similar to Multivariate Analysis of Covariance (a more common approach for such analysis), which would also treat the outcome as a latent variable. MANCOVA, however, would treat the pretest scores as two separate error-free covariates, and we saw in the chapter on error (Chapter 15) why it is dangerous to treat error-laden exogenous variables as if they were error-free.

The example was next analyzed as a multi-group model, with the explicit analysis of means and intercepts. Factor loadings and measured variable intercepts were constrained to be equal across groups (control versus experimental). The latent pretest means were constrained to zero for both groups; because there was random assignment to treatment groups the two groups should be equal on the latent (true, underlying) hot flash pretest. In one analysis the latent posttest intercepts (mean posttest score adjusted for pretest score) were constrained to zero for both groups, a model consistent with no treatment effect for group membership. In another analysis (actually, here it was the first analysis), the latent posttest intercept was allowed to vary for the hypnosis group, a model consistent with a treatment effect for the hypnosis group. In MG-MCAS, one group's latent means and intercepts must be set to zero, and the values for the other group (or groups) can be freely estimated; the difference represents the difference as a result of group membership. This model fit much better than did the no-treatment effect model, and the value for the intercept difference was large and statistically significant. Hypnosis led to a large and statistically significant reduction in hot flash frequency, severity, and interference.

The value for the difference in intercepts in the MG-MACS model was similar, but not identical to, the value of the path from treatment group to Hot Flash outcome in the dummy variable model. The difference in these two coefficients was a result of assumptions that were made but were untested in the dummy variable model. When these assumptions were made explicit in the MG-MACS model (by constraining variances, covariances, and slopes to be equal across groups), the estimate of treatment effect was identical in the two models. These constraints also led to a much worse-fitting model, however, a finding that illustrated that the assumptions made for the dummy variable model were probably not valid. This, then, is also an advantage of the MG-MACS approach: it allows the testing of assumptions that are made but not tested in the dummy variable approach to estimating mean structures.

For these examples we analyzed raw data. It is also possible to conduct both types of analyses using matrix input. The matrices look different for the different analyses, however. For the dummy variable approach a single matrix is used, and one of the variables in the matrix represents the dummy (group membership) variable. It is not necessary to explicitly analyze means structures. With the MG-MACS approach, separate matrices are needed for each group, and the grouping variable (control versus experimental group) does not appear in the matrix. You must explicitly estimate means and intercepts in the MG-MACS approach, however, because group differences show up as differences in means and intercepts of the latent variables. The matrices input for the MG-MACS approach must also include a row of means for the measured variables.

EXERCISES

1. Reproduce the hot flash analyses used in this chapter, both the dummy variable model and the MG-MACS models. Make sure your results match mine. Are there additional models you might test?
 2. Figure 19.25 shows a starting model for a MG-MACS analysis of the effect of Homework on 12th-grade GPA. The model for boys is shown. A starting model for Amos (with variable names but without cross-group constraints) is available on the website (www.tzkeith.com). Also on the website are the raw data for analysis (homework means.sav).

Further develop this model so that you can conduct a MG-MACS analysis (add a group for girls, make the correct cross-group constraints). In the initial model constrain latent means/intercepts for Family Background, Previous Achievement, Homework, and Grades to zero for both groups. In a second model allow the Homework and Grades intercepts to vary for girls. Do girls have significantly higher or lower levels of true homework (once other variables are controlled)? Grades? In a third model allow the effect of Homework on Grades to vary across groups. Does homework have the same effect on grades for boys and girls, or does the effect of homework depend on sex?

Interpret your findings. Make sure you answer the questions asked in the preceding paragraph.

3. Figure 19.26 shows a dummy variable model designed to test the effect of Sex on the change in Locus of Control from 8th to 10th grade (or 10th-grade Locus controlling for

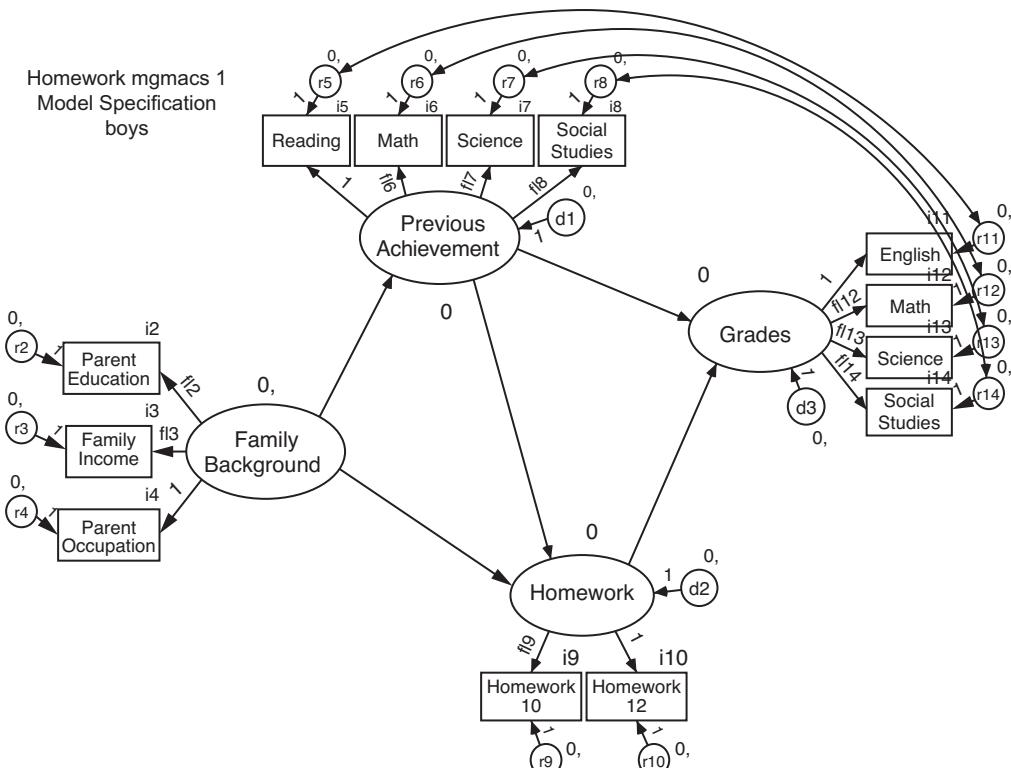


Figure 19.25 Initial MG-MACS model to study levels of homework and grades and effects of homework on high school grades for boys versus girls.

8th-grade Locus). Analyze the model using the NELS data. You should recode the Sex variable or create a new Female dummy variable so that boys are coded 0 and girls 1. Analyze the model. Do boys or girls have higher (more internal) locus of control in 10th grade?

Analyze a MG-MACS version of this model. Test both for intercept differences and for differences in the effect of Achievement on Locus of Control in 10th grade.

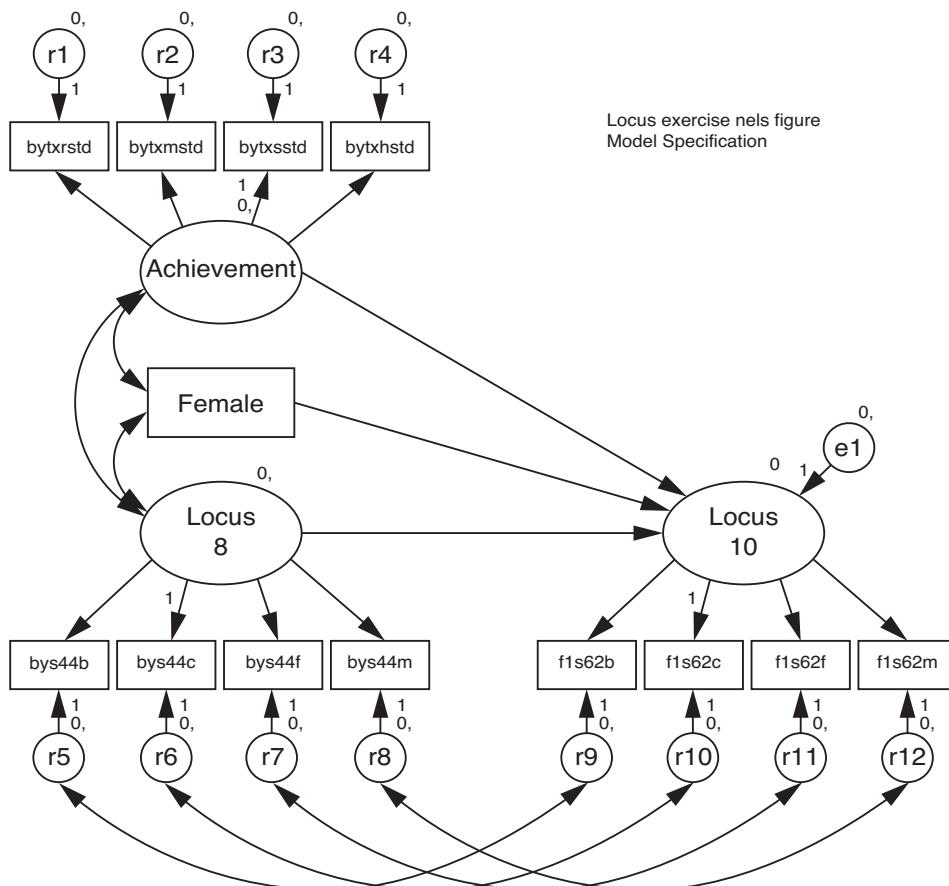


Figure 19.26 Initial dummy variable model to study differences in locus of control for boys versus girls in grades 8 and 10.

Notes

- 1 In Amos, this is accomplished by clicking on “Estimate means and intercepts” in the estimation tab in the “View→Analysis Properties” menu. The graphic input and output shown here are in Amos format. In order to obtain identical estimates with Amos to those shown, you will need to change one setting: View→Analysis Properties→Bias. Choose “unbiased” for both “covariances supplied as input” and for “covariances to be analyzed.” With small samples, the default settings in Amos will give divergent results from regression. The reasoning for the difference is explained in the Amos manual (see example 16, pp. 242–243 in the version 22 manual, or search for “unbiased”).
- 2 For this third model I constrained the path from Hypnosis to Posttest to zero, while still allowing the covariance between Hypnosis and Pretest (as in the second model). Because we rejected the “Pretests Vary” model, it would make more sense to constrain the covariance to zero for the third model and compare it to the initial model. That model would not run, however (problems with identification).

20

Confirmatory Factor Analysis II

Invariance and Latent Means

Invariance Testing With Means	475
<i>Measurement Invariance Steps</i>	478
<i>Alternative Model Specification</i>	494
<i>Invariance Testing Without Means</i>	499
<i>Higher-Order Models</i>	500
Single-Group, MIMIC Models	505
Summary	509
Exercises	510
Notes	512

Now that we have introduced the topic of latent means in SEM, we can revisit the topic of CFA, with the addition of latent means analysis in CFA. We will do so within the framework of invariance testing, a topic first introduced in the initial discussion of multi-group SEM. This is an important topic that needs additional exploration. Although it is possible, perhaps even common, to test for aspects of invariance without examining measured and latent means, here we will first focus on invariance testing with means and intercepts. Thus, this chapter will focus in some detail on the steps needed to test for invariance in constructs across groups, including invariance in measured variable intercepts, which will allow the testing of differences in latent means. Part of this discussion will concern what is tested conceptually at each step, and why one would want to do such testing. We will then back up a little and focus in less detail on the steps you might take if you were interested in invariance testing without focusing on means. Finally, we will (as in the previous chapter) see how some of this same information can be obtained by the addition of a categorical variable to the analysis, but with the addition of some assumptions that may or may not be valid.

INvariance TESTING WITH MEANS

In the introductory chapter on CFA we focused on an example from the intelligence literature in part because the topics of factor analysis and intelligence are so intertwined. We will do so again. The example is drawn from research by Matthew Reynolds and colleagues in which they were interested in possible sex differences and similarities in general and specific intelligences (Reynolds, Keith, Ridley, & Patel, 2008). Previous research had shown some consistent differences across the sexes (e.g., males generally perform at a higher level on measures of spatial rotation), but also areas with no differences, and plenty of inconsistencies across

studies. Reynolds and colleagues analyzed data from children ages 6 through 18 from the standardization of the Kaufman Assessment Battery for Children—Second Edition (KABC-II) (Kaufman & Kaufman, 2004). They used a higher-order model of intelligence in order to study both general intelligence and five more specific intellectual abilities. They reasoned that one reason for inconsistencies in research findings was that researchers often had studied measured variables, such as composite scores, that were likely clouded by the specific measures used. Latent variables should provide more accurate estimates of any true differences.

Here, we will use data from one age group (ages 15–16), and with a slightly different focus. Specifically, we are interested in testing whether the KABC-II measures the same set of constructs for boys and girls in this age group. For the sake of presentation, we will focus on fewer constructs, and only on first-order factors. The data (correlation matrices, means, and standard deviations) are in the first two worksheets in the Excel file “kabc cfa matrices.xls.” The third worksheet will be used later in the chapter. The small amount of data that were missing were imputed.

Table 20.1 shows a brief description of the various KABC-II subtests used in this chapter, and Figure 20.1 shows the constructs these tests supposedly measure, that is, the

*Table 20.1 Description of KABC-II Subtests for Youth Ages 15 to 16. Adapted from “Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACs and MIMIC models” by M. R. Reynolds, T. Z. Keith, K. P. Ridley, & P. G. Patel, *Intelligence*, 36, 236–260. Copyright 2008 by Elsevier.*

<i>Subtest</i>	<i>Description</i>
Riddles	Examinee points to or names objects or ideas described by examiner
Verbal Knowledge	Points to a picture that illustrates the meaning of a vocabulary word or the answer to a general information question
Expressive Vocabulary	Names pictured objects
Gestalt Closure	Describes the pictured object or action from incomplete black and white drawings
Triangles (Untimed)	Arranges two-colored foam triangles to match a pictorial model
Block Counting	Counts blocks in pictures when some blocks are clearly visible and others are implied or only partially visible
Rover	Determines the most efficient route for a dog to find a bone on a grid. The route must take into account various obstacles.
Rebus	Examiner teaches the meaning of rebuses (pictures representing words); the examinee reads a series of rebuses, which form a sentence or phrase
Rebus Delayed	Reads a series of rebuses 15–25 minutes after initial training
Atlantis	Examiner teaches names for cartoon fish and objects; the examinee points to the correct picture when the examiner subsequently names them
Atlantis Delayed	Points to the Atlantis objects 15–25 minutes after initial training
Word Order	Examiner states object names, examinee touches pictures of the objects in the same order. Later items have an intervening interference task.
Number Recall	Recalls digits spoken by examiner
Hand Movements	Repeats a series of hand motions made by examiner

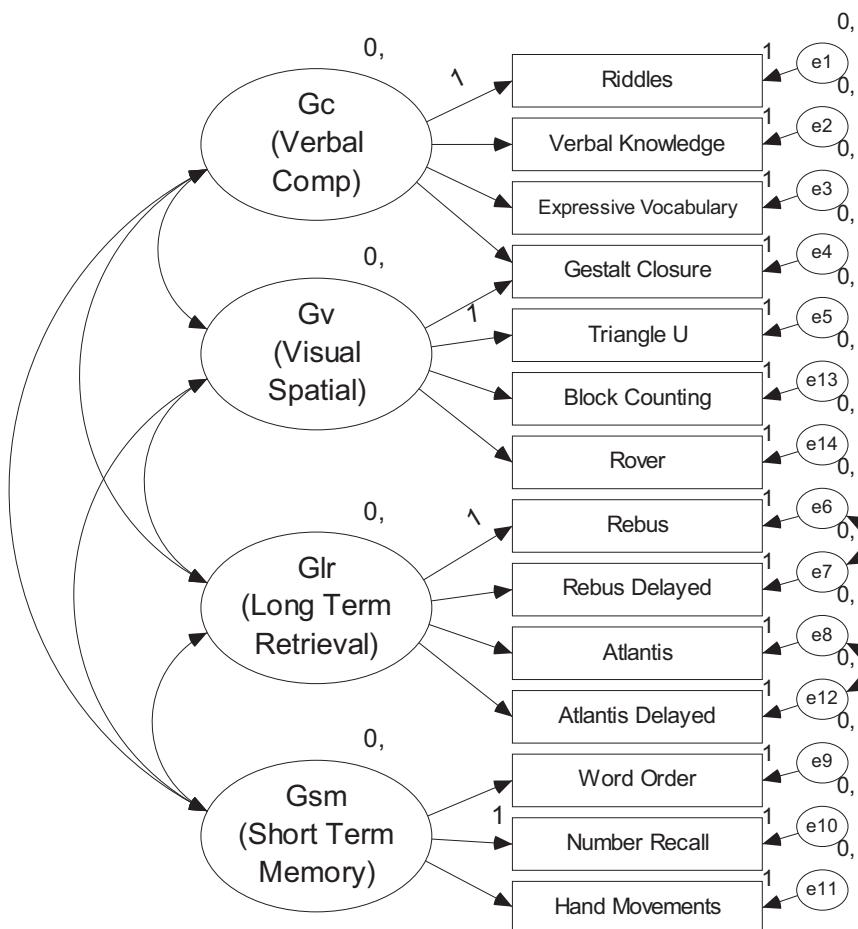


Figure 20.1 Factor structure of the KABC-II for 15- to 16-year-olds.

expected factor structure. The construct Gc is also known as crystallized intelligence, or may be referred to as verbal comprehension or verbal reasoning. Gv represents visual-spatial reasoning, Glr long-term storage and retrieval, and Gsm short-term memory. As shown, most subtests appear to measure a single construct, with the exception of Gestalt Closure, which is thought to require both visual-spatial and verbal skills. This makes sense, given that for Gestalt Closure the child is supposed to describe or name incomplete pictures. The model allows the residuals of two pairs of tests to covary, because one in each pair is a delayed version of the other. So, for example, for the Rebus Delayed subtest, children are asked to recall names associated with symbols initially presented in the Rebus subtest. Note that the KABC-II is also designed to measure other abilities not analyzed here.

Again, our interest in this chapter is whether the constructs measured by the KABC-II are measured in the same way, or are measurement invariant, across the sexes. We are interested in the invariance of measurement across groups. As we shall see, there are different levels of measurement invariance that can be modeled with CFA, from fairly loose definitions of invariance to quite strict. We will focus on the steps needed to test these levels of invariance and what each level means.

Measurement Invariance Steps

Configural Invariance

The first step in invariance testing is often referred to as configural invariance. The model shown in Figure 20.1 is estimated via a multi-group model but without parameter constraints across groups (other than the reference variable indicator of 1 for each factor, and the same pattern of fixed at zero versus free loadings). In other words, for this level of invariance we simply specify that the same factor model holds for both groups. There are no specifications that the values of factor loadings must be the same across groups, just that the same pattern of loadings holds. As we shall see, the χ^2 of this model will be the same as if we had analyzed each group separately and summed the χ^2 values. This level of invariance is generally called configural invariance, meaning the structure of what is measured by the test shows the same configuration across groups (boys and girls in this example).¹

Note, then, in Figure 20.2 that the same factor configuration is specified for boys and for girls, with the model for boys to the left and the model for girls to the right. The model is the same as in Figure 20.1, with the exception that the shorter titles are used for the factors. For both sexes, one factor loading is set to 1 to set the scale for the latent ability variables. Although it is not necessary to estimate mean structures for this level of invariance (and we will examine such models later in the chapter), we have done so here. The means of all latent variables (the ability constructs and the residuals/error) are fixed to zero. Although it may not be obvious from the figures, the intercepts for the measured variables (the subtests) are freely estimated for both groups.

The graphic standardized results for this model are shown in Figure 20.3. Because this is a multi-group model, there is only one set of fit indices shown, representing the fit of the model across both groups. Note that a single set of fit indices is provided in Amos; other programs may also provide some of the fit indices separately by group (e.g., Mplus provides the contribution of each group to the χ^2). The model shows a good fit for girls and boys,

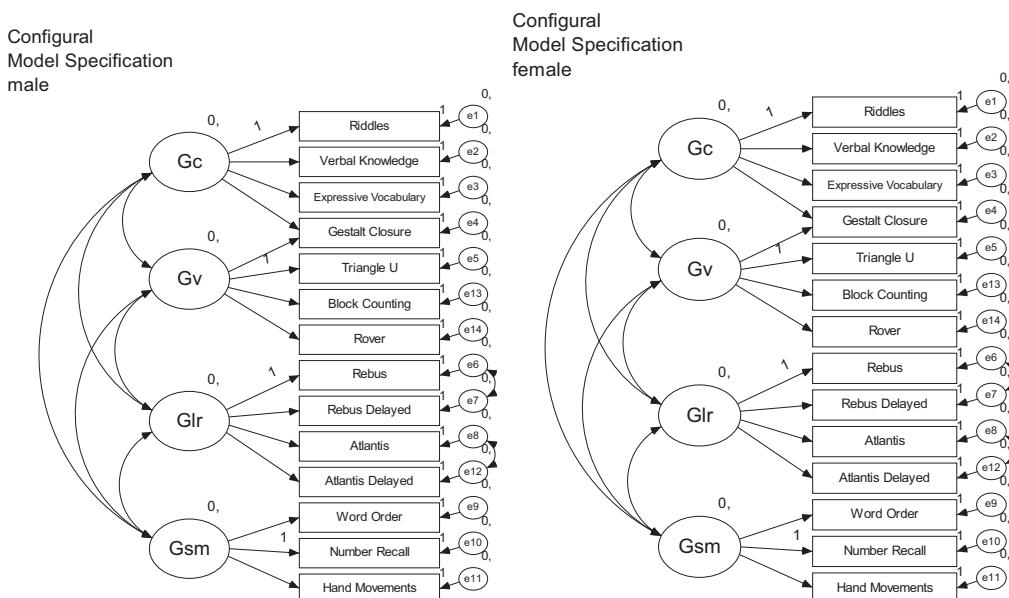


Figure 20.2 Configural invariance model. The same factor structure is specified for males and females, but no cross-group constraints are made.

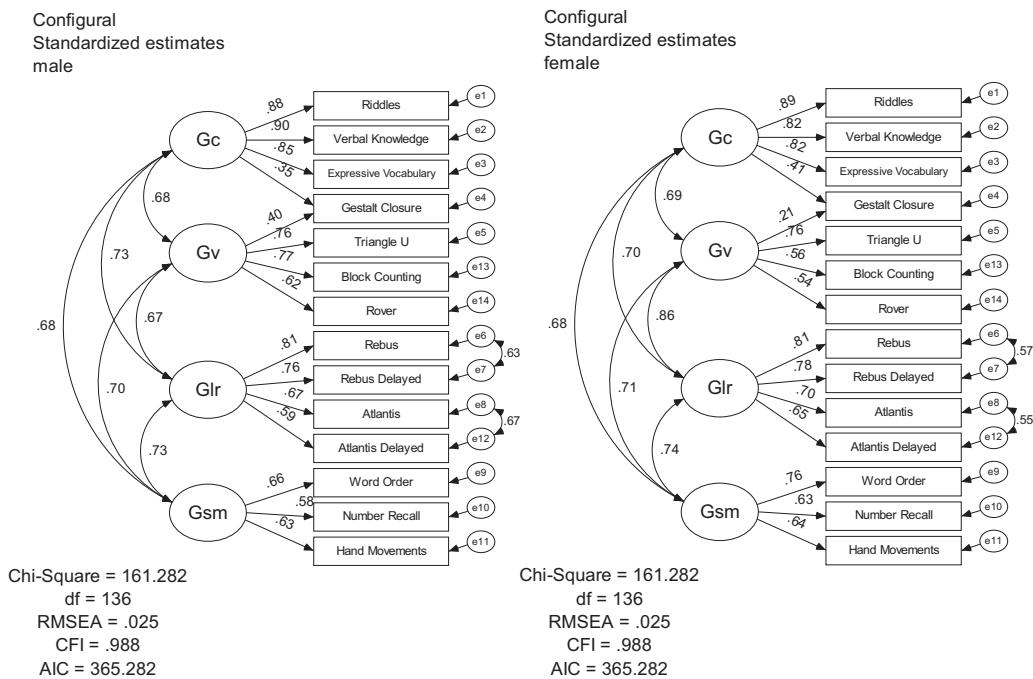


Figure 20.3 Configural invariance results, standardized estimates. The results for boys are shown to the left, and girls to the right.

with RMSEA = .035 (corrected for 2 groups), SRMR = .047, and CFI = .988. Fit indices for this and subsequent models are also shown in Table 20.2. We would likely accept this model as providing a good baseline for subsequent model comparisons.

As shown in the Figure, the model produced similar standardized estimates for both girls and boys. All values are reasonable, with most factor loadings and factor correlations of substantial and reasonable magnitude. Perhaps the biggest difference is the factor loading of the Gestalt Closure subtest on the Visual-Spatial (Gv) factor: .40 for males and .21 for females. We should not over-interpret this difference, however. First, recall that to compare differences across groups, we should compare unstandardized estimates rather than standardized ones. Second, we will test whether the unstandardized loadings are statistically significantly different across the sexes in the next step in invariance testing.

As noted previously, the χ^2 for the configural invariance model should be the sum of the χ^2 's for the boy and girl models if run separately. The fit information for these models is also shown in Table 20.2. It appears that the model fit well for boys and girls, and the χ^2 for the configural invariance model is almost identical to the summed value for males and females separately (161.282 vs. 161.281). The same relation holds for the AIC, whereas the configural CFI and SRMR are closer to averaged values. The RMSEA appears somewhat better for the Configural Invariance model than for the separate male and female models, but recall that RMSEA should be corrected for the number of groups (Steiger, 1998). The column labeled RMSEA* shows the corrected RMSEA, the reported RMSEA multiplied by the square root of the number of groups ($RMSEA \times \sqrt{2}$). This correction is necessary in Amos (but not in Mplus) as of this writing; check whatever program you are using to determine whether it is needed.

It is common—but certainly not universal—to test the factor structure of each group separately either before or after testing the configural invariance model, and some authors recommend this be done routinely (e.g., Brown, 2015). This approach has much to recommend

Table 20.2 Tests of Invariance of Factor Structure for Males and Females, Ages 15–16 on the KABC-II

Model		χ^2	df	$\Delta\chi^2$	Δdf	p	RMSEA	RMSEA*	SRMR	CFI	AIC
1.	Configural	161.282	136			.025	.035	.047	.988	365.282	
1a.	Male	83.047	68			.039	.039	.047	.987	185.047	
1b.	Female	78.234	68			.031	.031	.043	.990	180.234	
2.	Metric	172.236	147	10.954	11	.447	.024	.034	.988	354.236	
3.	Intercept (means vary)	181.305	157	9.069	10	.526	.023	.033	.051	.989	343.305
4.	Subtest residuals	204.322	173	23.017	16	.113	.025	.035	.050	.986	334.322
5.	Factor variances	208.217	177	3.895	4	.420	.024	.034	.056	.986	330.217
6.	Factor covariances	214.034	183	5.817	6	.444	.024	.034	.055	.986	324.034
7.	Factor means	243.743	187	29.709	4	<.001	.032	.045	.057	.974	345.743
7a.	Gc means equal	214.347	184	.313	1	.576	.024	.034	.055	.986	322.347
7b.	Gv means equal	220.222	184	6.188	1	.013	.026	.037	.055	.983	328.222
7c.	Glr means equal	218.091	184	4.057	1	.044	.025	.035	.055	.984	326.091
7d.	Gsm means equal	217.654	184	3.620	1	.057	.025	.035	.056	.985	325.658
7e.	Gc and Gsm means equal	221.081	185	7.047	2	.029	.026	.037	.056	.984	327.081

*RMSEA corrected for the number of groups

Note: Models 2 through 7 are compared to the previous model. Models 7a through 7e are compared to Model 6.

it, given that it tests whether the chosen model fits each group well and thus allows model adjustments prior to testing for configural invariance. I thought that this step was not necessary in this case, because the factor structure is fairly well understood for this test. In addition, the Configural Invariance model fit well, so I would likely just move to the next step in invariance testing. On the other hand, if this were an analysis of a relatively new or unexplored measure, or if we had questions resulting from this first step, a separate analysis by group would help to understand the nature of the group differences. Perhaps the model fits well for one group but not the other. Perhaps one subtest should have a cross-loading for one group but not the other, or two subtests should have correlated errors for one group. A change or two may be quite reasonable, and, if minor, may still allow a conclusion of configural invariance, or partial configural invariance. Any change that leads to a *substantive* difference in interpretation across groups suggests a lack of invariance. Of course, what constitutes a “substantive” difference will likely be a matter of opinion; the excellent references concerning invariance testing throughout this chapter will provide guidance on this and other topics. Please note, however, that *if you need to make changes in the factor structure (number of factors, cross-loadings, error covariances), this is the time to do it*; subsequent models simply add equality constraints across groups.

Metric Invariance

The next step in invariance testing is often referred to as Metric Invariance, or factor loading invariance. It is also known as weak factorial invariance (as opposed to the next step, strong factorial invariance) (Meredith, 1993; Meredith & Teresi, 2006).² For this step, the loadings of the subtests on the factors are constrained to be equal for males and females. The setup for this step is shown in Figure 20.4. Note that the loading for Verbal Knowledge on Gc is set to gcl1 (for Gc loading 1) for both males and females. All other loadings are also constrained to be equal across groups, except those used to identify the latent factors, which were already set to 1 in both groups. Note that all other aspects of the factor model, including unique

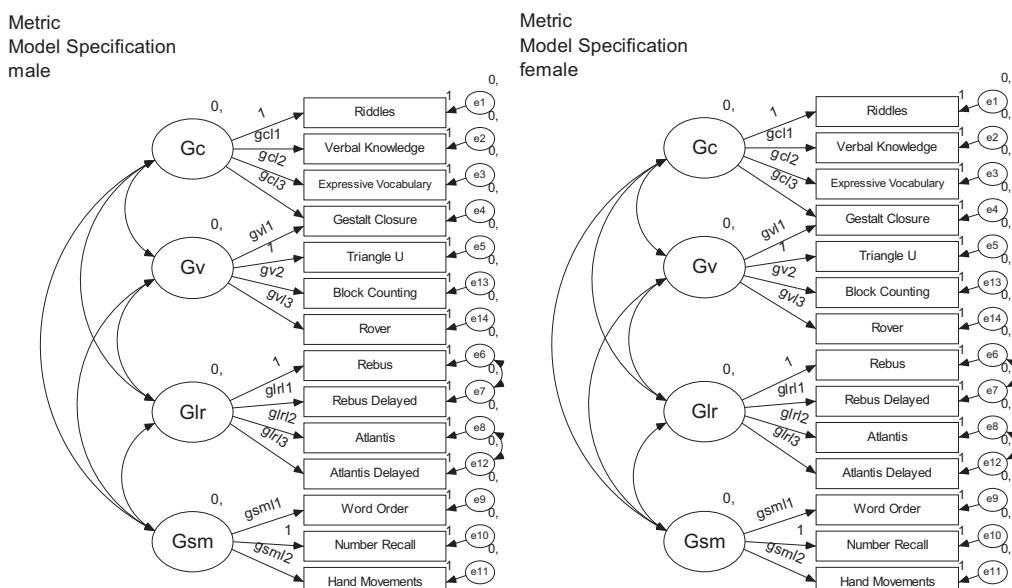


Figure 20.4 Model setup for a test of metric invariance across the sexes. Factor loadings are constrained to be equal across the two groups.

variances of subtests (e1 through e11), factor variances, and factor covariances, were allowed to vary across groups. It is not necessary to estimate means and intercepts at this step (to be discussed in more depth later), but if means and intercepts are estimated, the intercepts are allowed to vary across groups, but factor means are constrained to zero for both groups (as was done here).

Table 20.2 shows the fit indices for this model. As shown in the Table, the model fit the data well ($\text{RMSEA} = .034$, $\text{SRMR} = .051$, $\text{CFI} = .988$). This model is nested with the Configural model, and the increase in χ^2 for the Metric invariance model was not statistically significant ($\Delta\chi^2(11) = 10.954$, $p = .447$). Thus, we would likely accept the factor loading equality constraints in the Metric invariance model as reasonable. As will be explained later in this chapter, it is also common to use other fit indices (e.g., ΔCFI) to compare invariance models.

Given metric invariance, what does it mean? With metric invariance, the unstandardized factor loadings are the same for both groups. The factor loadings tell us about the relation of the measured variables to the latent factors. This level of invariance means that the scales of the latent variables are the same for both males and females. This finding, in turn, means that for each unit change in the latent variable, it is reasonable to assume that scores on the subtests increase by the same amount for males and females. So, for example, if the true level of long-term retrieval (Glr) increases by 10 points, scores on the Rebus subtest will increase by 10 points for males and 10 points for females (because the unstandardized loading for both groups is 1.0). Likewise, with this example, scores on the Atlantis subtest would increase by 9 points for both males and females (the unstandardized loading for both groups = $.916 \times$ a 10-point increase in Glr = 9.16).

Conversely, imagine what it would mean if metric invariance did not hold. If metric invariance did not hold for the Glr factor, that would mean that a 10-point increase in the latent variable would result in a different point increase on the subtest for adolescent boys versus girls. As an analogy, imagine a fishing competition where the biggest fish wins the prize. Imagine that I measure my fish using a meter stick and you measure yours using a yard stick. I find that my fish is 35 units long compared to your fish, which is only 25 units long. I win, correct? No; if converted to the same units, *the same metric*, my fish is 35 cm long, whereas yours is 64 cm. The scaling needs to be the same for the two instruments (tests, scales, rulers) to measure the same construct. Metric invariance means using the same scale for both groups, which in turn means the factors represent the same “thing” across groups. As another example, imagine measuring the temperature in two different cities daily for a month but you use a Fahrenheit thermometer in one city and a Celsius scale in the other. The comparisons would make no sense because the scales of measurement are different.

As noted when we first discussed multi-group analysis, with this level of invariance we can conclude that the latent variables have the same meaning and represent the same constructs across groups. In SEM, this level of invariance is the minimum level needed in order to compare the effects (the paths) of one latent variable on another. Thus, if this were an SEM rather than a CFA model, we could now validly compare the effects of one latent variable on another across groups (e.g., Gc on Glr). In CFA (and SEM), if metric invariance is achieved, it is reasonable to compare factor variances and covariances across groups (Brown, 2015).

If metric invariance is not achieved, it is possible to test for partial metric invariance (Byrne, Shavelson, & Muthén, 1989) by allowing one (or several) factor loadings to differ across groups, while constraining all other loadings to be equal. We would likely pursue this option if the metric invariance step resulted in considerable decrement in model fit. The modification indices could be used to help isolate loadings that should be allowed to vary across groups. Any such subtest should be freed in the intercept invariance step as well. Given the complexity of such models and the number of comparisons being made, we might also choose to use a more conservative level of statistical significance (e.g., $p < .01$ or $.001$) for rejecting the hypothesis of invariance at each step. For this reason, some researchers have

suggested using ΔCFI or Δ in some other fit index as opposed to $\Delta\chi^2$ to evaluate invariance tests (Cheung & Rensvold, 2002). With this approach we might decide that a change in CFI of $-.01$ or more from one step to another signals a lack of measurement invariance at that step. In my experience, this ΔCFI criterion works well for invariance testing.

In the section (Chapter 7) on categorical and continuous variables in multiple regression we discussed the issue of test bias. Invariance testing is commonly used to answer an even more basic question about bias, sometimes referred to as bias in construct validity. Consider that if metric invariance did not hold in the present example we would be forced to conclude that the KABC-II measures different constructs (in some sense of the word) for males versus females. This test of bias answers a common question about tests and other scales across groups, commonly expressed along the lines of “Sure, the XYZ test likely measures intelligence for white middle-class students, but it probably measures something different, perhaps test-taking skill, for students from ethnic minority backgrounds.” Or, with the current example, “Sure, the Gestalt Closure, Triangles, Block Counting, and Rover subtests likely measure visual spatial reasoning for boys, but they probably measure exposure to such mechanical-spatial problems for girls.” In both cases, the questioner is suggesting that the constructs measured by the instrument differ across groups. Achieving metric invariance across groups suggests no such bias across groups (as we will see, however, other problems may still exist).

Intercept Invariance

Invariance testing with mean structures also goes by several names: scalar invariance, intercept invariance, or strong factorial invariance. I will use the term intercept invariance because this term makes clear what additional constraints are being made; I believe the term *scalar invariance* is more common. Intercept/scalar invariance includes all the constraints of metric invariance, plus the added constraint that the *intercepts* of the corresponding measured variables are constrained to be equal across groups. The setup for this step is shown in Figure 20.5 for females. Notice the values i_1, i_2, i_3 , and so on next to the subtests. These labels refer to the intercepts of each subtest, and the same labels are used for the male model, thus constraining the values of the measured intercepts to be equal for females as for males.

Note also that constraints on the factor means (G_c, G_v , etc.) have been removed for females. In the previous model, the metric invariance model, the latent means were set to zero for both males and females. For this step, the male latent factor means are still constrained to zero, but the female factor means are allowed to differ from the male factor means. This combination of free versus constrained parameters means that any differences in intercepts (and therefore means) on the subtests are the result of *true* differences in means of the latent variables (G_c, G_v , etc.) rather than something specific to that subtest. Figure 20.6 shows a portion of this same model using the RAM format where the triangle pointing to a factor or variable represents the estimate of its mean or intercept.

The fit of the intercept invariance model is also shown in Table 20.2. The unstandardized model for females is shown in Figure 20.7. As shown in the Table and the Figure, the intercept invariance model fit well. More important for our present purpose, the fit compared well to the metric invariance model, with $\Delta\chi^2$ not statistically significant and the ΔCFI is less than $-.01$ ($.988 - .989 = -.001$). Note also that the AIC is lower for the intercept invariance as opposed to the metric invariance model, also supporting the constraints imposed by this model.

More detailed model output for males and females is shown in Figures 20.8 through 20.10, with the male output on the left and female output on the right. There is a lot of detail contained in these figures so we will spend some time going over it. I encourage you to compare the output for the two groups on your own, and to conduct these various analyses to make sure you understand how to do them and how to interpret them. Figure 20.8 shows the unstandardized and standardized factor loadings (paths from the latent to measured variables) for the two

Intercepts
Model Specification
female

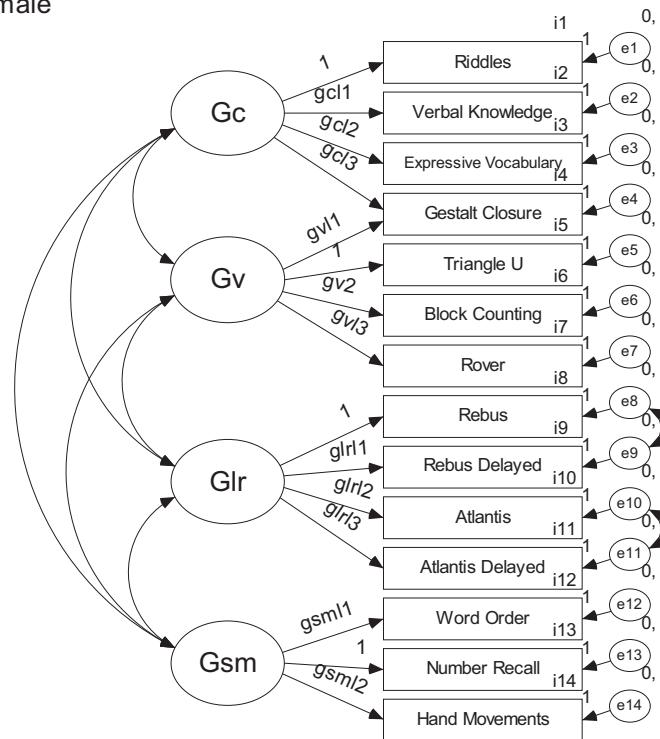


Figure 20.5 Specifying invariance for the measured variable intercepts. The same names are given to the intercepts for boys and girls, thus constraining these to be equal across the groups (intercept invariance). At the same time, the latent factor means are freely estimated for one group.

Portion of ram model
Model Specification
female

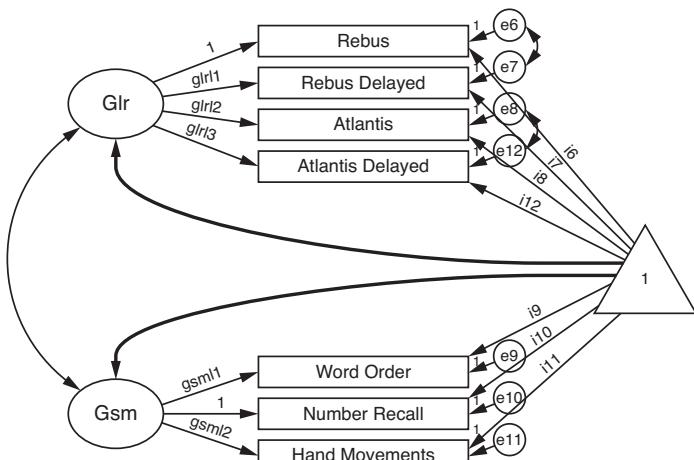


Figure 20.6 A portion of the intercept (scalar, strong) factorial invariance model using a RAM-type notation. Because all latent variables means were set to zero for males, the male model would have no arrows from the constant (triangle) to the factors. In addition, the paths from the constant to the subtests (representing the intercepts) are constrained to be equal for males and females.

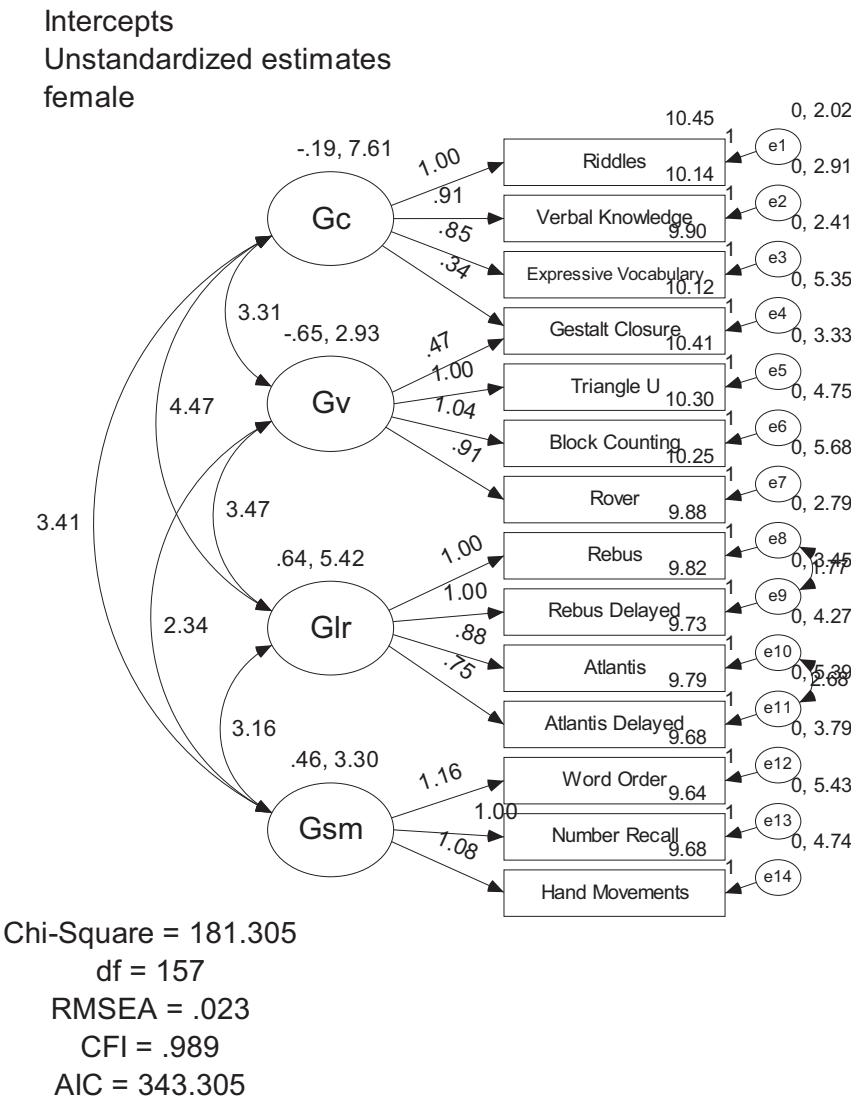


Figure 20.7 Intercept (scalar, strong) invariance results for females. Note that the values above the factors represent the differences in means from the group coded zero (males), followed by the factor variance.

groups. The tables of unstandardized loadings shows that these values indeed are constrained to be equal across groups, with those constraints originally made in the metric invariance model. The columns named “Label” show the labels attached to each parameter in the model setup. This is, of course, how equality constraints are made in Amos; other programs will use other methods.³ Note that males and females have the same labels for all factor loadings not constrained to 1. (It is also possible to give the parameters different names and then tell the program to constrain them to be equal; see the Amos manual for more detail.) The second set of tables show the standardized loadings. If you compare these across groups you will see that they are similar but not identical for boys and girls. Why? Again, it is the unstandardized values (paths, loadings, etc.) that are constrained to be equal across groups, not the standardized values. Recall from multiple regression that a standardized coefficient depends on the unstandardized coefficient AND on the variances of the variables involved. Just as in regression, $\beta = b \frac{SD_x}{SD_y}$ or $= b \sqrt{\frac{V_x}{V_y}}$, none of the variances is constrained to be equal in this model, so the standardized loadings are not equal across groups.

Estimates (male - intercepts equal)

Regression Weights: (male - intercepts equal)						
		Estimate	S.E.	C.R.	P	Label
RIDDLES	<--- Gc	1.000				
VERB_KNO	<--- Gc	.908	.046	19.861	***	gcl1
EXP_VOC	<--- Gc	.847	.045	18.664	***	gcl2
TRIAN_UN	<--- Gv	1.000				
REBUS	<--- Glr	1.000				
REBUS_D	<--- Gir	.999	.047	21.261	***	gir1
ATLANTIS	<--- Gir	.880	.084	10.486	***	gir2
WORD_ORD	<--- Gsm	1.161	.134	8.664	***	gsm1
NUM_REC	<--- Gsm	1.000				
ATLANT_D	<--- Glr	.749	.079	9.535	***	gir3
BLOCK_C	<--- Gv	1.038	.098	10.600	***	gv2
HAND_MOV	<--- Gsm	1.075	.132	8.162	***	gsm2
ROVER	<--- Gv	.909	.099	9.161	***	gv3
GESTALT	<--- Gc	.395	.073	4.620	***	gcl3
GESTALT	<--- Gv	.467	.115	4.062	***	gv1

Estimates (female - intercepts equal)

Regression Weights: (female - intercepts equal)					
		Estimate	S.E.	C.R.	P
RIDDLES	<--- Gc	1.000			
VERB_KNO	<--- Gc	.908	.046	19.861	*** gcl1
EXP_VOC	<--- Gc	.847	.045	18.664	*** gcl2
TRIAN_UN	<--- Gv	1.000			
REBUS	<--- Glr	1.000			
REBUS_D	<--- Glr	.999	.047	21.261	*** glr1
ATLANTIS	<--- Glr	.880	.084	10.486	*** glr2
WORD_ORD	<--- Gsm	1.161	.134	8.664	*** gsm1
NUM_REC	<--- Gsm	1.000			
ATLANT_D	<--- Glr	.749	.079	9.535	*** glr3
BLOCK_C	<--- Gv	1.038	.098	10.600	*** gv2
HAND_MOV	<--- Gsm	1.075	.132	8.162	*** gsm2
ROVER	<--- Gv	.909	.099	9.161	*** gv3
GESTALT	<--- Gc	.335	.073	4.620	*** gw3
GESTALT	<--- Gv	.467	.115	4.062	*** gw1

Standardized Regression Weights: (male - intercepts equal)

		Estimate
RIDDLES	<---	Gc .884
VERB_KNO	<---	Gc .903
EXP_VOC	<---	Gc .843
TRIAN_UN	<-->	Gv .789
REBUSES	<---	Glr .820
REBUS_D	<---	Glr .770
ATLANTIS	<---	Glr .611
WORD_ORD		Gsm .685
NUM_REC	<---	Gsm .588
ATLANT_D	<---	Glr .601
BLOCK_C	<-->	Gv .730
HAND_MOV		Gsm .604
ROVER	<-->	Gv .618
GESTALT	<---	Gc .382
GESTALT	<-->	Gv .373

Standardized Regression Weights: (female - intercepts equal)

		Estimate
RIDDLES	<-- Gc	.889
VERB_KNO	<-- Gc	.827
EXP_VOC	<-- Gc	.833
TRIAN_UN	<-- Gv	.684
REBUS	<-- Gir	.813
REBUS_D	<-- Gir	.782
ATLANTIS	<-- Gir	.704
WORD_ORD	<-- Gsm	.734
NUM_REC	<-- Gsm	.615
ATLANT_D	<-- Gir	.601
BLOCK_C	<-- Gv	.632
HAND_MOV	<-- Gsm	.668
ROVER	<-- Gv	.547
GESTALT	<-- Gc	.329
GESTALT	<-- Gv	.265

Figure 20.8 Detailed results, intercept invariance tests.

Figure 20.9 shows the values for the measured intercepts across groups. Again, for this model the intercepts for each measured variable have been constrained to be equal for males and females. Because the intercepts are constrained to be equal across groups, it is possible to allow the means of the latent variables to differ across groups. Without the intercept constraints, we could not allow the means of the latent variables to differ across groups, because the model would be under-identified (we would be using the 14 subtest means to estimate both 14 intercepts and 4 factor means). As in the previous chapter, what we are saying is that any differences that are shown on the means of the various subtests are a result of true mean differences in the latent variables. The latent mean differences are shown for females in the lower part of the Figure. Recall that the latent means are set to zero for males and the values shown for females thus represent the differences from zero for females. Thus adolescent girls differ from boys by $-.194$ points on the latent Gc factor, meaning they score lower by 2/10 of a point, compared to boys. This value is not statistically significant, however, meaning that we should probably consider the true value to be zero, or not different from that of boys. Two of the latent mean differences were statistically significant, however: those for Gv and Glr. These findings suggest that boys score statistically significantly higher on the latent visual-spatial reasoning factor ($-.648$) and that girls score statistically significantly higher on the latent long-term retrieval factor ($.638$). We will return to these findings and delve more deeply into them later.

Figure 20.10 shows the information concerning covariances, correlations, and variances across groups. None of the values has been constrained across groups, but we can add such constraints in subsequent models.

Intercepts: (male - intercepts equal)				
		Estimate	S.E.	C.R.
RIDDLES		10.449	.270	38.696 ***
VERB_KNO		10.143	.244	41.656 ***
EXP_VOC		9.898	.235	42.187 ***
GESTALT		10.116	.196	51.520 ***
TRIAN_UN		10.414	.210	49.680 ***
REBUS		9.877	.223	44.251 ***
REBUS_D		9.822	.231	42.556 ***
ATLANTIS		9.730	.228	42.613 ***
WORD_ORD		9.676	.217	44.639 ***
NUM_REC		9.642	.206	46.824 ***
HAND_MOV		9.677	.216	44.874 ***
ATLANT_D		9.795	.203	48.228 ***
BLOCK_C		10.297	.228	45.162 ***
ROVER		10.254	.221	46.420 ***

Intercepts: (female - intercepts equal)				
		Estimate	S.E.	C.R.
RIDDLES		10.449	.270	38.696 ***
VERB_KNO		10.143	.244	41.656 ***
EXP_VOC		9.898	.235	42.187 ***
GESTALT		10.116	.196	51.520 ***
TRIAN_UN		10.414	.210	49.680 ***
REBUS		9.877	.223	44.251 ***
REBUS_D		9.822	.231	42.556 ***
ATLANTIS		9.730	.228	42.613 ***
WORD_ORD		9.676	.217	44.639 ***
NUM_REC		9.642	.206	46.824 ***
HAND_MOV		9.677	.216	44.874 ***
ATLANT_D		9.795	.203	48.228 ***
BLOCK_C		10.297	.228	45.162 ***
ROVER		10.254	.221	46.420 ***

Means: (female - intercepts equal)				
		Estimate	S.E.	C.R.
Gc		-.195	.351	-.554 .580
Gv		-.648	.260	-2.489 .013
Glr		.638	.307	2.079 .038
Gsm		.464	.246	1.890 .059

Figure 20.9 Intercept invariance detailed results, part 2.

Covariances: (male - intercepts equal)				
		Estimate	S.E.	C.R.
Gc	<-->	Gv	4.228	.742 5.701
Gc	<-->	Glr	4.900	.815 6.009 ***
Gsm	<-->	Gc	3.393	.665 5.102 ***
Gv	<-->	Glr	3.191	.601 5.309 ***
Gsm	<-->	Gv	2.419	.499 4.848 ***
Gsm	<-->	Glr	2.695	.546 4.934 ***
e8	<-->	e9	1.812	.624 2.906 .004
e10	<-->	e11	3.936	.686 5.737 ***

Covariances: (female - intercepts equal)				
		Estimate	S.E.	C.R.
Gc	<-->	Gv	3.314	.587 5.650 ***
Gc	<-->	Glr	4.475	.748 5.985 ***
Gc	<-->	Gsm	3.413	.643 5.309 ***
Gv	<-->	Glr	3.474	.566 6.142 ***
Gv	<-->	Gsm	2.343	.462 5.072 ***
Glr	<-->	Gsm	3.164	.594 5.325 ***
e8	<-->	e9	1.767	.537 3.291 .001
e10	<-->	e11	2.679	.542 4.945 ***

Correlations: (male - intercepts equal)				
		Estimate		
Gc	<-->	Gv	.677	
Gc	<-->	Glr	.725	
Gsm	<-->	Gc	.678	
Gv	<-->	Glr	.675	
Gsm	<-->	Gv	.691	
Gsm	<-->	Glr	.711	
e8	<-->	e9	.613	
e10	<-->	e11	.678	

Correlations: (female - intercepts equal)				
		Estimate		
Gc	<-->	Gv	.702	
Gc	<-->	Glr	.697	
Gc	<-->	Gsm	.681	
Gv	<-->	Glr	.871	
Gv	<-->	Gsm	.753	
Glr	<-->	Gsm	.748	
e8	<-->	e9	.570	
e10	<-->	e11	.558	

Variances: (male - intercepts equal)				
		Estimate	S.E.	C.R.
Gc		8.930	1.262	7.078 ***
Gv		4.372	.779	5.611 ***
Glr		5.114	.946	5.409 ***
Gsm		2.806	.666	4.216 ***
e1		2.494	.426	5.861 ***
e2		1.663	.317	5.249 ***
e3		2.603	.387	6.733 ***
e4		3.594	.460	7.819 ***
e5		2.657	.475	5.599 ***
e8		2.495	.631	3.955 ***
e9		3.504	.732	4.786 ***
e10		6.652	.902	7.373 ***
e12		4.283	.686	6.242 ***
e13		5.315	.736	7.224 ***
e14		5.652	.795	7.110 ***
e11		5.074	.685	7.406 ***
e6		4.124	.633	6.515 ***
e7		5.860	.784	7.470 ***

Variances: (female - intercepts equal)				
		Estimate	S.E.	C.R.
Gc		7.609	1.049	7.256 ***
Gv		2.933	.583	5.034 ***
Glr		5.424	.943	5.752 ***
Gsm		3.298	.747	4.414 ***
e1		2.021	.389	5.197 ***
e2		2.907	.432	6.723 ***
e3		2.414	.366	6.597 ***
e4		5.346	.640	8.354 ***
e5		3.334	.487	6.840 ***
e8		2.787	.567	4.918 ***
e9		3.446	.635	5.430 ***
e10		4.273	.606	7.053 ***
e12		3.794	.620	6.120 ***
e13		5.429	.729	7.449 ***
e14		4.739	.678	6.991 ***
e11		5.391	.697	7.737 ***
e6		4.754	.645	7.377 ***
e7		5.685	.719	7.911 ***

Figure 20.10 Intercept invariance detailed results, part 3.

Okay, we have now established intercept, scalar, or strong invariance across groups. What does that mean at a practical level? One meaning was already presented: intercept invariance means that any difference in means across the groups on the subtests (measured variable) are the result of *true differences* in the underlying latent variables, not to something specific about the subtest. Intercepts are the mean values on the dependent variable (the subtests) for those with a value of zero on the independent variable (the factors). Another way of thinking about what intercept invariance means is that each measured variable has the same zero point for males as for females. That is, the scale for the measured variables starts at the same place. Metric invariance means that the scales use the same metrics across groups; intercept invariance means that the scales start at the same point.

Imagine if this were not the case, if the scales did not have the same starting point. Imagine, for example, if your speedometer were broken such that speed did not register until you were going 10 mph, and it consistently registered 10 mph under your actual speed. Imagine all the tickets you would get! Your measured speed would be due, in part, to your true (latent) speed but would also be due, in part, to having an incorrect starting point for speed measurement. Alternatively, consider measuring temperature in two different cities daily for a month. In one city you use a thermometer with a Celsius scale, but in the other you use one with a Kelvin scale (the same metric but with a zero point of -273° Celsius). The average temperatures will be different in part because the scales have different zero points.

Also at a practical level, intercept invariance is assumed (but rarely tested) any time we wish to compare means on some composite variable. In other words, any time we compare means across groups on some composite, we are assuming—whether we know it or not—that strong measurement invariance holds. Brown (2015) used the example of items designed to measure agoraphobia (unreasonable fear of being in places where escape is difficult, such as crowds or open spaces). We might expect women to express more fear of walking alone in isolated areas (one indicator of agoraphobia), compared to men, even when they have the same level of underlying agoraphobia. If so, this would show up as a difference in intercepts for this item across sexes. If not tested or taken into consideration in research on agoraphobia, this difference could lead to erroneous conclusions. If 4–5 agoraphobia items were simply summed, we might conclude that women had higher levels of agoraphobia when in fact they only differed on this single item. This example also illustrates what a difference in intercepts often means: that there is some more specific factor (e.g., fear of attack) that influences an item beyond the more general factor (agoraphobia).

We discussed bias in construct validity in connection with metric invariance (or a lack of metric invariance). A lack of intercept invariance would likewise suggest construct bias across groups. Suppose we had found a lack of intercept invariance for, say, the Rover subtest on the Gv factor (partial intercept invariance, see below). That would mean that one sex was scoring systematically higher or lower on this test, even after taking into account the differences in the latent Gv mean. That finding, in turn, would suggest that the Rover test was not a fair measure of Gv skills for the lower scoring group, because the test systematically underestimated that group's scores. Likewise, in this case, a composite score using this test would show systematic bias for one group versus the other.

There are several additional points to consider concerning the topic of intercept invariance. First, you may question whether the metric invariance and the intercept invariance models are truly nested. After all, nested relations are those in which one model can be derived from the other by imposing parameter constraints. In going from metric to intercept invariance, we have imposed 14 constraints (the intercepts for one group constrained to be equal to those of the other) but have also freed four previous constraints (the latent factor mean differences were freely estimated for females). Good for you if you wondered about this, but the models are indeed nested. An alternative method for estimating the *metric*

invariance model would be to constrain the intercepts of the reference variables (those with a factor loading of 1, e.g., Riddles, Triangles) to be equal across groups, but free the factor means for girls. This alternative metric invariance model would have the same degrees of freedom (because 4 constraints are added and 4 dropped) and an identical fit. The results are the same. To go from this model to the intercept invariance model then merely requires that the remaining 10 intercepts be constrained equal across groups (for an additional 10 *df*).

Second, it is possible, and indeed common, to test the first two invariance steps without estimating means structures. Model fit and model results will be the same as those shown here. Of course it is necessary to estimate means and intercepts to test for intercept invariance. The point here is, however, that it is perfectly acceptable to estimate configural and metric invariance without estimating mean structures and then compare those models to an intercept invariance model.

Third, in my experience, complete intercept invariance is often harder to fulfill than metric invariance (and I have talked to others far more knowledgeable than I am about this topic who have reported the same thing). When complete intercept invariance is not achieved, one option is to test for partial intercept invariance by freeing selected intercept constraints for one or more groups (Byrne et al., 1989). Candidates for such model relaxations may be found through inspection of findings (e.g., modification indices) or based on theoretical grounds (Reynolds & Keith, 2013). As already noted, the most common reason for partial intercept invariance is the existence of unmodeled minor specific (or common) factors. Recall the agoraphobia example. As with any ad hoc model comparisons, results-based partial invariance relaxations should be done sparingly. It is easier to come up with reasons for partial invariance after the fact than it is prior to model testing!

How many loadings or intercepts can differ and a model still be considered partially invariant, as opposed to not invariant? One rule of thumb is that a factor should have at least one other invariant indicator (other than the reference variable) in order to consider it partially invariant (Byrne et al., 1989). For more information concerning the topic of partial invariance, see Byrne and colleagues (1989), Gregorich (2006), Reynolds and Keith (2013), or Vandenberg and Lance (2000). Another common solution to this potential problem is to recognize that the $\Delta\chi^2$ test we have been using to compare models may be too sensitive and to use an alternative (e.g., ΔCFI) for invariance tests (Cheung & Rensvold, 2002). Intercept invariance is often not supported when there are many groups being compared. A method known as “multiple group factor analysis alignment” has been proposed as an alternative to traditional invariance testing in such cases, and allows the estimation of group factor means and variances even when intercept invariance is not supported (Asparouhov & Muthén, 2014; for an example, see Marsh et al., 2017).

Fourth, and finally, when item level analysis is conducted (e.g., the agoraphobia example), and intercept invariance is not achieved for an item, this finding is evidence of what is known as differential item functioning (DIF) in the psychometric literature.

Residual Invariance

The final step, at least in the current presentation, in invariance testing requires that the residual variances (and covariances, if any) for the measured variables also be equal across groups. Meredith (1993) termed this “strict” factorial invariance. It is also sometimes referred to as “invariant uniquenesses” (Vandenberg & Lance, 2000). This level of *factorial* invariance is consistent with *measurement* invariance, that is, all differences in the means and variances of the observed scores are completely explained by mean and variance/covariance differences in the latent factors.

Not all writers consider residual invariance necessary, or the next step, in invariance testing (Vandenberg & Lance, 2000). Metric invariance is needed prior to intercept invariance, and

intercept invariance is needed prior to testing for differences in latent means. But strict factorial (residual) invariance need not follow intercept invariance testing immediately. Many writers would test for substantive differences in latent means, or factor variances and covariances next, deeming invariance in residuals a relatively minor, and perhaps unlikely, issue (cf. Brown, 2015). I present residual invariance as the next step here for several reasons. First, configural, metric, intercept, and residual invariance variance all focus on different aspects of *measurement invariance*; other aspects of invariance (latent means and variances) can be considered *structural aspects* of invariance (Byrne et al., 1989; Vandenberg & Lance, 2000). Or, said differently, configural through residual invariance focus on how the measured variables relate to the latent variables; the other aspects focus on the latent variables themselves. Second, if we were to select two subsamples at random from a larger group, we should not necessarily expect equality of factor variances and covariances (Meredith, 1993; Widaman & Reise, 1997), so these should be considered later steps. Finally, the order of these subsequent invariance tests should, in most cases, make little difference in findings. So, using the present example, changing the order of invariance tests (e.g., testing latent means next versus last in the series) made very little change in the χ^2 associated with each step. This makes sense if we have faith in the power of these models to separate the different aspects of measurement, such as unique variances, latent variances, covariances, means, and so on.

For the subtest Residual Invariance model, the residual variances (also known as errors, or unique variances) for the 14 KABC-II subtests were constrained to be equal for males and females. The two subtest residual covariances (between Rebus and Rebus Delayed and Atlantis and Atlantis Delayed) were also constrained to be equal for the two sexes. These constraints can be seen in the labels used for these parameters in Figure 20.11. The figure is for the male subsample; the same labels were used for the female subsample, thus constraining the values to be equal across the sexes. The fit of the residual invariance model is shown in Table 20.2. Once again, the model fit well and the change in χ^2 was not statistically significant. The additional constraints are “worth” the cost of the slight increase in χ^2 , and we would likely accept the Residual Invariance model as a reasonable representation of the cross-sex structure of the KABC-II. The ΔCFI criterion would also support the residual invariance step.

Residual invariance is more difficult to obtain than are metric and intercept invariance. If it does not hold, then, as with other types of invariance, it may be possible to achieve partial residual invariance by allowing some of the measured variable residuals to vary across groups. Residual invariance is also not as important as are the previous forms of invariance. As noted by Widaman and Reise, metric (weak) and intercept (strong) invariance are most important “for most substantive research questions” and residual (strict) invariance is “nice but not necessary” (1997, p. 296). Metric invariance is needed in order to compare the effects of one variable on another across groups (i.e., to compare paths in SEM), and intercept invariance is needed to compare mean structures across groups. The addition of residual, or strict, invariance means, however, that “group differences in the factor means and variances account fully for all group differences in subtest scores.” (Reynolds & Keith, 2013, p. 45; cf. Meredith & Teresi, 2006). Recall the distinction made earlier between measurement invariance and factorial invariance. Here, we have been demonstrating testing for factorial invariance as a way of demonstrating measurement invariance. Residual, or strict, “invariance is consistent with measurement invariance because group differences would only be attributed to group differences in the latent variables” (Reynolds & Keith, 2013, p. 74; cf. Meredith & Teresi, 2006). Just as strong invariance allows one to make valid comparisons of factor and observed means, strict invariance allows one to make valid comparisons of variances and covariances of the observed variables (Gregorich, 2006). This final statement is the case because the observed variables are affected both by the factors and by the errors of measurement (the values constrained in the strict invariance model).

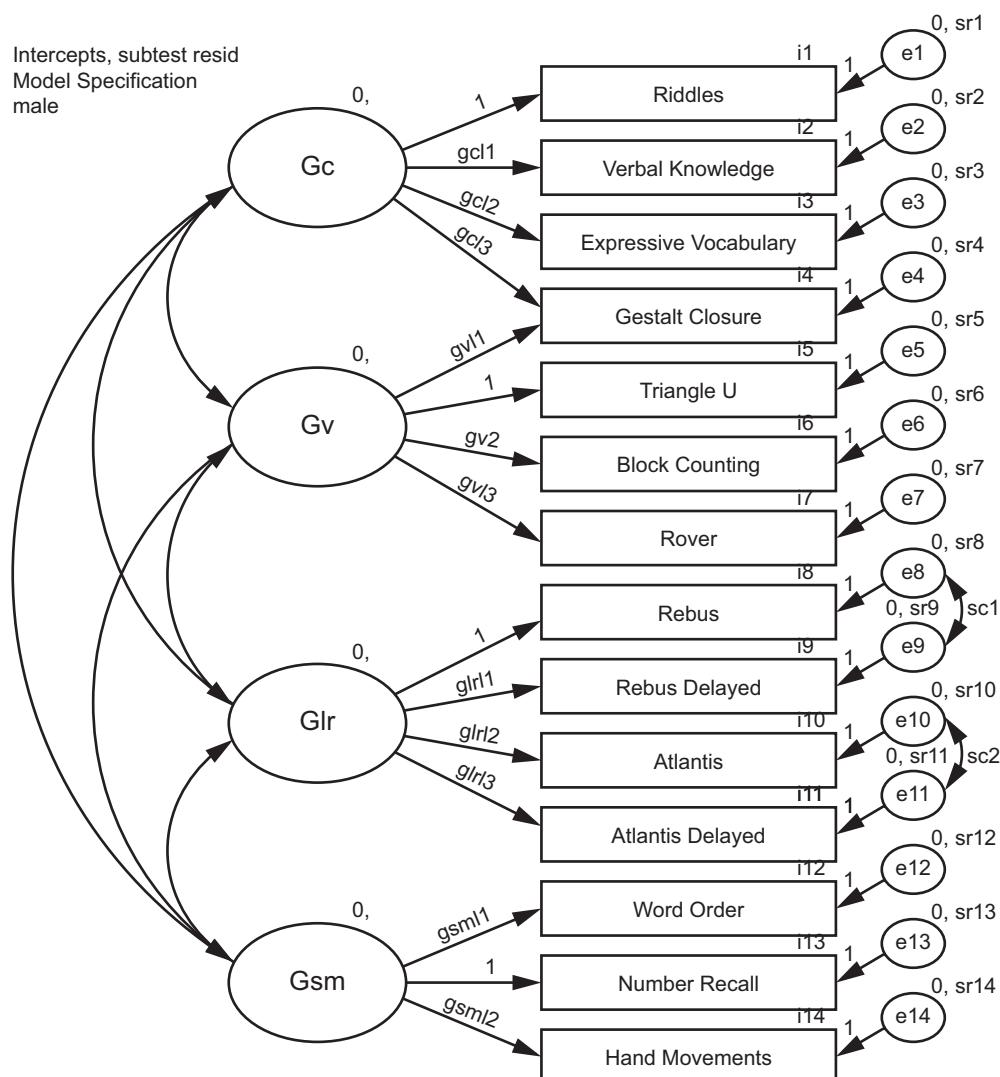


Figure 20.11 Residual (strict) invariance. Subtest residual variances constrained to be equal across groups.

Structural Invariance: Factor Variances Equal

Testing for invariance in measurement residuals completes the steps involved in testing for measurement invariance. Subsequent steps test for *substantive differences* in structural aspects of the CFA model, that is, the characteristics of the latent variables (variances and means) and how they relate to one another (covariances). These often reflect substantive research questions about the nature of constructs of interest, such as whether and how the constructs differ across groups. In contrast, the measurement aspects of invariance ask whether the measurement instruments work equally well across groups (to accurately assess the constructs of interest). So, with the current example, Reynolds and colleagues' primary interest was whether there were differences for males and females in their mean levels of different aspects of intelligence (2008); previous research had suggested differences favoring both males and females in various aspects of intelligence. Research has also suggested that males may overpopulate

the two ends of the normal curve, and thus show greater variance on intelligence than females (Johnson, Carothers, & Deary, 2008). Reynolds and colleagues also tested this possibility.

Because measurement invariance and structural invariance have different orientations, some methodologists recommend using different criteria to judge model fit. Little, for example, suggested the possibility of using a “modeling rationale,” and focusing on what I have called stand-alone fit indices to judge the overall fit of the measurement models, but then using a “statistical rationale” (e.g., the statistical significance of $\Delta\chi^2$) to compare the fit of the structural model (Little, 1997, pp. 58–59). Others have provided rules of thumb for judging changes in fit indices such as CFI in tests of invariance (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Another possibility would be to use different criteria for judging measurement invariance versus testing substantive hypotheses (e.g., $p < .01$ versus $.05$, respectively). My current approach is generally to use ΔCFI of less than $.01$ for tests of measurement invariance, and $\Delta\chi^2$ for the substantive (structural invariance) tests.

When testing for differences in these structural parameters, most methodologists begin by constraining the factor variances to be equal (see Vandenberg & Lance, 2000 for variations in these recommendations, however), followed by factor covariances. This makes sense to study how variables vary before examining how they covary with one other. Figure 20.12 shows the model for females with these and all other parameters constrained to be equal (we

All + all means
Model Specification
female

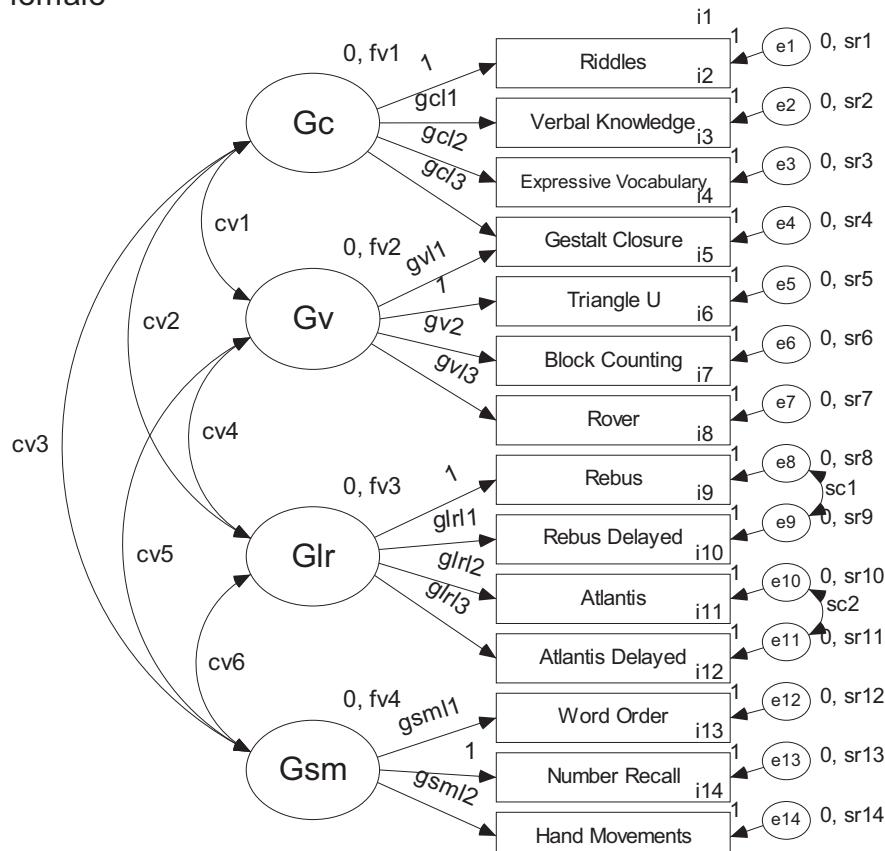


Figure 20.12 Latent factor variances, covariances, and means constrained to be equal across groups.

will refer to this Figure for the next two models, as well). The values above and to the right of the four factors (fv1, fv2, etc.) represent the constraints on the factor variances.

As shown in Table 20.2, this first set of structural constraints (factor variances equal) resulted in an increase in χ^2 of 3.895, which was not statistically significant with 4 degrees of freedom. The overall fit of the model was also good. I would accept these constraints as reasonable and conclude that the variances of the four latent factors are equal for 15- to 16-year-old boys and girls. Males and females show the same degree of variability (the same width for the normal curves) for verbal reasoning, visual-spatial reasoning, and so on. As noted earlier, because the factor variances are now constrained to be equal (in addition to the unstandardized factor loadings), the standardized loadings are also now the same across groups (although this equivalence is not shown here).

It would also be possible to test the equality of these factor variances one at a time, to determine if each factor in isolation showed equality of variances across the sexes. This could be planned in advance, or as a response to decrement in model fit at this step. Thus if we had found a statistically significant increase in χ^2 at this step, the next step would likely have been testing each variance, in turn, to determine which among them showed differences in variances across groups. For those that were statistically significant, we would likely have concluded that 15- to 16-year-old boys and girls had different degrees of variability on the underlying latent construct. Note again that the conclusions from this (and subsequent) invariance steps focus on the *constructs* being measured rather than the measuring instrument. Invariance in factor variances (and covariances and means) thus should not be expected, even given an excellent measuring instrument. “Still, if imposing invariance constraints on the $\hat{\psi}$ matrices [the variance-covariance matrix of latent factors] results in little worsening of fit, the resulting model—with $\hat{\psi}$ invariant across groups—is elegant” (Widaman & Reise, 1997, p.298). So far, our model is quite elegant!

Factor Covariances Equal

Figure 20.12 also shows equality constraints on the factor covariances (cv1, cv2, etc.). As shown in Table 20.2 (Model 6), this addition of 6 equality constraints also resulted in a small, but not statistically significant, increase in χ^2 , and all other model fit indices continue to look good. We can conclude that the degree that each factor relates to the others is equal across groups.

If both factor variances and factor covariances are equal across groups, then the factor correlations are equal across groups. Correlations, after all, are standardized covariances, standardized by taking into account the SDs, or variances, of the two variables. It is possible to test factor correlations (as opposed to covariances) explicitly by adding phantom variables (cf. Little, 1997), but that procedure is beyond the scope of this book. You may be tempted to think you could do this using the standardized (UVI constraint) model in the initial CFA chapter, but that method would confound earlier tests of invariance with the test of the equality of the factor correlations.

Factor Means Equal

The final step in this series is to test the equality of the means of the latent variables across groups. The model shown in Figure 20.12 shows the setup for this model also. Note that the factor means for females are all constrained to zero in this model, and recall that the factor means have been constrained to zero for males in every model tested. Thus for this model, not only are the intercepts constrained to be equal across groups, but the factor means are constrained to be zero for both groups. The fit statistics are shown in Table 20.2 (Model 7). As shown in Table 20.2 (Model 7), this set of constraints resulted in a statistically significant increase in χ^2 , along with a noticeable decrement in model fit according to the other indices. This finding suggests males and females show differences in their mean level of one or more of

the latent variables Gc, Gv, and so forth. This finding is also consistent with the detailed output from the intercept invariance model, which suggested that females had statistically significantly higher long-term retrieval (Glr) abilities and that males had higher visual-spatial (Gv) abilities.

The next five models (7a through 7e) probed these findings further. Models 7a through 7d tested each factor individually for mean differences. Constraining the Gv factor means to be equal for males and females resulted in a statistically significant increase in χ^2 (Model 7b), as did the model in which the Glr means were constrained to be equal (Model 7c). Again, the findings are consistent with those from the intercept invariance model. In contrast, a model constraining the Gc factor means to be equal did not result in a statistically significant degradation in model fit, nor did a model in which the Gsm means were constrained to be equal across the sexes. Interestingly, however, the model in which Gc and Gsm means were both constrained to be equal (but Glr and Gv means were allowed to vary) also fit worse ($\Delta\chi^2$ statistically significant) than did equal factor covariances model (Model 6). Presumably, although neither Gc nor Gsm were that different for males and females, constraining them in combination pushed the $\Delta\chi^2$ over the cut-point for statistical significance. What should we conclude? I think that taken together, the evidence suggests that we should consider Gc and Gsm to have equivalent means and variances. In contrast, long-term retrieval (Glr) abilities and visual-spatial reasoning (Gv) are measured equally well for males and females, but adolescent boys and girls show different levels of these abilities. Girls appear to have higher long-term retrieval abilities and boys appear to have higher visual-spatial skills.

It is worth reiterating that it is not necessary to test for equality of variances and covariances prior to testing for differences in factor means. It is necessary to demonstrate intercept invariance (or partial intercept invariance) prior to testing for differences in latent means. Likewise, metric (or partial metric) invariance is needed prior to testing for differences in factor variances and covariances.

To review, this section has served two purposes: it introduced the estimation of means and intercepts in confirmatory factor analysis, and it fleshed out in more detail the process of invariance testing in CFA. The demonstration of invariance in the measurement of constructs is an important topic, and an important use of multi-group CFA. Whether we realize it or not, such invariance is assumed in all research that makes cross-group comparisons; our research is much stronger if we demonstrate such invariance. As shown in previous chapters (and fleshed out here), metric invariance is needed in order to compare the effects of one latent variable on another (paths) across groups (Chapter 18), and intercept invariance is needed in order to compare latent means and intercepts across groups (Chapter 19). You may be tempted to think that these issues apply only to latent variable analysis, but such thinking is shortsighted. Most of the variables in our research are really latent variables. That is, we are generally interested in constructs and their effects on each other, and our various measures are only imperfect measures of those constructs. Our research will provide much stronger evidence of influences if we can demonstrate invariance across the groups being compared.

Alternative Model Specification

Here I have used the ULI approach to factor identification as the basis for the initial model specification. This is, I believe, the most common approach, and forms the basis for the automatic invariance-testing in Amos and Mplus. One disadvantage to that approach is that one indicator per factor (the one with the loading fixed to 1 across groups) is not tested for statistical significance, and is constrained equal across groups for all models. An easy solution if this is a concern is to re-conduct the analysis using a different reference variable and make sure all results are still acceptable. For the present example, all model fits and all conclusions were the same when different subtests were used as the indicators for each factor. Some values, such as unstandardized loadings and estimates of factor means, differed because different scaling was used, but standardized results and conclusions about models and the statistical significance of

model parameters were the same. Be aware that there are other alternatives for model specification, however. See Brown, 2015 and Kline, 2016 for suggestions.

Variance/Covariance Matrix of Measured Variables

Many methodologists, starting with Jöreskog (1971), actually recommend a comparison of the covariance (variance/covariance) matrices across groups as a first step in invariance testing. Why, you may wonder? Consider how we solve CFAs; what is the “fuel” for the analyses? CFAs are solved from the covariance matrices; the covariance matrices are used to estimate factor loadings, residual variances, and factor variances and covariances. Thus, if the covariance matrices are the same across groups, then the factor structures (excluding mean structures) must also be the same across groups. Said differently, any factor solution is contained within the covariance matrix. A test of the equivalence of the covariance matrices thus tests whether the instrument measures the same constructs across groups, but *without specifying* exactly what those constructs are. If this test is extended to include constraints on the means of the measured variables, then this test of the equivalence of the moment matrices (means, variances, and covariances) also subsumes the intercept invariance and factor mean equality steps.

Figure 20.13 shows a model designed to test the equivalence of the covariance and moment matrices across groups. The model appears complex (or at least, cluttered), but it really only includes three sets of constraints:

1. The variances of the measured variables (subtests), labeled vvv1_1 through vvv14_1 in the model shown for males (and labeled vvv1_2 through vvv14_2 for females, although not shown here);
2. The covariances of the measured variables, labeled ccc1_1 through ccc91_1 for boys and ccc1_2 through ccc91_2 for girls; and
3. The means of the measured variables, labeled m1_1 through m14_1 for boys and m1_2 through m14_2 for girls.

(For those using Amos, these constraints can be made automatically using the multiple-groups option in the analysis tab.)

Table 20.3 shows the fit indices for the models tested. The first model made no constraints across groups. This model simply estimates the means, variances, and covariance in the models. There are no constraints within groups; there are, for example, no factor loadings constrained to a value of zero, as in a CFA. There are also no constraints across groups. There are no constraints whatsoever, and thus this model has zero degrees of freedom and fits perfectly. There is really no need to include these values in the table, but I did so in order to make it clear that this “no constraints” model is our baseline for subsequent comparisons. For the second model, the variances and the covariances among the measured variables are constrained to be equal for males and females (this includes the first two sets of constraints in the previous numbered list). As shown in the table, this model had an excellent fit to the data (e.g., CFI = 1, RMSEA = 0), and even the $\Delta\chi^2$ from the model was non-significant. (I generally would not expect this for a model of this complexity and would likely place more emphasis on the stand-alone fit indices for this model.) This finding is consistent with our more detailed invariance testing, which suggested that the factor loadings, residual variances, factor variances, and factor covariances were all equal for males and females. This test is equivalent to Box’s M test for the equality of covariance matrices.

If the variances/covariances equal model did not fit well, what would that mean? It would mean that one or more aspects of measurement (loadings, residual variances, etc.) were not equal across groups (and those aspects of measurement were out of whack enough to mess up the fit of the entire model). If this were the case, our next steps would need to be the detailed invariance testing as summarized in Tables 20.2 and 20.4.

Test matrices

Model Specification

male

Most General Model

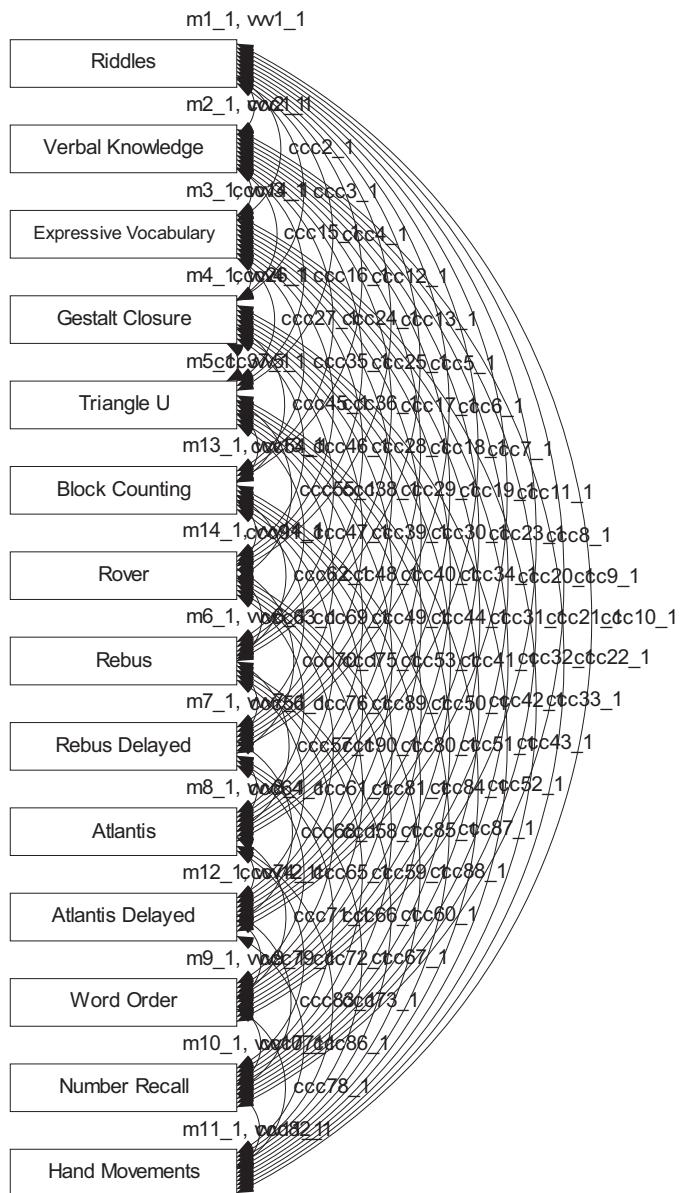


Figure 20.13 Testing the equivalence of the moment matrix across groups.

Table 20.3 Testing the Equivalence of Variance/Covariance and Moment Matrices

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	RMSEA	RMSEA*	SRMR	CFI	AIC
1. No constraints	.000	0			.000	.000	.000	1.000	476.000	
2. Variances & covariances equal	98.155	105	98.155	105	.669 .000		.000	.038	1.000	364.155
3. Means equal	139.586	119	41.431	14	.000 .024		.034	.038	.991	377.586

*RMSEA corrected for the number of groups

Note: All models compared to the previous model.

Table 20.4 Steps for testing for invariance with means and intercepts.

Model	Also Known As	Model Constraints Across Groups	Meaning	Practical Implications, if Met
0. Invariant matrices	Equality of moment matrices	Variances, covariances, and means of measured variables. There are no latent variables in this model.	The instrument measures the same constructs across groups (without demonstrating what those constructs are).	Can compare the effects of one variable on another (paths in SEM) across groups. If not supported, the detailed steps below are needed to find the source of the misfit. This omnibus test can mask differences in specific parameters, however.
Measurement Invariance				
1. Configural invariance	Pattern invariance	Same pattern of fixed and free loadings. Factor means constrained to zero for all groups.	Factors similar across groups	Can compare effects of one variable on another (paths in SEM) across groups.
2. Metric invariance	Weak measurement invariance, factor loading invariance	Factor loadings constrained equal. Factor means set to zero for all groups.	Construct is on the same scale for different groups; Same constructs measured across groups; Any differences in variation of the measured variable are due to latent variables	Step required before testing intercept invariance.
3. Intercept invariance	Strong measurement invariance, scalar invariance	Metric invariance + intercepts. Factor means constrained to zero for one group only .	Scale has the same starting point (zero point, intercept) across groups; Any differences in means on the measured variables are due to differences in latent means	Can validly compare the means on the latent (or composite) variable across groups. This step required before comparing latent means.

(Continued)

Table 20.4 (Continued)

Model	Also Known As	Model Constraints Across Groups	Meaning	Practical Implications, if Met
4. Residual invariance	Strict measurement invariance; uniqueness invariance; invariant error variances	Intercept invariance + measured variable residual variances (and covariances, if any)	Any differences in the measured variables are the result of the latent variables	Can validly compare variances & covariances for the measured variables
Structural Invariance				
5. Factor variances equal	Invariant factor variances	Metric invariance + variances equal	Do the latent variables have the same variance across groups? (this and subsequent models may be used to test substantive questions about differences across groups)	The normal curves for the latent variables will be equally wide or narrow
6. Factor covariances equal	Invariant factor covariances	Variances equal + covariances equal	Do the latent variables have the same relations among each other across groups?	If both variances and covariances are equal across groups, correlations among factors are also equal across groups
7. Factor means equal	Equal factor means	Intercept invariance + latent variable means constrained to zero for both groups	Are the true means on the latent variables equal across groups?	True means are equal on the constructs of interest. Used to test substantive questions about group differences on the latent means.

The third model shown in Table 20.3 constrained the means of the subsets to be equal for males and females. As shown in the Table, this set of constraints resulted in a statistically significant increase in χ^2 . If this were the criteria by which we judged the model, we would conclude that we needed to investigate this finding further. We could do so by testing whether the model misfit was due to a difference in some of the subtest means (via testing of intercept invariance), or if the model misfit could be explained by a more general difference in latent means on some or all of the latent factors. Our previous invariance testing suggests that the difference in means is a result of a significant difference for boys versus girls on the latent Gv and Glr factors.

It is interesting that if we had used only stand-alone fit indices to judge the final model in Table 20.3, we would have concluded that variances, covariances, and means are all invariant across the two groups. In this example, the addition of mean constraints led to a decrement in model fit but not enough of a decrement to make the overall fit bad. This finding is analogous to an omnibus F test being non-significant in ANOVA but some of the more detailed comparisons of means being statistically significant.

As noted, many methodologists suggest this as the first step in invariance testing (and I often use it in my own research). Many also suggest that if this model fits well across groups then detailed invariance testing is not needed. This makes perfect sense. But the current example also illustrates that the overall matrix may fit well (in this case, the moment matrix), but that detailed invariance testing may show differences in some aspects of the model. That is, an overall good fit at this step may mask actual group differences in measurement or structural invariance at the more specific steps. Which route should you take with your research? As always, carefully consider the purpose of your research and proceed based on the questions you want to answer. In the research on which this example was based, the researchers wanted to know if boys and girls differed in their mean levels of different types of intelligence (Reynolds et al., 2008). It thus made sense to conduct detailed comparisons of intercepts and factor means. If, on the other hand, your purpose is to compare across groups the effects of one variable on another, then a demonstration of well-fitting variance/covariance matrices across groups may be enough to establish invariance before your comparison of effects. If you are not sure which route to take, then the detailed invariance testing may well be the safest.

Table 20.4 summarizes these invariance steps when means and intercepts are estimated. The table also reviews the meaning of invariance at each step. The comparison of moment matrices is listed as step zero because it may or may not be used.

Invariance Testing Without Means

Many researchers conduct invariance testing without testing for differences in mean structures. Suppose, for example, you are interested in testing for invariance on a measure administered in different countries, with the measure standardized separately by country. All measured variables will have the same mean across countries, because the measure was developed that way (e.g., Chen, Keith, Weiss, Zhu, & Li, 2010). Or perhaps you are only interested in whether a latent variable (e.g., Homework) has the same effect on another latent variable across groups (e.g., Grades, as in Chapter 17). In this case, intercepts and means are not of interest. Whatever the reason, not all researchers include tests of invariance in mean structures in their tests of invariance (see, however, Little (1997) for an argument for always including mean structures). This section briefly outlines the steps in invariance testing when mean structures are not included.

When mean structures are not part of invariance testing, invariance testing can follow these steps:

1. configural invariance (same factor patterns)
2. metric invariance
3. invariance of measured variable residual variances and covariances

4. equality of factor variances, and
5. equality of factor covariances.

We could also add testing of equivalence of variance/covariance matrices as step zero. Again, not all writers would conduct these in the same series. But most would likely agree that step 2 (metric invariance) is the most important, and is required prior to subsequent steps. Practically, you can accomplish such invariance testing, in part, by simply turning off the option of testing means and intercepts in your model (e.g., by un-clicking the “estimate means and intercepts” box in Amos, or by including the command MODEL=NOMEANSTRUCTURE in Mplus, assuming, in both cases, that you have no missing data). These steps are summarized in Table 20.5.

Table 20.6 shows the fit of the models listed previously with the data for the KABC-II for ages 15–16. Note that most of the fit indices are identical for the first three models (Equal matrices through Metric Invariance) as they were in the analyses using mean structures (shown in Tables 20.2 and 20.3). Even when means and intercepts are analyzed they are not really considered in the model for these steps (comparison of variance/covariance matrices, configural invariance, and metric invariance). Because of this similarity in fit, some researchers conduct the first three steps without estimating means and intercepts and then add that estimation for the remaining steps from Table 20.4. I recommend that if you are going to analyze mean structures, go ahead and do so through all of the steps.

Note that the AIC is the exception to this rule of the same fit with and without the estimation of means; it differs for all models in Tables 20.2 and 20.3 versus 20.6. The AIC differs because its formula relies on the number of parameters in the model, rather than just the degrees of freedom, and the models have more parameters when means and intercepts are analyzed. This same caveat applies to the other, related fit indexes, as well (e.g., aBIC).

Higher-Order Models

In the previous chapter on CFA we also analyzed a higher-order version of the CFA model. As noted in that chapter, theory underlying some constructs, including intelligence, would suggest that those constructs are better understood by including a second- or even higher-order construct that helps explain the first order constructs. The theory underlying the KABC-II is hierarchical in nature, and thus a hierarchical or higher-order model is justified when doing research on this instrument. The steps involved in invariance testing are easily generalizable to a higher-order model.

Figure 20.14 shows a higher-order version of the KABC-II model. Note the components of this model: the first-order factor loadings, second-order factor loadings, subtest intercepts, intercepts for first-order factors, subtest residual variances and covariances, unique factor variances (and covariances, if applicable) for first-order factors (e.g., e15 through e18), and the second-order factor mean (for g). As in previous models, we can constrain these parameters to be equal across groups to test for different aspects of invariance.

Table 20.7 shows the steps involved in testing invariance for a higher-order model such as shown in the Figure. As with previous models, the ordering of some of the steps is not fixed. I have put the second-order steps in a similar sequence to the first-order steps, but you might change these depending on the purpose of your research. Reynolds and colleagues (2008), for example, were primarily interested in latent mean differences for males versus females on the various types of intelligence; as a result, they saved comparison of the second-order factor mean and the first-order factor intercepts for the last two steps in their analyses. We will not go through these steps here, but you will have the opportunity to do so in the Exercises!

Before we move to the next topic, I do want to make two points. The first point is that the second-order portions of the factor model, second-order factor loadings, first-order

Table 20.5 Steps for Testing for Invariance Without Considering Means and Intercepts

Model	Also Known As	Model Constraints Across Groups	Meaning	Practical Implications, if Met
0. Invariant matrices	Equality of covariance matrices	Variances and covariances of measured variables. There are no latent variables in this model.	The instrument measures the same constructs across groups (without demonstrating what those constructs are).	Can compare the effects of one variable on another (paths in SEM) across groups. If not supported, the detailed steps below are needed to find the source of the misfit. This omnibus test can mask differences in specific parameters, however.
Measurement Invariance				
1. Configural invariance		Same pattern of fixed and free loadings. Factor means constrained to zero for all groups.	Factors similar across groups.	
2. Metric invariance	Weak measurement invariance; factor loading invariance	Factor loadings constrained equal	Construct is on the same scale for different groups. Same constructs measured across groups. Any differences in variation of the measured variable are due to latent variables.	Can compare effects of one variable on another (paths in SEM) across groups.
3. Residual invariance	Uniqueness invariance; Invariant error variances	Metric invariance + measured variable residual variances (and covariances, if any)	Any differences in the variances and covariances of measured variables are the result of the latent variables.	
Structural Invariance				
4. Factor variances	Invariant factor variances	Metric invariance + variances equal	Do the latent variables have the same variance across groups? (This and the next model may be used to test substantive questions about differences across groups.)	The normal curves for the latent variables will be equally wide or narrow.
5. Factor covariances equal	Invariant factor covariances	Variances equal + covariances equal	Do the latent variables have the same relations among each other across groups?	If both variances and covariances are equal across groups, correlations among factors are also equal across groups.

Table 20.6 Tests of Equivalence of Factor Structure Without Consideration of Means and Intercepts

Model	χ^2	df	$\Delta\chi^2$	Δdf	p	RMSEA	RMSEA*	SRMR	CFI	AIC
0. Equal matrices	98.155	105			.000	.000	.038	1	308.155	
1. Configural	161.282	136			.025	.035	.047	.988	309.282	
2. Metric	172.236	147	10.954	11	.447 .024	.034	.051	.988	298.236	
3. Subtest residuals	194.894	163	22.658	16	.123 .026	.037	.051	.985	288.894	
4. Factor variances	198.633	167	3.739	4	.442 .025	.035	.057	.986	284.633	
5. Factor covariances	204.635	173	6.002	6	.423 .025	.035	.056	.986	278.635	

*RMSEA corrected for the number of groups

Note: Models 2 through 5 compared to the previous model.

Higher order Model Specification female

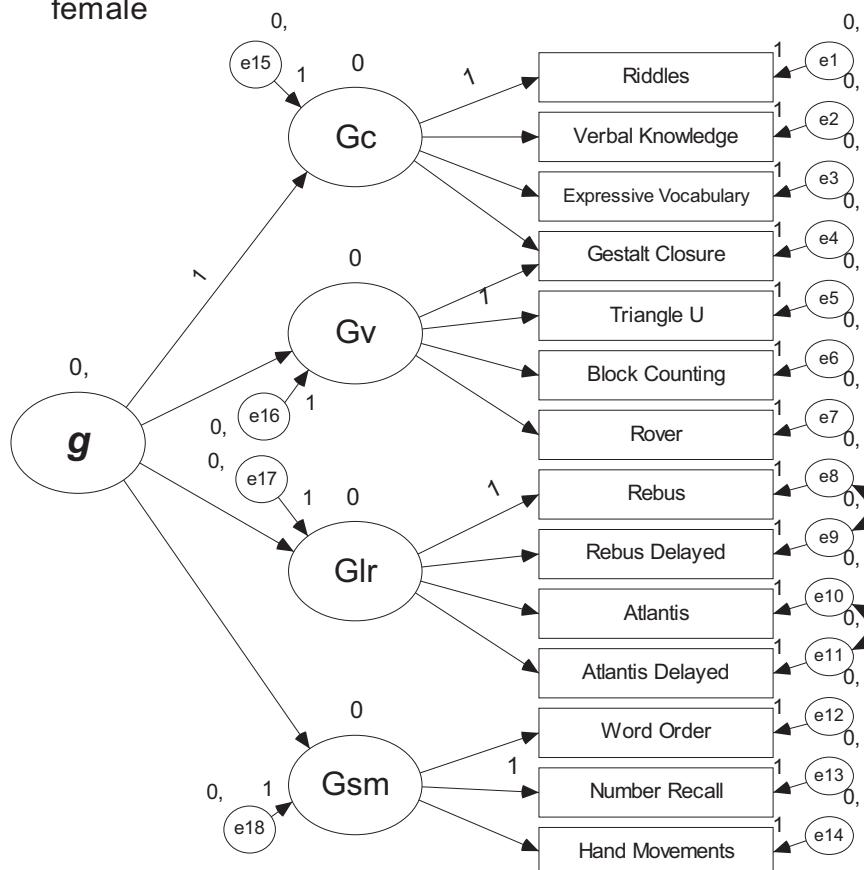


Figure 20.14 Higher order model for the KABC-II.

Table 20.7 Steps for testing for invariance of a higher-order model.

<i>Model</i>	<i>Also Known As</i>	<i>Model Constraints Across Groups</i>	<i>Meaning</i>	<i>Practical Implications, if Met</i>
0. Invariant matrices	Equality of moment matrices	Variances, covariances, and means of measured variables. There are no latent variables in this model.	The instrument measures the same constructs across groups (without demonstrating what those constructs are).	Can compare the effects of one variable on another (paths in SEM) across groups. If not supported, the detailed steps below are needed to find the source of the misfit.
Measurement Invariance				
1. Configural invariance		Same pattern of fixed and free loadings for both first- and second-order factors. First-order factor intercepts and second-order means constrained to zero for all groups.	Factors similar across groups	Can compare effects of one variable on another (paths in SEM) across groups. Step required before testing intercept invariance.
2. Metric invariance (first-order)		Factor loadings constrained equal. Factor intercepts and means set to zero for all groups.	First-order constructs are on the same scale for different groups; Same constructs measured across groups; Any differences in variation of the measured variable are due to latent variables	Can compare effects of one variable on another (paths in SEM) across groups. Step required before testing intercept invariance.
3. Intercept invariance	Strong measurement invariance, scalar invariance	Metric invariance + intercepts. First-order factor intercepts constrained to zero for one group only . Second-order means constrained to zero for both groups.	Scale has the same starting point (zero point, intercept) across groups; Any differences in means on the measured variables are due to differences in latent means for both groups.	Can validly compare the means on the latent (or composite) variable across groups. This step required before comparing latent means.

(Continued)

Table 20.7 (Continued)

<i>Model</i>	<i>Also Known As</i>	<i>Model Constraints Across Groups</i>	<i>Meaning</i>	<i>Practical Implications, if Met</i>
4. Residual invariance	Strict measurement invariance; uniqueness invariance; invariant error variances	Intercept invariance + measured variable residual variances (and covariances, if any) equal across groups	Any differences in the measured variables are the result of the latent variables	
5. Second-order loadings equal		Step 4 + second-order loadings set equal across groups. Second-order mean(s) set to zero for all groups	The second-order factor has the same meaning across groups	
6. First-order intercepts equal		Step 5 + first-order intercepts set to zero for both groups. Second-order mean set to zero for one group only.	No mean difference across groups on the first-order latent variables	
7. First-order unique variances equal		Step 6 + first-order unique variances equal across groups	The unique aspects of the first-order factors (that not explained by the second-order factor) are the same across groups	
8. Second-order variances equal		Step 7 + second-order factor variances equal across groups	The second-order factors are equally variable across groups	
9. Second-order factor means equal		Step 8 + second-order factor means equal across groups	Means equal across groups on the second-order factor	

intercepts, and so on, are considered aspects of the structural model rather than the measurement model. There are two reasons for this categorization. First, these parameters are estimated from the first-order factor covariances and means (just like the first-order loadings, and so on are estimated from the covariances and means of the measured variables). The first-order factor covariances and means are considered part of the structural model in Table 20.4, so why should they not be in Table 20.7? Second, the term “structural model” is generally used to refer to how latent variables relate to other latent variables. The second-order portion of the factor model also deals with exactly that: the relation of one set of latent variables on another.

The second point concerning the second-order invariance model concerns the nature of the first-order intercepts. If it is still unclear why these intercepts reflect the difference in means on the unique aspects of the first-order factors, then review the explanation of means and intercepts in the previous chapter. Second, consider an alternative method for setting up steps 4–6. The most common method (that shown in the table) would set the first-order intercepts for one group to zero and allow the intercepts to vary for the other group (with the second-order factor means set to zero for all groups). In these models, the differences in first-order intercepts then reflect the differences on the first-order factors controlling for the second-order factors. The first-order intercepts are then set to zero for both groups in step 7 to set the means of the first-order factors equal (controlling for the second-order factors) to see what happens to the fit of the model (here the second-order factor means are allowed to vary for all but one group). An alternative method for specifying this same model would be to set the first-order intercepts to zero for both groups (steps 4–6), but allow the *means of the unique factors* (e15 through e18) to vary for one group (females, in this example). The resulting difference in means of the unique aspects of the first-order factors using this method will show the same values as the difference in first-order intercepts using the original method. Then in step 7 we would constrain the means of the unique factors to be equal for both groups. Again, the results should be the same using either method. The second method may make it clearer exactly what is being compared in the different models.

In chapter 18 we also tested a different hierarchical structure, the bifactor model. Because all factors in a bifactor model are first-order factors, the invariance steps for this model are the same as those shown in Table 20.4 (although one might test for invariance for broad versus the general factor loadings in two or more steps).

SINGLE-GROUP, MIMIC MODELS

The previous chapter illustrated the consistencies (and inconsistencies) in the MG-MACS and the single-group/dummy variable approaches for testing models with mean structures. It is also possible to test some, but not all, aspects of invariance using a dummy variable approach. These models have a special name in CFA: they are commonly known as MIMIC models, with MIMIC an acronym for Multiple Indicators and Multiple Causes. MIMIC models are those in which (one or more) measured variables influence one (or more) latent variables, with those latent variables having multiple indicators. Alternatively, you can think of MIMIC models as a CFA model with covariates, or a CFA model with one or more measured variables influencing the factors.

A MIMIC version of the first-order KABC-II model is shown in Figure 20.15. (Note that this model is still referred to as a MIMIC, that is, multiple cause, model even though there is only a single cause, Female.) In this model we analyze a single group rather than conducting a separate (but connected) analysis for males and females. Instead of two groups, the group variable is contained in the model as a single sex variable (`sex_d` in the matrix; coded 0 for males and 1 for females and thus labelled Female in the figure). The data for this analysis are in the same Excel file as were the MG-MACS data (“`kabc cfa matrices.xls`”) but as the third worksheet; the data include the correlation matrix, means, and *SDs*.

Kabc-ii mimic
Model Specification

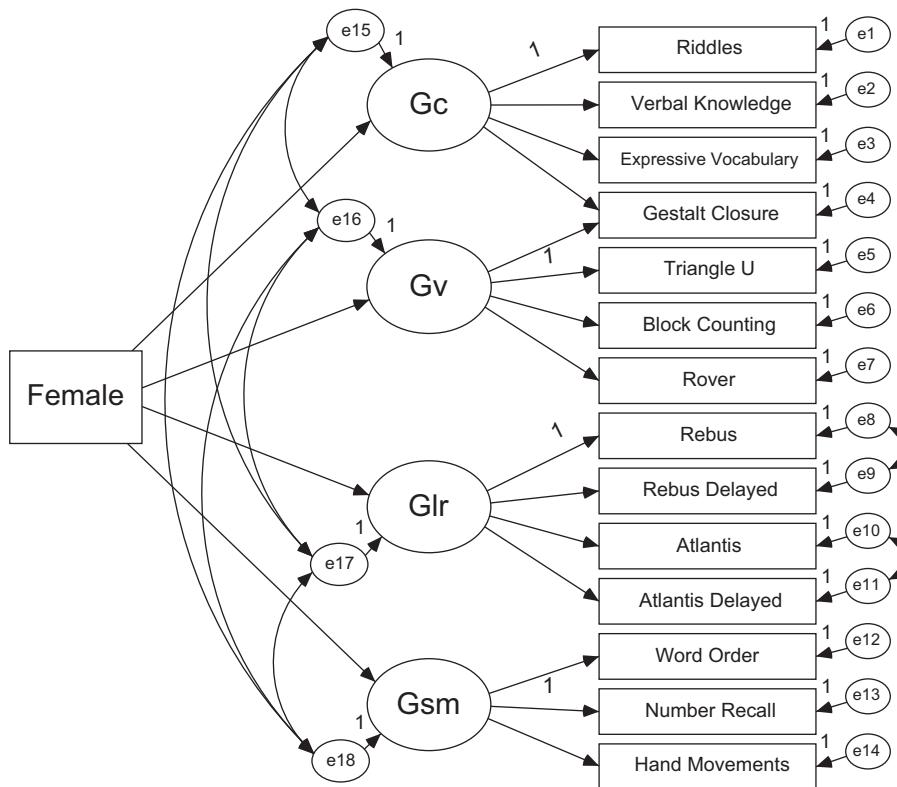


Figure 20.15 Testing for group differences in latent means using a MIMIC approach. Many aspects of measurement invariance are assumed, rather than tested. Intercept invariance can be tested, however.

Essentially, the model is a standard first-order CFA model but with a single measured, categorical variable, Female, influencing each of the first-order factors. Whereas our first-order model had correlations (covariances) among factors, in this model these correlations show up as correlations among the disturbances. Endogenous variables can't be correlated, but their disturbances can. These correlated disturbances are something that, in my experience, novices often forget, but you would be quickly alerted to the inadequacies of a model without them in the present case by the poor fit. (How poor a fit, you may wonder? Without the correlated disturbances, $CFI = .801$ and $RMSEA = .133$. Sex clearly does not account for the correlations among the first-order factors!)

The unstandardized figural results are shown in Figure 20.16. The model fits the data well according to most of our rules of thumb. The unstandardized estimates are shown because I want to compare the difference between males and females on the latent variables. In the MG-MACS model these were estimated by the difference in latent means for females as compared to males. In the MIMIC model they are estimated by the paths from the Female variable to each of the first-order factors. For the latent G_v factor this path is $-.65$ (or $-.653$ to 3 decimals). Because males are coded 0 and females 1, this means that females score $.653$ points lower on the G_v factor than do males. Think of this as the "effect" of going from being a male (coded 0) to being a female (coded 1); this one point change resulted in a $.653$ point decrease on the latent G_v factor.

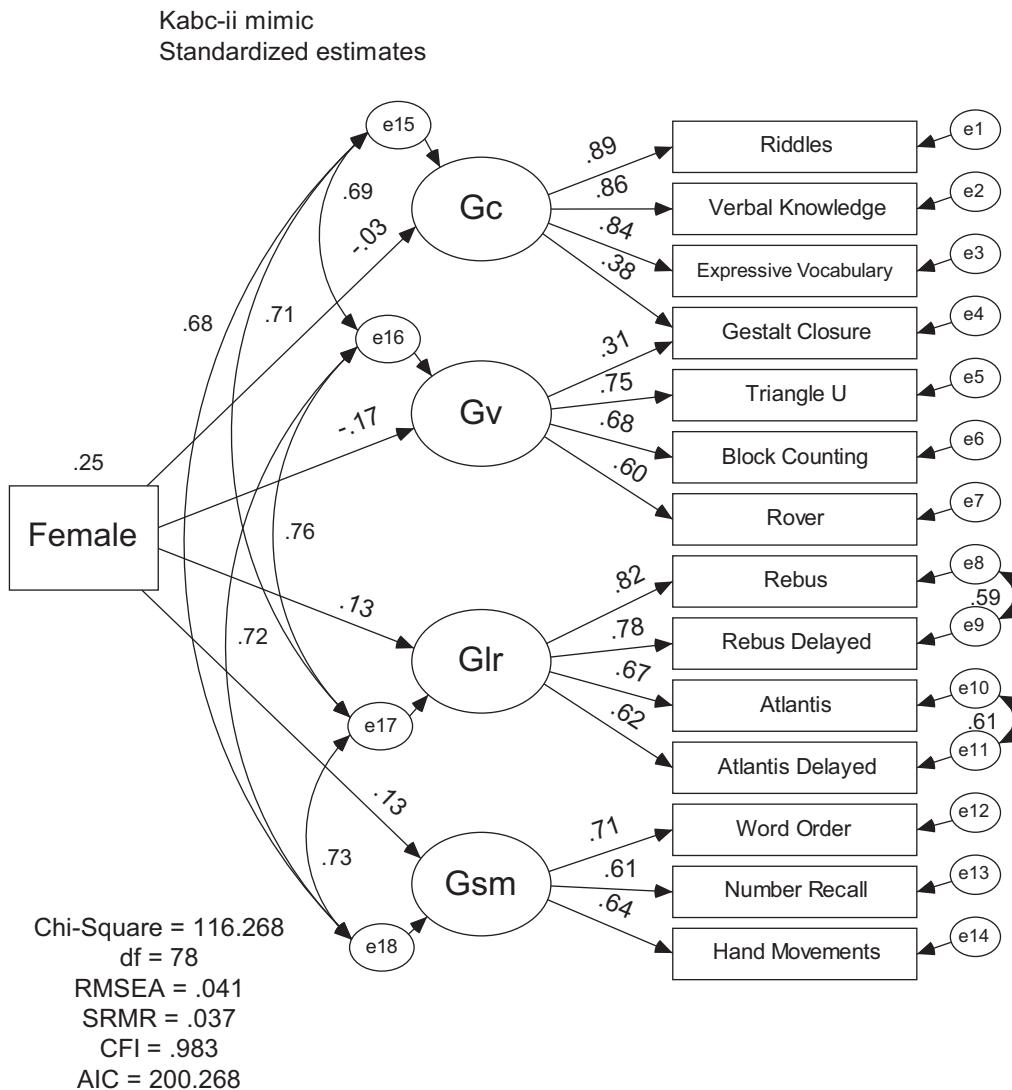


Figure 20.16 MIMIC results. Compare the unstandardized paths with the differences in latent means shown in Figure 20.9.

This finding is identical to the findings for the difference in means for step 7 (factor covariances equal) in the MG-MACS invariance testing: a .653 point difference favoring males. This is the model in which all parameters except the latent means were constrained to be equal. In fact, as shown in Figure 20.17, all estimates of mean differences are the same for the MIMIC model (on the left) as for the penultimate MG-MACS model (on the right). For the MIMIC model, these estimates of differences show up in the table of unstandardized paths from exogenous variables to endogenous variables (the values of interest are bolded); in the MG-MACS model they are in the table of mean differences for females as compared to males. Compare the other difference we found in the MG-MACS model, the difference favoring females on the Glr factor. The value shown on left side of Figure 20.17 is .623, the same as the value on the right (MG-MACS) side (.623).

Why did I compare the MIMIC results to those of step 7 in the MG-MACS analyses? This is the model in which all parameters—factor loadings, measured variable intercepts and residuals,

Regression Weights: (MIMIC)

			Estimate	S.E.	C.R.	P
Gv	<---	Sex_d	-.653	.262	-2.492	0.013
Gsm	<---	Sex_d	.466	.246	1.896	0.058
Gc	<---	Sex_d	-.196	.350	-.560	0.575
Glr	<---	Sex_d	.623	.305	2.040	0.041
RIDDLES	<---	Gc	1.000			
VERB_KNO	<---	Gc	.908	.046	19.626	***
EXP_VOC	<---	Gc	.848	.045	18.669	***
TRIAN_UN	<---	Gv	1.000			
REBUS	<---	Glr	1.000			
REBUS_D	<---	Glr	1.005	.047	21.241	***
ATLANTIS	<---	Glr	.900	.087	10.368	***
WORD_ORD	<---	Gsm	1.155	.134	8.623	***
NUM_REC	<---	Gsm	1.000			
ATLANT_	<---	Glr	.777	.081	9.618	***
BLOCK_C	<---	Gv	1.011	.098	10.349	***
HAND_MOV	<---	Gsm	1.079	.132	8.145	***
ROVER	<---	Gv	.905	.099	9.149	***
GESTALT	<---	Gc	.358	.076	4.712	***
GESTALT	<---	Gv	.426	.120	3.547	***

Means: (MAG-MACS, female - 7. factor cov)

	Estimate	S.E.	C.R.	P
Gc	-.196	.350	-.560	.576
Gv	-.653	.263	-2.487	.013
Glr	.623	.306	2.037	.042
Gsm	.466	.246	1.893	.058

Figure 20.17 Detailed output for the MIMIC model versus the MG-MACS model. Again, compare the paths from Female to the latent factors with the values of the latent means shown in in the right half of the figure.

and factor variances and covariances—except the factor means are constrained to be equal across groups. Those same constraints are also made in the MIMIC model, whether we know it or not. Because there is only one group for the MIMIC analysis, there is only one set of factor loadings, residuals, factor variances, and factor covariances; they are the same for males and females. What about the measured variable intercepts? Recall how differences in latent means show up in the MIMIC model: as paths from the Female categorical variable to the first-order factors. If the model included any differences in measured variable intercepts, these could likewise be modeled in the MIMIC model via paths from Female to the subtests. The fact that there are no paths from Female to any of the measured variables means that the measured variable intercepts are constrained to be equal across groups. In contrast, if you wished to test for *partial* intercept invariance, you could do so by including paths from Female to one or more measured variables.

If you consider this correspondence between models a little more completely, it becomes clear all of the assumptions the MIMIC model makes but does not test. The MIMIC model assumes that:

- 1) factor loadings,
- 2) measured variable intercepts,
- 3) measured variable residuals,
- 4) factor variances, and
- 5) factor covariances

are all invariant across groups. Only one of these—the measured variable intercepts—can be tested across groups in the MIMIC model. If these assumptions are valid, then the MIMIC model will provide valid estimates of mean differences on the latent factors.

Said differently, the MIMIC model assumes that the variance/covariance matrix of the measured variables is the same across groups. And this assumption can (and was) tested in the comparisons of matrices (Table 20.3). Thus if your main interest in model comparisons was to determine whether boys and girls differed on mean levels for any of these intelligences, then it would likely be reasonable to compare covariance matrices across groups. If

this comparison showed a reasonable fit, you could then compare latent means in a MIMIC model. You could test for intercept invariance by comparing the MIMIC model shown to a model with paths drawn from Female to all subtests (but not pointing to any of the latent factors). In contrast, if your primary interest is in establishing invariance in measurement across groups, you should go through the detailed invariance steps as shown in Table 20.4. For considerably more detail comparing MG-MACS and MIMIC models see Hancock (1997) or, for this same example, Reynolds et al. (2008).

SUMMARY

This chapter focused on the estimation of mean structures in confirmatory factor analysis. We did so by focusing on testing for invariance across groups in CFA, an important topic, and a needed step, before latent variables can be compared across groups. In the previous chapter on latent means, we noted that it was necessary to constrain factor loadings and measured variable intercepts to be equal across groups in order to be able to test for differences in *latent* variable means and intercepts. As explicated in this chapter, this prerequisite step goes by the label of intercept, or scalar, or strong factorial, invariance. It is worth exploring the topic of invariance in more depth, and the meaning of the information obtained at each step. We did so here by examining the structure of a common intelligence measure, the KABC-II, for adolescent boys versus girls.

When mean structures are estimated, common steps for establishing measurement invariance include:

1. Configural invariance, in which the same pattern of fixed and free loadings is tested across two or more groups.
2. Metric invariance, also known as weak factorial invariance or factor loading invariance, in which the values of the (unstandardized) factor loadings are constrained to be equal across groups. If established, this means that the scaling of the measure is the same across groups, meaning that a one unit change in the underlying latent variable results in the same change in the measured variables for the two (or more) groups. Metric invariance is needed in order to make valid comparisons of factor variances and covariances across groups (in CFA) or to make valid comparisons of effects (paths) across groups (in SEM).
3. Intercept invariance, also known as scalar invariance, or strong factorial invariance, in which the values of the measured variable intercepts are constrained to be equal across the groups. At the same time, the latent variable means are freely estimated in all but one group. This set of constraints says that any differences in the measured variable means are the product of differences in the true means of the underlying, latent variables. Intercept invariance also means that the measured variables have the same starting point (intercept) across groups. Intercept/strong factorial invariance is needed in order to make valid comparisons of factor means across groups (in CFA) or to make valid comparisons of latent variable (or composite variable) means and intercepts (in SEM).
4. Residual invariance, also known as strict measurement invariance, in which the measured variable residual variances (and covariances, if there are any) are constrained to be equal across groups. This set of constraints says that the errors of measurement are the same across groups. If residual invariance is established, it means that any and all differences in the measured variables are a result of the latent variables.

Be aware that not all writers would put these steps in the same order, and others would suggest other steps (e.g., testing the equivalence of the moment matrices).

As noted, if metric invariance is established, it is possible to test the equality of latent variable variances and covariances in CFA, or the equality of effects in SEM. If intercept invariance is established, it is possible to test for differences in latent means and intercepts across groups. These tests generally address substantive hypotheses about the nature of the latent variables, however, rather than aspects of how well the observed variables measure the constructs of interest (measurement invariance).

It is possible to conduct invariance tests without reference to means and intercepts. The reasons one might do so and the steps involved are discussed in the chapter. In addition, it is possible to conduct invariance testing on higher-order models, in which the first-order factors are considered as indicators of a second-order one.

It is tempting to think that the topic of invariance is applicable only if you are interested in validating some measurement instrument, but that is emphatically not the case. Just as it is important to attend to measurement whenever you conduct research, it is important to attend to invariance whenever you compare groups in research. The comparison of the effects of one variable on another across two or more groups (e.g., paths in SEM, regression coefficients in MR) presumes that there is metric invariance across the groups. The comparison of means of latent or composite variables across groups presumes that there is intercept invariance across groups on the variable being compared. This admonition applies whether those groups are based on some variable of interest, such as sex in the present chapter, or when we wish to compare treatment groups in experimental research. It applies equally to nonexperimental, quasi-experimental, and experimental research. Invariance is important, and you now have the tools to test for it. The sources cited in this chapter are great places to turn for additional information on this topic.

EXERCISES

1. Conduct the analyses outlined in this chapter. The data are in the file labeled “kabc cfa matrices.xls.” The first two worksheets include the matrices for males and females separately. The third worksheet contains the matrix with sex as a part of the matrix (for the MIMIC model). See the website (www.tzkeith.com) for initial setup for these models for Amos and Mplus.
2. Conduct the higher-order invariance tests as outlined, but not detailed, in this chapter. Make sure your degrees of freedom match those shown in Table 20.8 below for each model:

Table 20.8 Degrees of Freedom to the Higher-Order Invariance Tests, Exercise 2

Configural	140
Metric	151
Intercept	161
Residual	177
Second-order metric	180
Second-order intercept	183
Second-order residual	187
Second-order variance/covariance	188
Second-order means	189

3. Test the invariance of shorter self-concept and locus of control scales from NELS across sex. The proper composition of these scales is represented in Figure 20.18. A reminder of the item wording is shown in Table 20.9. The data are in the excel file “sc loc matrix 2.xls” with a separate tab for boys and girls (there is also a tab with the full matrix). Note that the four self-concept items have been reversed so that for all items a high score represents a positive self-concept or an internal locus of control. Note also that the figure shown below represents a “conceptual model” only; it does not include errors or other crucial model details. (Your model should include those details, however!)
- a) Test this initial model across groups (configural invariance). You do not need to estimate means & intercepts at this step (but I recommend that you do so).
 - b) Add a covariance between the errors of measurement for items BYS44Dr and BYS44Er.

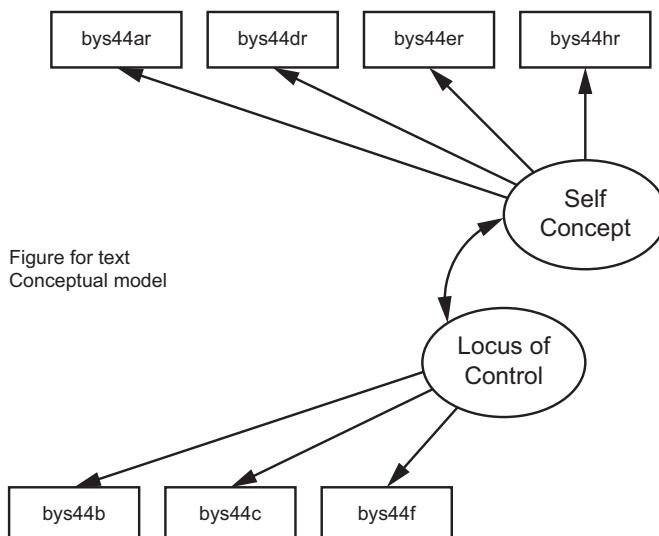


Figure 20.18 Conceptual model for Exercise 3: invariance of self-concept and locus of control for boys and girls.

Table 20.9 Self-Concept and Locus of Control Items for Exercise 3. Each item's response choices ranged from 1, strongly agree, to 4, strongly disagree. As shown, positively worded items were reversed so that for all items a high score represents a high self-concept or a high (internal) locus of control.

Variable	Label
bys44ar	I FEEL GOOD ABOUT MYSELF, reversed
bys44b	I DON'T HAVE ENOUGH CONTROL OVER MY LIFE
bys44c	GOOD LUCK MORE IMPORTANT THAN HARD WORK
bys44dr	I'M A PERSON OF WORTH, EQUAL OF OTHERS, reversed
bys44er	I AM ABLE TO DO THINGS AS WELL AS OTHERS, reversed
bys44f	EVERY TIME I GET AHEAD SOMETHING STOPS ME
bys44hr	ON THE WHOLE, I AM SATISFIED WITH MYSELF, reversed

- c) Test metric invariance across groups.
- d) Test intercept invariance (aka scalar or strong invariance; you *do* need to estimate means & intercepts at this step).
- e) Based on the text output, does it appear that boys and girls differ in their overall latent mean level of self-concept? Locus of control? On what did you base this conclusion? If you concluded that there were differences, which sex scored higher, and by how many points?
- f) Test the equivalence of the latent means for locus of control using a model constraint and fit statistics. Do boys and girls differ in their levels of locus of control? Why do you come to that conclusion?
- g) Test the equivalence of the latent means for self-concept using a model constraint and fit statistics. Do boys and girls differ in their levels of self-concept? Why do you come to that conclusion? Briefly interpret your findings from questions f and g (e.g., who, if anyone, scored higher?).
- h) Provide a table of fit statistics for the models listed in steps a through g. Be sure to list the corrected RMSEA.
- i) Would you be willing to accept configural invariance (step a)? Metric invariance? Intercept invariance? Briefly explain why or why not.

Notes

- 1 I say this is the “first step” in invariance testing, but it need not be. Many use a test of the equality of variance/covariance matrices as the first step. It is also reasonable to start with a very strict model and gradually free parameter constraints. The ordering of some of the other steps is fixed (we need to establish intercept invariance prior to testing for latent mean differences, for example), but for others different orders are recommended by different methodologists (Vandenberg & Lance, 2000). I will point out likely variations as the chapter progresses.
- 2 Strictly speaking, the next three types of invariance (weak, strong, and strict) are all forms of metric invariance (Widaman & Reise, 1997) because they all refer to the metric (scaling) of the instrument. I believe most writers use the term “metric invariance” as used here to refer to factor loading, or weak measurement invariance.
- 3 Although Mplus uses this method for general equality constraints, it has various defaults and short-cuts for invariance testing. See www.tzkeith.com for examples associated with this chapter and www.statmodel.com for details. One nice short-cut for automatic invariance testing is the addition of the command ANALYSIS: MODEL=CONFIG METRIC SCALAR, which produces fit indices and unstandardized output for these three models. Amos also can automate the testing of invariance, although I don’t always agree with the ordering the program chooses. Figure 20.13 is a product of this approach (Multiple Group Analysis in the Analyze menu).

21

Latent Growth Models

Unconditional, Simple Growth Model	517
Conditional Growth Model, or Explaining Growth	524
Additional Issues	528
<i>Data Requirements</i>	528
<i>Variations in Model Specifications</i>	529
Other Methods of Analyzing Growth Data	530
Summary	531
Exercises	532
<i>Note</i>	533

This chapter will cover latent growth modeling (LGM), also known as latent growth curve modeling. Such models will enable us to more closely and clearly study the process of longitudinal change. So, for example, we can use latent growth models to study the process of learning in children, the developmental trajectory of behavior problems in youth, the decline in cognitive functioning in old age, or even the developmental trajectory of children's height.

Consider some of our previous examples. In the early chapters in this text, we studied the effects of various influences (including homework) on achievement. Later, we examined the potential effects of different variables on achievement, while controlling for previous achievement, essentially asking whether these variables influenced the change in achievement over some time frame. Our panel models in Chapters 14 and 18 were designed to assess the potential effects of locus of control and achievement on each other, over time. Perhaps underlying all these models was a more basic question: what influences actual *growth* in learning and achievement? Although this may have been an underlying question in these examples, we were not able to get at that question directly. With LGM we will be able to do so more directly. With LGM it will be possible to study the influences on initial level of achievement and also the growth in achievement. And it will be possible to study the influence of growth in achievement and learning on other, subsequent variables.

We will start the chapter by revisiting, yet again, the topics of slopes and intercepts that we first addressed in the early chapters of this book. The graph shown in Figure 21.1 shows Math test data for 10 children from the Early Childhood Longitudinal Study (<https://nces.ed.gov/ecls/kindergarten.asp>). The test is designed to measure “conceptual knowledge, procedural knowledge, and problem solving” with math items ranging from simple (number knowledge) to advanced (algebra) (DiPerna, Lei, & Reid, 2007, p. 372). The scores form a

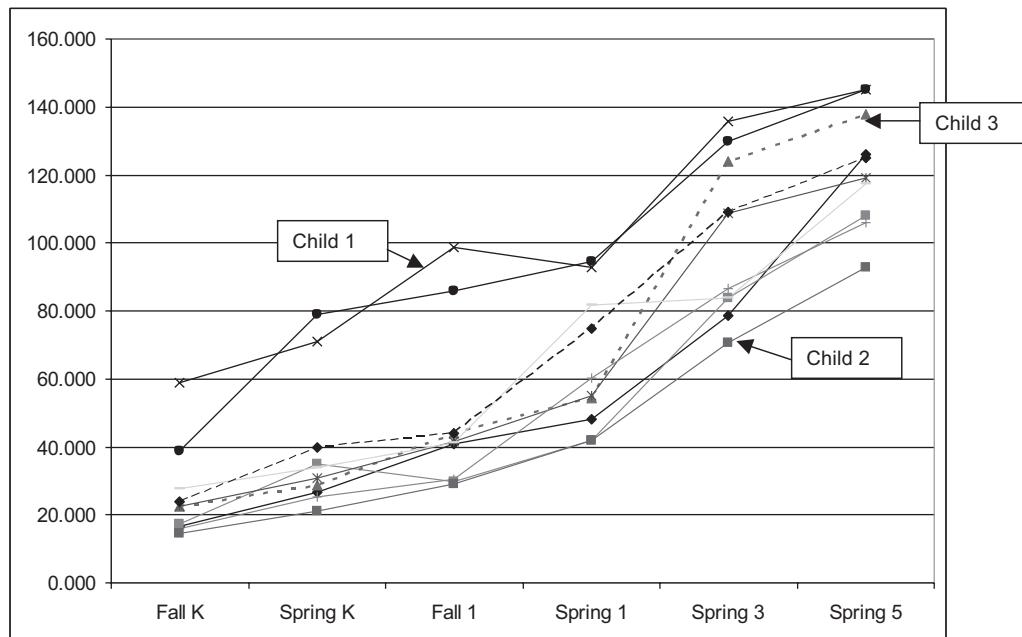


Figure 21.1 Math scores from Kindergarten to 5th grade for 10 students from the Early Childhood Longitudinal Study.

continuous measure of math skill and knowledge, with each child measured during the fall and spring of Kindergarten, fall and spring of first grade, spring of third grade, and spring of fifth grade. More difficult items were included in the third and fifth grade assessments, but these assessed the same skills. Note the line labeled Child 1; this girl started Kindergarten with relatively well-developed math skills, and then showed steady improvement. Child 2, in contrast, began Kindergarten with a lower level of math skills, and fell further behind as she progressed through the Spring of fifth grade. This is an illustration of the infamous Matthew effect, named after the Bible verse (Matthew 25:29): “For everyone who has will be given more, and he will have an abundance. Whoever does not have, even what he has will be taken from him” (New International Version).

The raw data are interesting, but if they form coherent patterns, perhaps we can summarize and help explain them. One method for doing so would be to correlate children’s scores time 1, time 2, and so on. Or we could regress Grade 5 scores on scores from grades K, 1, 2, and 3 to see if there was something unique about Kindergarten-level scores in explaining Fifth-grade math skills. This sort of focus has been the orientation taken so far in this text.

Another option would be to conduct a separate regression and create a separate regression line for each child. In this orientation, the data for each child would look like that shown in Table 21.1, with the scores on each administration of the math test shown in the first column, and time of administration shown in the second column. In this arrangement, the Fall K administration is time 0, the Spring K administration is time 1, and so on through time 5 for the Spring fifth grade administration of the math test. The data shown are for Child 1. You can regress this child’s math scores on the time of those math scores. If you were to do so, you would obtain the following regression equation:

$$\begin{aligned} \text{Predicted Reading} &= a + b \times \text{Time} \\ &= 56.128 + 17.715 \times \text{Time} \end{aligned}$$

This is the regression equation that describes Child One's math growth. Her initial predicted level of math achievement is described by the intercept, a score of 56.128. Recall that the intercept describes the predicted level on the dependent variable (Math) for a score of zero on the independent variable (Time). In the current setup, a level of zero on the independent variable represents Time 0, or testing in the Fall of Kindergarten. Figure 21.2 shows the regression line for Child 1, and compares it to the raw data for this same child. The slope for this regression line (17.715) is even more interesting than was the intercept: it represents our prediction of growth for this girl in math from one measurement to the next. Child one shows an average growth in math of 17.7 points from one measurement to the next. We could

Table 21.1 Math scores from Figure 20.1 for child 1.

<i>Math</i>	<i>Time</i>
58.810	0
71.070	1
98.800	2
92.950	3
135.600	4
145.270	5

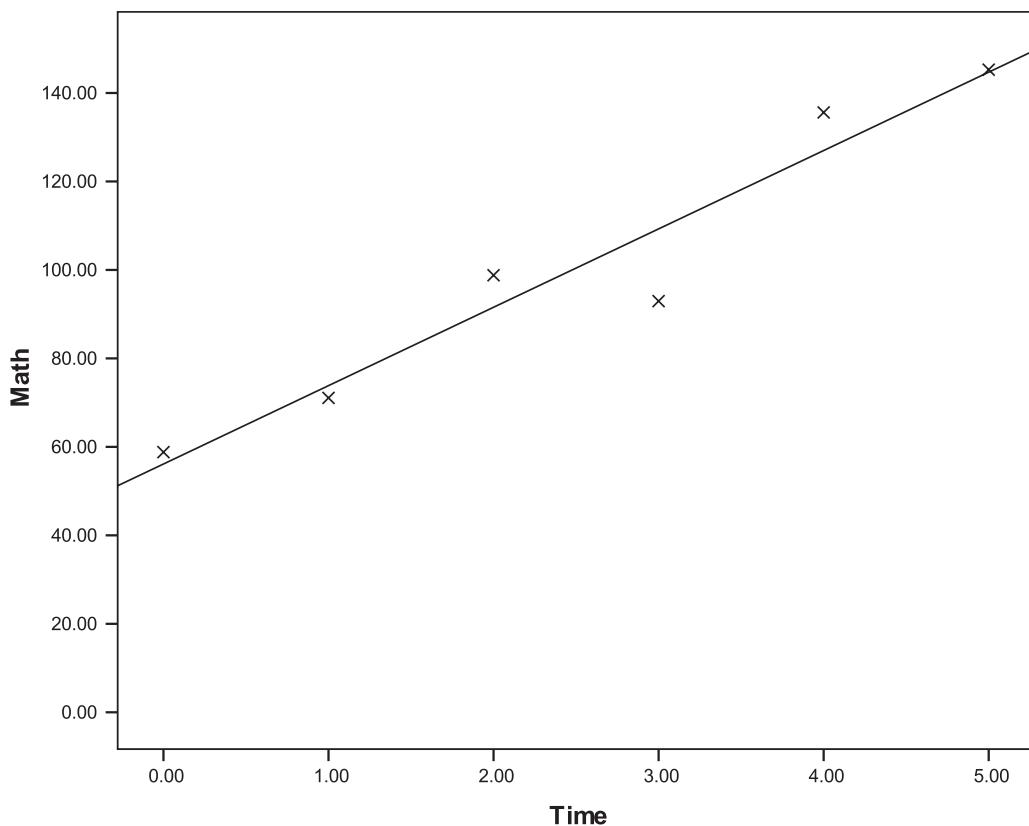


Figure 21.2 Child 1's math scores and a regression line of those math scores over time.

use this line to predict that her math score at time 6 would be approximately 163 (ignore for now the fact that the time intervals are uneven). Note that the regression line indeed does a good job of describing the raw data for this child.

We can conduct a similar regression for each of the ten children whose data are shown in Figure 21.1. The results of those 10 regressions are shown in Figure 21.3; each line represents the regression of math scores on time for an individual child. For each child, the intercept represents our prediction of his or her initial level of math. And the slope for each regression line represents our prediction of each child's growth in math. Again, note Child 2's regression line. She started Kindergarten with a Math score of approximately 5.6 (intercept), among the lowest beginning levels for the ten children shown. (Note that the intercepts—the zero point on the X axis—are slightly to the right of the Y axis.) Even worse, her rate of growth in math is lower than the other children shown in the Figures (a slope of 15.8 versus 19.6, the average slope). This child needs some sort of math intervention if she has not yet had one!

Look at Child 3, however (dashed line). Although she started K with lower math skills than the other children (intercept = 6.3), her growth in math exceeded the average (slope = 24.9), so that by grade 5 her math score was higher than the average. Whatever her teachers and parents are doing to teach her math certainly seems to be working!

We could, as was implied in the previous paragraph, average the intercepts and slopes for these 10 children to get an idea of the average starting level of math for these children, and the average growth in math from K to 5th grade. This would tell us something interesting about average initial level of math knowledge in Kindergarten as well as the average growth. We could also look at the standard deviations of the intercepts to get an estimate of individual differences in initial math knowledge, and we could look at the standard deviation of the slopes (the b 's) to get an estimate of individual differences in growth.

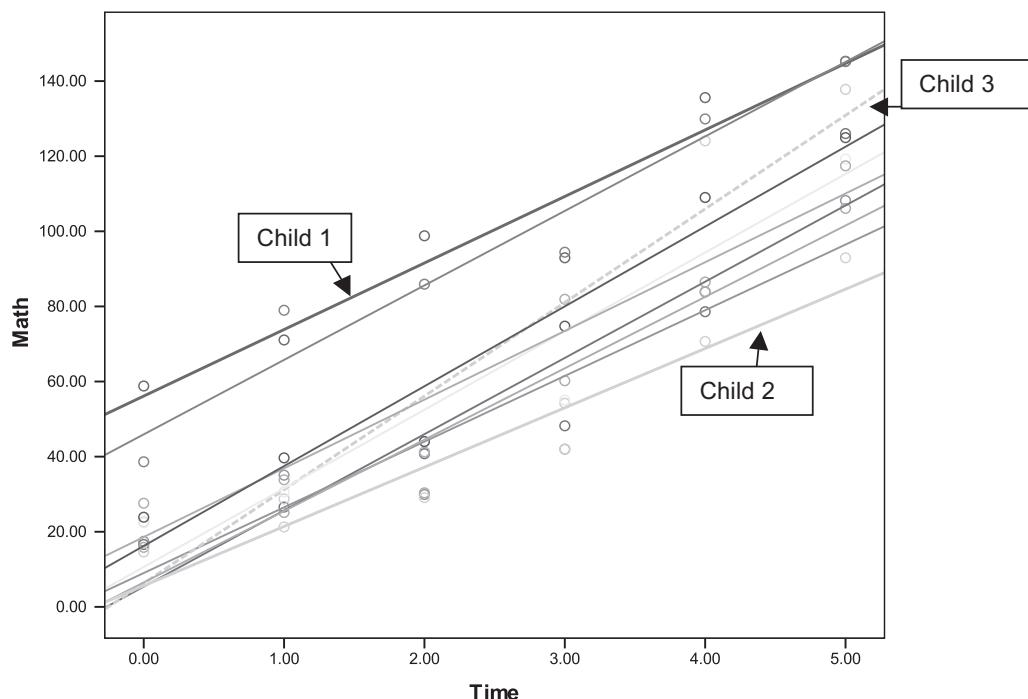


Figure 21.3 Regression lines for 10 students from ECLS. Math scores are regressed on time of test administration (from K through 5th grade).

Even better, we could use latent variable methods to get closer to the “true” (latent) intercept, or starting point, for these children and their “true” (latent) growth, or slope. This is what LGM does. We now turn to an example to illustrate the method.

UNCONDITIONAL, SIMPLE GROWTH MODEL

Recall that many SEMers urge a two-step process for estimating complex latent variable models: first estimate the measurement model and then add in the structural model. Most users of LGM follow a similar process in which they first estimate the growth portion of the model, and then add in influences on growth or the variables affected by growth. This is the approach we will take here. The initial model, the portion focusing just on the growth aspects of the model, is often referred to as the unconditional model, meaning that the growth aspects do not depend on (are not conditional on) other variables in the model. This initial model is sometimes simply referred to as the “change model” (Kline, 2016). Here, I will refer to this model as the simple growth model (with the understanding that “simple” is a relative term!).

The file “math growth final.sav” includes simulated data for 1000 children’s math scores (and other information). The data are loosely based on research by DiPerna, Lei, and Reid (2007). In that study, the researchers used data from the ECLS to study growth in mathematics skills from Kindergarten through grade 3. The researchers were interested in the possible influences on that growth from child behavior (e.g., teacher ratings of internalizing and externalizing behavior) and other characteristics (e.g., cognitive ability) measured in Kindergarten. The simulated data we will use includes measures of math skills for five equally spaced time points (math skill in the Fall of grades K, 1, 2, 3, and 4), in raw score units. Also included are child and parent variables measured in Kindergarten. These are Female (0=male, 1=female), Parent Education (highest years completed by either parent, from 11th grade through a PhD, coded 20), Cognitive Ability (on a standard IQ scale, $M=100$, $SD=15$), and the age at first assessment (in months). The descriptive statistics for these data are shown in Figure 21.4. Note the steadily increasing means for the math scores; indeed, as shown in Figure 21.5, these means barely depart from a straight line (the advantage of simulated data!).

Figure 21.6 shows the setup to estimate the simple growth model designed to understand growth in mathematics skills. The model, as shown, suggests that the observed math scores of these 1000 children are the product of two latent variables: their intercept, which we can also think of as the initial level of math skills, and their slope, that is, their growth from one observation to the next. I have here labeled these latent variables as the intercept and slope,

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
math1	1000	39.871148	107.642878	70.13811900	9.254201322
math2	1000	46.730771	113.480272	79.61711846	1.02109861E1
math3	1000	49.790040	128.998535	88.54294497	1.12984577E1
math4	1000	56.005987	136.398892	96.95733842	1.28533857E1
math5	1000	64.127584	150.988747	1.05854571E2	1.46851370E1
sex	1000	0	1	.51	.500
age	1000	61.12	74.51	68.5182	2.02936
ParEd	1000	11.00	20.00	16.1330	1.39329
Cognitive	1000	37.00	148.00	101.00010	14.99453
Valid N (listwise)	1000				

Figure 21.4 Descriptive statistics for the simulated math K-4 data.

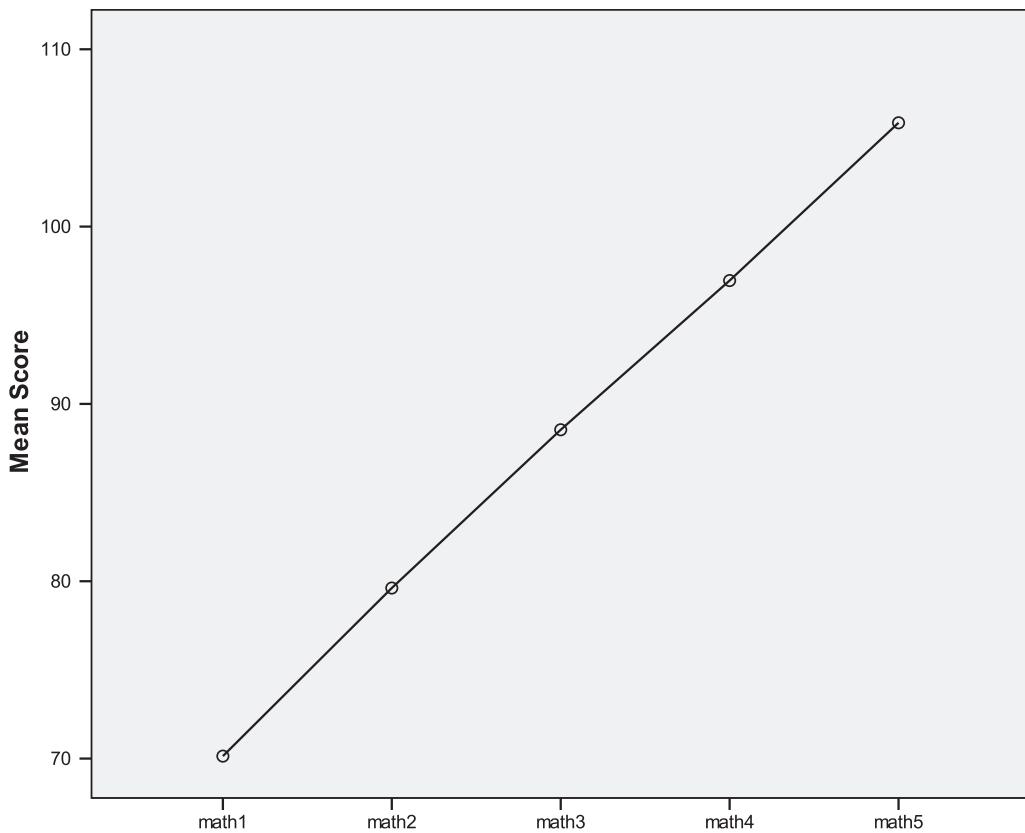


Figure 21.5 Average math scores over time.

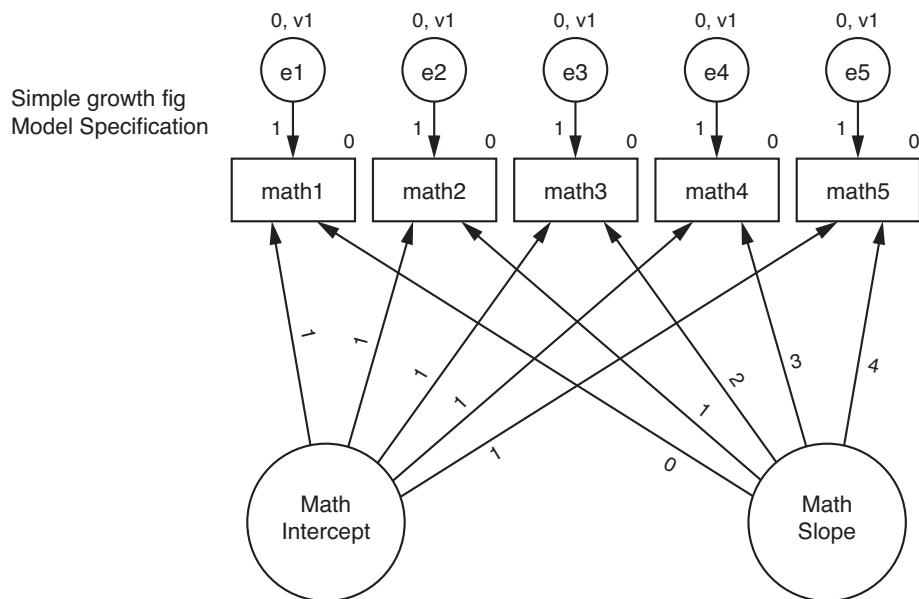


Figure 21.6 Initial simple, or unconditional, growth model of Math scores.

but names along the lines of “Initial Math Level” and “Math Growth” would be equally valid (although of course by initial we are referring to Fall of Kindergarten). The main things we will seek to estimate in this model are the means and variances of both of these latent variables. Consider what these represent: the mean for the intercept will represent the average level of beginning math skills, cleansed of measurement error and conceptually similar to the average of intercepts from our previous regression example. The variance of the latent intercept is slightly trickier but represents the variation in intercepts across the children. Look back at Figure 21.3: do all children have the same intercept? No, of course not, and we would not expect all children to enter Kindergarten with the same math knowledge. We can estimate that variation in the intercepts via the variance of the latent intercept. The mean for the latent slope variable represents the average systematic level of growth for these 1000 children across the assessments. As with the intercepts, not all children have the same slope; some show more growth, some show less. The estimate of the variance of the latent slope variable serves to estimate the degree of variation in growth these children show. And as with other latent variable analyses, the errors and unique variances of the measurements are separated out in the analysis, allowing the intercept and slope variables to more closely approximate the underlying variables of true interest (e.g., the real rate of change in Math skills).

The model and its constraints look similar to some of our CFA models with mean structures but also has some important differences. As in previous models, the means of the residuals, or measurement errors and unique variances ($e1$ through $e5$), are set to zero. Note the differences between this and previous models, however. First, note the “factor loadings,” all of which are constrained. The paths from the Math Intercept latent variable to each of the math tests are all constrained to 1. One way of estimating the intercept in an ordinary regression is to regress the Y variable on the X variables plus a constant variable of 1 (this happens behind the scenes when we conduct a regression). Setting the intercept-to-measured-variable paths to 1 forces this latent variable to be an intercept in LGM. The paths from the Math Slope latent variable to the measured math tests are constrained to linearly increasing values, starting with zero (0, 1, 2, 3, etc.). The first loading set to 0 indexes math 1 as the starting point for the growth. These constraints make the second latent variable serve as a latent slope variable. We will discuss alternatives to these and other specifications later in the chapter.

Note also that the intercepts for the measured math tests are all set to zero. This is a departure from what we normally do with mean structures. Less obvious, but related, the means for the latent intercept and slope variables are *not* constrained to zero. This is our first instance of a single-sample analysis with mean structures in which we have not set the latent variable means to zero. We have not done so because this is one of our primary interests in LGM: estimating the latent intercept and the latent slope. We are able to accomplish this in a single group analysis because we set the measured variable intercepts to 0 (the measured intercepts are shown above the measured variable rectangles, on the right side), essentially forcing the means to the latent variable level rather than the measured variable level.

The model allows the latent intercept and latent slope to be correlated. Intercepts and slopes are often correlated. In the present example, recall the Matthew penalty: the academically rich get richer and the poor get poorer. If this is indeed the case with math skills, then we would expect the two variables to be positively correlated (although the correlation will depend, in part, on how the slope loadings are assigned). In other areas of research it is not unusual to see the intercept and slope negatively correlated, so that those who start at a lower level catch up with those who started at a higher level.

The model also sets each of the residuals to be the same (because each is assigned a value of $v1$), a specification that says that the errors and unique variances are equal from one time point to the next. Interestingly, this is one of the assumptions that would be made, but not tested, in a repeated measures ANOVA. We can test this assumption in LGM by comparing

this model to one with these constraints freed. It is also possible to allow the residuals to be correlated with those at the next time point to allow for a “lag one autocorrelation.” It makes sense that adjacent measurements might be correlated for other reasons other than the trajectory of growth; a student, for example, may remember some of the items from the previous test administration. To review, here are the steps to specify this model:

1. Include two latent variables affecting the longitudinal data of interest, one representing the intercept, or starting level, and one representing the slope, or growth. Constrain the paths from the Intercept to each measured longitudinal variable to 1. Constrain the paths from the Slope to the measured longitudinal variables in a linear fashion. Start with a constraint of zero and increase each subsequent path by one (0, 1, 2, 3, and 4 in the present example). There are many alternatives to this specification, but this is a common method, especially within a rather restricted developmental time period.
2. Constrain the means of the residuals (errors) to zero (if this has not been done automatically).
3. Constrain the intercepts of the longitudinal measured variables to zero.
4. Freely estimate both the means and the variances of the Intercept and Slope.
5. Allow the Intercept and Slope to correlate.
6. Constrain the variances of the residuals to be equal (can be relaxed).

Let's also briefly review what this model says about the data. According to this model, the five math scores for this sample are a product of two influences: first, the initial level of math skills (intercept), and second, the growth in math skills that the children experience as a result of development and education. These two influences (plus error) are the primary source of the increase in scores over time, and of their variation and covariation with one another. The model also allows one other reason for the covariation among score: the intercept and slope are presumed to covary. The fit of the model against the data will tell us the degree to which this explanation for the scores is consistent with the data.

Figure 21.7 shows the unstandardized output for this initial unconditional model. Both the TLI (.995) and the SRMR (.006) suggest an excellent fit of the model to the data, but the

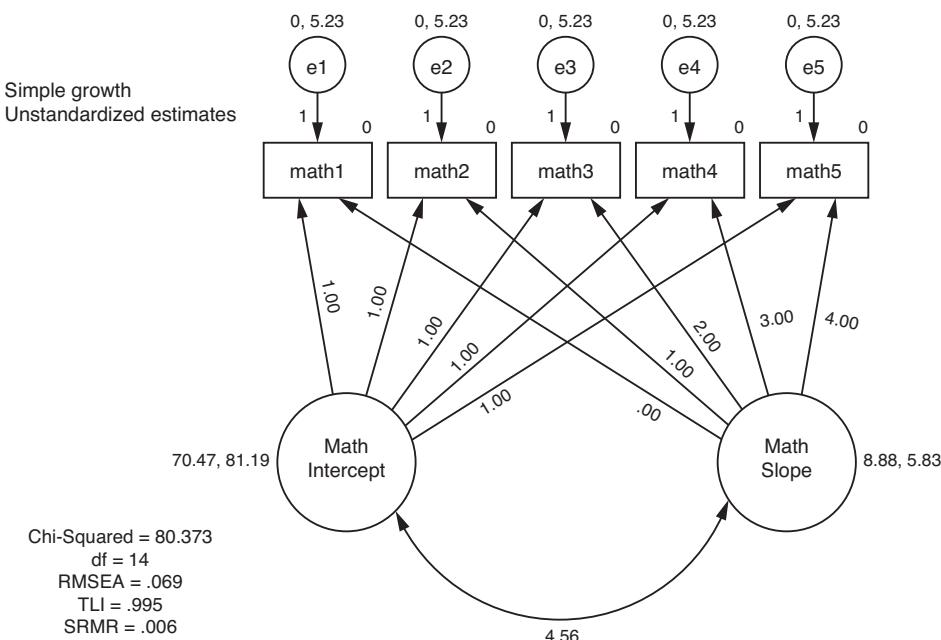


Figure 21.7 Unstandardized output, initial simple growth model

RMSEA is a little higher than we would like (.069). We know, however, that RMSEA tends to be inflated in small df models; in addition, the 90% CI for the RMSEA (.055–.084) suggests that we can reject the hypothesis on a not-close fit (because it does not encompass the value .10). Most other parameters were statistically significant and reasonable. The model has 14 df . There are 20 moments in the matrix (15 variances and covariances, 5 means), and these are used to estimate the means and variances of the Intercept and Slope, 1 error variance (1 because all were constrained to the same value), and the Intercept and Slope covariance. Degrees of freedom = $20 - (4+1+1) = 14$.

Figure 21.8 shows a revised version of this initial simple growth model. This model removed the equality constraint for the residual variances. This model also uses alternative names from the previous one: Initial Math Level rather than Math Intercept and Math Growth rather than Math Slope. This model shows improvement in fit over the initial model ($\Delta\chi^2 = 11.798$ [4], $p = .019$), but the RMSEA and the TLI both get slightly worse (because they reward parsimony). Despite the improvement in fit, I think I would likely stick with the initial model, given its elegance and parsimony. Indeed, I might not have even bothered to compare this second model given the overall good fit of the initial model.

We will thus accept the initial model (Figure 21.7) as a reasonable simple growth model and examine the output in more detail. A portion of this more detailed output is shown in Figure 21.9. This portion of the output, showing the unstandardized and standardized paths from the two latent variables to the five math measured variables, is not particularly interesting. Unlike our previous analyses, all of these values have been constrained, so there are no standard errors or significance levels. It is useful to examine the unstandardized values, however, just to make sure that all of the constraints were as they should be. You can see that the paths from the latent intercept (starting level) were all constrained to 1, and that the paths from the latent slope (growth) variable, were indeed constrained to 0, 1, 2, 3, and so on. The standardized coefficients from the level and growth variables are generally less of interest in LGM than in SEM.

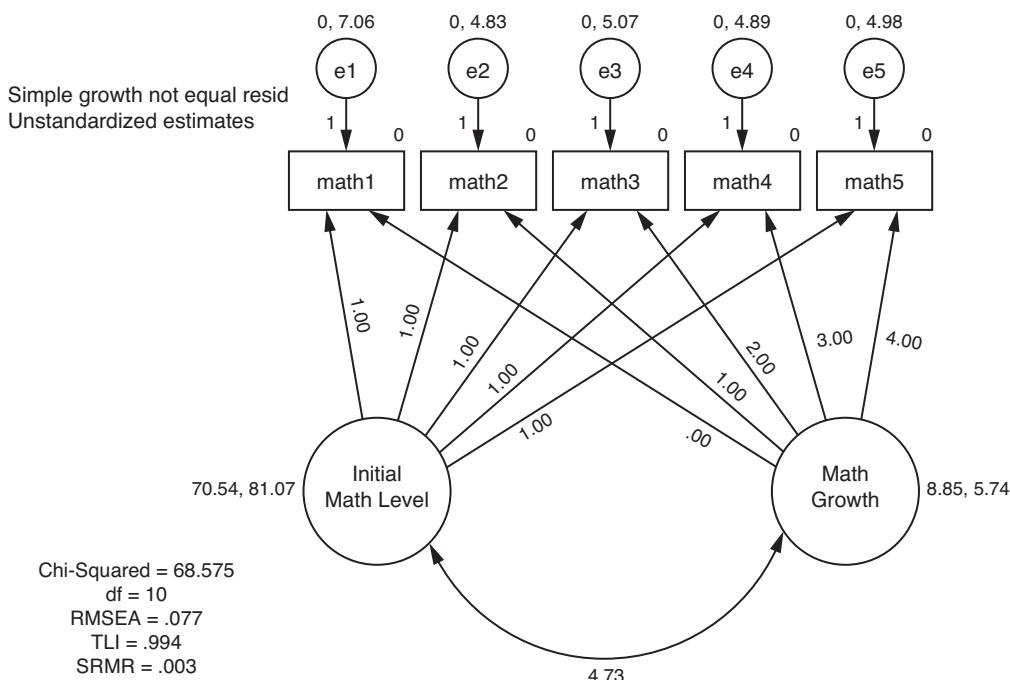


Figure 21.8 Revised simple growth model. Here, the equality constraint on the residual variances was relaxed.

Regression Weights: (equal residuals)

			Estimate	S.E.	C.R.	P	Label
math1	<---	Math_Intercept	1.000				
math1	<---	Math_Slope	.000				
math2	<---	Math_Intercept	1.000				
math2	<---	Math_Slope	1.000				
math3	<---	Math_Intercept	1.000				
math3	<---	Math_Slope	2.000				
math4	<---	Math_Intercept	1.000				
math4	<---	Math_Slope	3.000				
math5	<---	Math_Intercept	1.000				
math5	<---	Math_Slope	4.000				

Standardized Regression Weights: (equal residuals)

			Estimate
math1	<---	Math_Intercept	.969
math1	<---	Math_Slope	.000
math2	<---	Math_Intercept	.895
math2	<---	Math_Slope	.240
math3	<---	Math_Intercept	.797
math3	<---	Math_Slope	.427
math4	<---	Math_Intercept	.699
math4	<---	Math_Slope	.562
math5	<---	Math_Intercept	.613
math5	<---	Math_Slope	.657

Figure 21.9 Detailed results for the initial growth model. These results correspond to the model shown in Figure 21.7. There is not much to see here, but it is worth checking.

The primary output of interest includes the means and variances of the latent Math Intercept and Math Slope variables. These are shown in Figures 21.7 (mean, variance next to the two latent variables) and 21.10 (Figure 21.10 also includes the standard errors and z values). The baseline latent mean, or Initial Math Level, was 70.47, a value close to the average for the first math measured variable (70.14). The two are not identical because the latent Intercept takes into account errors of measurement, deviations from a linear trajectory, and so on. This value is our best estimate of the true initial average level of math achievement in these 1000 children. The variance of Initial Math Level (the intercept) is 81.19, a value that is statistically significant. This finding means that there is considerable variability in children's initial level of math skills. Presumably, some of the variables we add in the next step, variables designed to explain initial level and growth in math skills, will help explain a portion of this variation.

Figures 21.7 and 21.10 also show the mean (8.88) and the variance (5.83) for the latent Math Growth (or Slope) variable; both are statistically significant. The statistically significant value for mean Math Growth means that these children, on average, show statistically significant growth. The fact that the variance is statistically significant means that there is indeed considerable variation in the individual slopes for these children. So, for example (as in Figure 21.3), some children have fairly steep slopes (fast growth), and some children have much more shallow slopes (slower growth). There is enough variation in these slopes so that the value is statistically significant.

Means		Estimate	S.E.	C.R.	P	Label
Math_Intercept		70.467	.291	242.547	***	
Math_Slope		8.877	.080	111.354	***	

Covariances			Estimate	S.E.	C.R.	P	Label
Math_Intercept	<-->	Math_Slope	4.559	.741	6.153	***	

Correlations			Estimate
Math_Intercept	<-->	Math_Slope	.210

Variances		Estimate	S.E.	C.R.	P	Label
Math_Intercept		81.187	3.774	21.513	***	
Math_Slope		5.826	.284	20.486	***	
e1		5.228	.135	38.710	***	v1
e2		5.228	.135	38.710	***	v1
e3		5.228	.135	38.710	***	v1
e4		5.228	.135	38.710	***	v1
e5		5.228	.135	38.710	***	v1

Figure 21.10 Detailed results for the initial simple growth model, continued. These tables show the findings of primary interest.

The final interesting finding included in these figures is the covariance (4.56, $p < .001$) and correlation (.21) between the latent Math Initial Level and Math Growth variables. As expected, this positive correlation means that children with higher levels of initial math skills also improve those skills at a faster rate. The correlation is not large, but it is statistically significant.

As noted earlier, this model says that the reasons for the scores that children get on these math tests are three-fold: their scores are a result of an overall initial level of math skill, plus growth over time, plus error (including deviations from linear growth). We can use this information to calculate the expected, or model-implied, means for each of the successive math test administrations. The implied mean score for math1 is the latent Initial Math Level (70.47). The implied mean score for math 2 is the initial Math Level plus 1 times the latent slope (8.88), and the implied mean for math3 is $70.47 + 2 * 8.88$, or 88.22 (within errors of rounding). These and the values for math4 and math5 are shown in Figure 21.11 (implied means), but they can also be calculated easily from Figure 21.7. Of course, the model-implied covariances and correlations could also be calculated from the data in Figures 21.7 and 21.9–21.10 using the tracing rule.

With this simple growth model we have modeled the development, or growth, or trajectory of children's math skills from Kindergarten through fourth grade (although with simulated data). The analyses suggest that a linear growth model can indeed explain this growth, and that the children show significant growth across this time period. There is also significant variation in this developmental process across the children in this sample, both in their initial starting level of math skills and in the amount of growth they show over time. Initial level and growth in math skills are weakly but positively correlated, meaning that children with higher initial levels of math skills also show more growth, on average, than do children who start out with lower levels of initial skills.

Implied Covariances					
	math5	math4	math3	math2	math1
math5	216.108				
math4	183.016	166.206			
math3	155.152	138.940	127.956		
math2	127.288	116.902	106.517	101.359	
math1	99.423	94.864	90.305	85.746	86.415

Implied Correlations					
	math5	math4	math3	math2	math1
math5	1.000				
math4	.966	1.000			
math3	.933	.953	1.000		
math2	.860	.901	.935	1.000	
math1	.728	.792	.859	.916	1.000

Implied Means					
	math5	math4	math3	math2	math1
	105.977	97.099	88.222	79.345	70.467

Figure 21.11 Covariances, correlations, and means implied by the initial simple growth model.

CONDITIONAL GROWTH MODEL, OR EXPLAINING GROWTH

I don't know about you, but I find these analyses pretty fascinating; imagine, we can actually model the process of growth and change! But as they'd say on the infomercial, wait—there's even more! Our next step will be to add other variables to the model to see if we can understand the variables that may *influence* initial math skills and growth in math skills.

Figure 21.12 shows a conditional growth model, labeled as such because now the growth parameters are conditional on, or depend on, four possible influences. These four new variables are Female, Parent Education, Cognitive Ability, and Age; the coding of these variables is described earlier in this chapter. Each of these variables is assumed to affect these children's initial levels of math skills and their level of growth; you can probably justify each of these paths fairly easily. Thus paths are drawn from each of these exogenous variables to the latent Initial Math Level and Math Growth variables.

The model setup builds on the previous analyses. Paths are drawn from the latent Initial Level and Growth variables to each of the measured math variables. Given the results of our previous analyses, no covariances are allowed among the residuals (errors), and the residual variances are constrained to be equal. As already noted, paths are drawn from each of the exogenous possible influences to the Initial Level and Growth variables. In the previous models we allowed the Intercept (Initial Math Level) and Slope (Math Growth) variables to correlate. In the conditional model these variables are now endogenous, and thus cannot correlate directly. Instead, their disturbances are correlated, accomplishing the same thing (some SEM programs don't make this distinction between the variables and their disturbances obvious). The Cognitive and Parent Education variables are allowed to correlate, but none of the other exogenous variables are expected to do so. If the Cognitive scores were raw scores you would expect them to correlate with Age, but they are age-corrected standard scores. All other aspects of the model setup are consistent with the previous analyses.

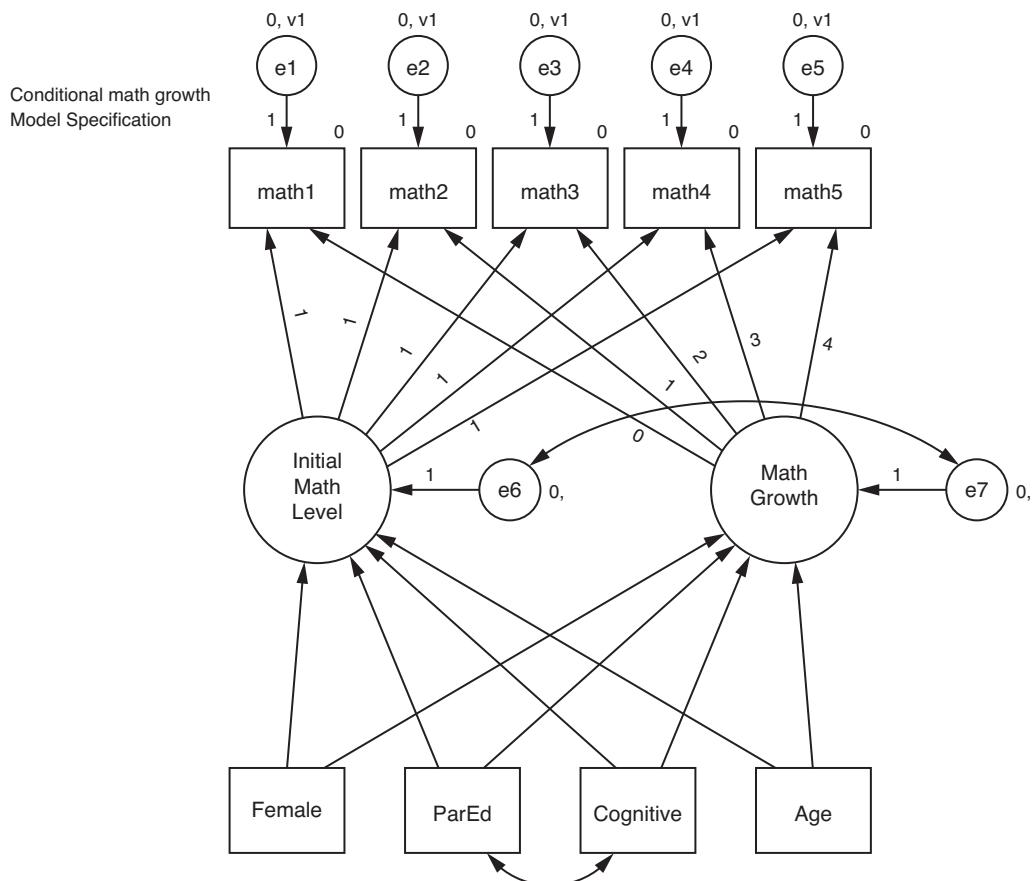


Figure 21.12 Model setup to test the influence of cognitive ability, parent education, and other background variables on math skills in Kindergarten and growth in math skills K-4.

Figure 21.13 shows the graphic output for this analysis, with the unstandardized solution in the top portion of the figure and the standardized solution in the lower portion. As shown in the Figure, all fit indexes suggest an excellent fit of the model to the data.

With the addition of the explanatory variables we now shift our attention from the aspects of the growth model, per se, to the influence of these new variables on growth. Some of these variables, such as Parent Education, are in a meaningful metric, whereas others, such as Cognitive Ability, are less meaningful. As with regression and other SEMs, the unstandardized metric is useful for variables that have a meaningful metric. For Cognitive Ability, the standardized effects (lower figure) are more interpretable. The standardized effects are also useful for comparing the relative effects of one variable with another. The standard errors and statistical significance of the effects are shown in Figure 21.14

The results shown in Figure 21.13 suggest that Female had positive effects on both Initial Math Level and on Math Growth. The Sex to Initial Level path of .77 means that girls (coded 1) scored .77 points higher on the latent Initial Math Level variable than did boys. However, these effects of Female on Initial Level and on Growth were not statistically significant (Figure 21.14), and thus should be considered as zero effects. In contrast, Parent Education had statistically significant effects on both Initial Level and Growth in math skills. For each additional year of parent education, children scored, on average, .89 points higher in

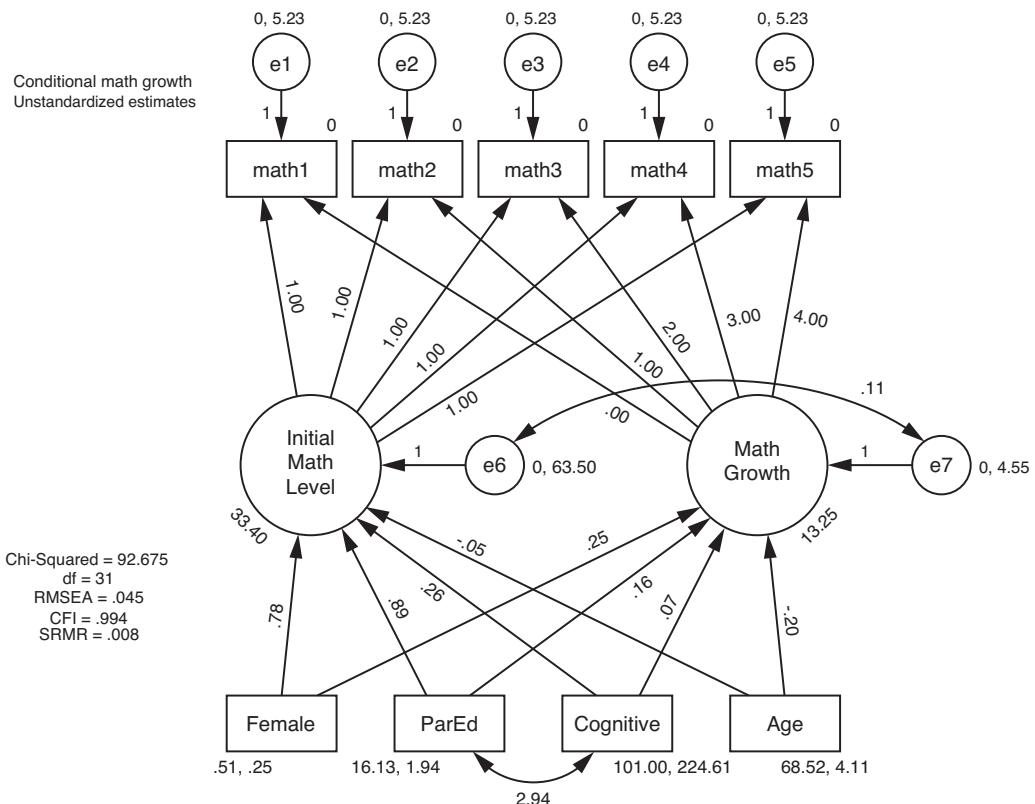
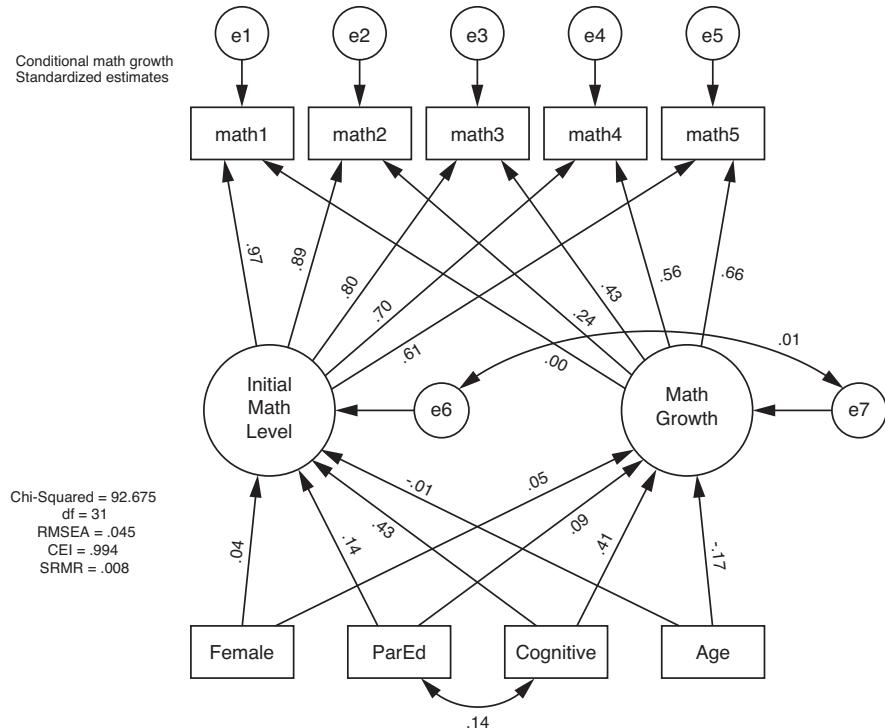


Figure 21.13 Influences of Sex, Parent Education, Cognitive Ability, and Age on initial level of math skills and growth in math skills grades K-5. The top model shows the unstandardized effects, the lower model the standardized effects

Regression Weights:

			Estimate	S.E.	C.R.	P	Label
Initial_Math_Level	<---	Sex	.777	.517	1.504	.133	
Initial_Math_Level	<---	ParEd	.887	.187	4.736	***	
Math_Growth	<---	ParEd	.160	.052	3.095	.002	
Initial_Math_Level	<---	Cognitive	.256	.017	14.702	***	
Math_Growth	<---	Cognitive	.067	.005	13.864	***	
Math_Growth	<---	Age	-.201	.035	-5.733	***	
Initial_Math_Level	<---	Age	-.051	.127	-.400	.689	
Math_Growth	<---	Sex	.246	.143	1.722	.085	
math1	<---	Initial_Math_Level	1.000				
math1	<---	Math_Growth	.000				
math2	<---	Initial_Math_Level	1.000				
math2	<---	Math_Growth	1.000				
math3	<---	Initial_Math_Level	1.000				
math3	<---	Math_Growth	2.000				
math4	<---	Initial_Math_Level	1.000				
math4	<---	Math_Growth	3.000				
math5	<---	Initial_Math_Level	1.000				
math5	<---	Math_Growth	4.000				

Figure 21.14 Effects of background variables on Initial Math Level and Math Growth (unstandardized estimates), standard errors, and statistical significance.

their initial math skills. In addition, for each additional year of parent education, children's growth in math skills increased by .16 points, after controlling for the other variables in the model (e.g., Cognitive Ability). Age had a statistically significant negative effect on Math Growth; children who were older at first measurement show less growth from one measurement to the next than were those who were younger (−.20 points per month increase in age). This finding could be a result of parents who perceive their children as not ready for Kindergarten waiting until they are older to start them in Kindergarten.

Cognitive Ability had statistically significant effects on Initial Math and on Math Growth. These effects were also quite large (see the lower portion of Figure 21.13). Children with higher levels of cognitive ability have higher initial levels of math skills; for each SD increase in cognitive ability, initial math skill increased by .43 SD . Children with higher cognitive ability also showed more growth than those with lower ability. Each SD increase in cognitive ability resulted in .41 of a SD increase in the math slope, or the growth in math skills from year to year. As shown in the standardized model, these are the largest influences on math skills and growth in math skills. (Keep in mind, of course, that these are simulated data. The findings for cognitive ability are fairly consistent with those reported by DiPerna et al., 2007, however). Parent Education had the next strongest influence on the Math intercept (Initial Level), .14, and Age was the second strongest on growth in math skills (slope), −.17.

Two other aspects of the results are worth mentioning. First, note that the covariance/correlation between the disturbances of the latent variables ($e6$ and $e7$) is considerably reduced from the previous correlation between the latent variables, and it is no longer statistically significant. The model suggests that this correlation was partially a result of Cognitive ability affecting both Initial Math Level (standardized effect .43) and Math Growth (.41). Using the tracing rule we learned when first discussing path analysis, you can see that this influence accounted for .17 of the correlation of the intercept and slope in the simple growth model. Thus cognitive ability is, to some degree, a common cause of initial math level and math growth. Or said differently, a large portion of reason that the initial level of math

skills and math growth are correlated is because both are affected by overall cognitive ability. Smarter children have higher initial levels of math and learn math more quickly and easily compared to less able children. Don't make too much of this original correlation or its reduction, however. If we had chosen a different initial level for the model (in the current model the initial level is indexed as Kindergarten, the time point with the zero path from the slope), the correlations in both the conditional and unconditional models would differ¹. The second point worth noting is that the variances of the math initial level and growth disturbances, e6 and e7, are still substantial and statistically significant (not shown here). The explanatory variables do not explain all of the variation in children's initial level of math skills or all of the variation in growth in those math skills.

Several of the paths and the correlation between disturbances were not statistically significant. Therefore, in a second model these paths (from Sex to both Initial Level and Math Growth, from Age to Initial Level) were removed, along with the correlation between e6 and e7. This revised model also showed an excellent fit to the data using our common criteria ($\chi^2 [35] = 98.36$, RMSEA = .043, TLI = .993, SRMR = .017), and the $\Delta\chi^2$ was not statistically significant ($\Delta\chi^2 [4] = 5.69$, $p = .22$). This trimmed model provides a more parsimonious explanation of the development and growth in math skills and the influences on that growth. The magnitude of the remaining influences was virtually unchanged from the initial model, so those will not be presented here (although I encourage you to conduct these additional analyses and interpret them).

ADDITIONAL ISSUES

Data Requirements

You should now have a basic understanding of LGM, how to conduct it, and how to interpret the results. Let's cover a few requirements for conducting LGM, and then we will talk about variations in setup and analysis.

Obviously, we need longitudinal data in order to conduct LGM. That is, we need repeated measures of the same individuals over time; measures of different individuals at each time point will not suffice (although for some inventive stringing together of two time-point data, see Ferrer & McArdle, 2004). Some other requirements for LGM data are:

1. At least three time-sequenced measures are needed, although four or more measures give you more degrees of freedom.
2. The measures need to measure the same construct at each time point and need to be on a metric capable of showing growth, such as raw scores. With many of our previous examples using achievement data we have used standardized scores as the metric. Such scores are available in the ECLS data ($M = 50$, $SD = 10$) but would not work because the standardization at each time point would destroy the growth aspects of the data. Thus, as noted, raw scores from an instrument that is administered repeatedly are often used in LGM. With the ECLS data, item response theory (IRT) methods were used to create a single continuous scale capable of showing growth. Even though assessments in the later grades included more advanced items than did earlier assessments, these measured the same domains as the earlier items and were placed on the same scale. The result is a short continuous measure of math skills appropriate for children in grades K through 5 in which 5th grade scores have the same meaning as K scores.
3. The time interval for longitudinal measurement needs to be the same for all participants in the study. Thus, every child in our simulated data was measured in the Fall of K, 1, 2, 3, and 4. As will be explored below, these time intervals need not be equal; that is, the research design could specify testing in Fall and Spring of Kindergarten, Spring

of first grade, Spring of third grade, and so on, as long as everyone was measured at the same intervals. There are ways of dealing with data where the time intervals are not equal, but those are beyond the scope of this introductory chapter (e.g., McArdle, Hamagami, Meredith, & Bradway, 2000; Mehta & West, 2000).

4. Raw data, as were used here, are always appropriate for LGM analysis. Matrix data will also work, if means are included; LGM is no different in this way than are other models in which latent means are estimated. The exercises at the end of the chapter will give you an opportunity to use both types of data.

Variations in Model Specifications

As noted above, the intervals from each measurement to the next need not be equal. So, for example, suppose you were interested in the growth of marijuana or other substance usage among adolescents (cf. S. C. Duncan, Duncan, Biglan, & Ary, 1998). You might administer the first and second survey one year apart and the third survey a year and a half later. If you expected growth in usage to be linear, you could set the slope (growth)-to-measured variable paths to 0, 1, and 2.5 rather than the more common 0, 1, and 2.

It is also not necessary to assume that growth is linear. Consider our current example, math achievement. Even with equally-spaced measurements, can we assume that growth in math skills is linear? It seems more likely, instead, that growth in math skills should be steeper for younger children (in Kindergarten) than for older (those in fourth grade). There are many possible ways to model non-linear growth. Recall testing for curves in regression lines; we can do something similar in LGM, by modeling one growth (slope) variable as linear growth, and another as quadratic growth. Hancock and Lawrence illustrate how to model quadratic and several other types of non-linear growth (Hancock, Harring, & Lawrence, 2013). It is also possible to constrain the first two slope loadings (e.g., to 0 and 1) and freely estimate the other slope loadings (known as a latent basis model). Researchers should consider the likely shape of growth prior to modeling, and we should always evaluate the raw data to consider possible departures from linearity (Willett & Sayer, 1994).

In the DiPerna and colleagues (2007) example referenced here, the authors were faced with different times between each assessment and the next and the possibility of nonlinear growth in math skills. They divided the sample in half and used the results from the first half to constrain the model in the second half. For the first half of the data (the calibration or training data), the values for the latent slope were constrained to 0 and 1 for the first two measured math variables. The remaining slope-to-measured variable paths were freely estimated (the latent basis model mentioned above). The values found in the calibration data (0, 1, 3.2, and 6) were used as parameter constraints in the validation (second half) data. The DiPerna et al. example also illustrates another variation in LGM: because the first slope-to-measured variable path is commonly set to zero, and this is the same as having no path at all, some researchers do not show this first measured variable as loading on the Slope variable at all.

It is possible to relax and add constraints. Here we to set the error variances of the measured variables equal at each time point (homoscedasticity), but it is possible to allow these to differ. It is also possible to allow correlations among the residuals.

There are additional constraints that can be made to the simple growth (unconditional) model to test specific aspects of that model. Some of the model constraints and their meanings include:

1. Error variances equal (discussed previously)
2. Zero correlation between the latent intercept and slope. If this model is supported, it means that the growth rate is uncorrelated with initial status.

3. Latent slope mean set to 0. If supported, this constraint would mean that the average growth was zero but that there may be variation in the amount of growth across individuals.
4. Variance of the slope variable fixed to zero. This model would suggest no individual differences in growth, that is, that everyone in the sample grew at the same pace.
5. Loadings for the latent slope variable fixed to a linear time metric (as was done in our example). This model suggests that growth is linear.
6. Mean and variance of slope both fixed to zero. This very strict “no growth” or “strict stability” model says that no one in the sample experienced growth (Stoolmiller, 1994).

Latent growth models are often more complex than those analyzed here. Here, we examined the effect of several possible influences on initial level of math skills and on growth in those skills. As already noted, it is possible to study the effects of the intercept and growth on other subsequent outcomes. It is also possible to study the relations of growth in one variable with growth in another with the inclusion of multiple intercept and slope variables. With repeated measures of cognitive ability, for example, we could study whether growth in cognitive skills was related to growth in math skills. More advanced methods, known as dynamic modeling or latent change score modeling, can test the effects, over time, of changes in variables on each other (McArdle et al., 2000; Reynolds & Turek, 2012).

It is also possible to model higher-order growth variables. Suppose for the example used here we had multiple measures of math at each grade, say measures of numbers, concepts, and geometry. It would then be possible to have a latent math skills variable at each grade rather than the measured variable included here. The intercept and slope variables could then be indexed by a series of latent variables, a method called the “curve-of-factors” approach. Alternatively, we could specify an intercept and slope for each of the math sub-constructs (numbers, concepts, geometry) and then a higher-order intercept and slope for general math skills (a factor-of-curves approach). If you are interested in reading further, the book *An Introduction to Latent Variable Growth Curve Modeling* includes much more detail (T. E. Duncan, Duncan, & Strycker, 2006). McArdle’s *Annual Review of Psychology* article does a great job of putting LGM models in the context of other longitudinal SEM models (McArdle, 2009), as does Little’s *Longitudinal Structural Equation Modeling* (2013).

OTHER METHODS OF ANALYZING GROWTH DATA

There are other methods of analyzing change and growth data such as those used here. A classic approach uses repeated measures ANOVA, or multivariate ANOVA to analyze repeated measures data. RANOVA assumes that errors of measurement (e1 through e5) are equal and independent, however, not always a reasonable assumption. As noted by Kline, the ability to model errors is a major advantage of LGM (Kline, 2016). ANOVA also generally focuses on categorical independent variables rather than the mix of categorical and continuous variables used here.

Our initial discussion of individual growth curves at the beginning of this chapter may have reminded you of the introduction (Chapter 11) to multilevel modeling (see also Chapter 22). Perhaps not surprisingly, multilevel modeling can also be used to analyze latent growth models (Singer & Willett, 2003). Consider that with LGM we are focused on understanding multiple measurements nested within individuals. This “nested within” language is exactly what multilevel modeling (MLM) is focused on. To use MLM to conduct LGM, the individuals are considered the second level of measurement, and the repeated measures are structured within individuals. Conceptually, at least, this approach is very much like what we did at the beginning of this chapter: conducting individual, time-related regressions, and pooling the results

of those regressions across individuals. One advantage of the MLM approach to LGM is that individuals need not have the same number of measurements or the same intervals between measurements, but those issues are easily overcome in SEM also. The use of SEM allows more complex (e.g., multiple related growth curves) and more flexible models, however.

SUMMARY

In previous chapters we used SEM to analyze various types of longitudinal models. Now, with the addition of latent means analysis we have expanded that focus to perhaps the most interesting longitudinal analysis so far: latent growth modeling. With LGM we are able to study and model more completely the actual process of growth and change, including the possible influences on growth and the possible effects of growth in some developmental process.

If we measured the same set of people on the same variable over time, it would be possible to conduct a regression for each person of their score on the measure on time. We could thus get a regression line of scores across time for each person. For each regression line, the intercept would represent the person's starting level on the variable, and the slope would represent his or her growth on the measure. If we then averaged the various intercepts and slopes across individuals we would be doing something conceptually similar to LGM through multilevel modeling.

To conduct LGM via SEM, we set up something that looks like a CFA, with the set of repeated measures being indexed by two latent variables, one representing the latent intercept (initial level on the repeated measures) and the other the latent slope (growth on the repeated measures). For the illustrative example used in this chapter, the repeated measures were (simulated) math test scores for a group of children, K through 4th grade. Our estimates for the mean and variance of intercept latent variable thus represented the average initial level of math knowledge for these children and the degree of variability from child to child. The mean and variance of the latent slope variable represented the average growth for the children and the degree of variation in that growth from child to child. A tabular summary of the meaning of, and alternative names for, the intercept and slope variables is shown in Table 21.2.

The model setup for LGM, while looking something like a CFA, is also a little different from what we are used to. The paths from the intercept variable to the repeated measures were all constrained to 1 (this forces this variable to be an intercept), and the paths from the latent slope variable to the repeated measures were constrained to sequential values of 0, 1, 2, and so on. The constraints on the slope loadings suggest linear growth and place the first measurement (Kindergarten) as the starting point for the growth. Finally, the intercepts for

Table 21.2 Meaning and alternative names of LGM intercept and slope variables.

<i>Latent growth variable</i>	<i>Meaning</i>	<i>Alternative names</i>
Intercept	Initial level of the construct	Initial (construct name) level Initial status Level
Slope	Growth or change in the construct	(Construct name) growth Linear growth Developmental trajectory Trend

the repeated measures (the math scores) were all set to zero. This allowed us the estimate the means for the latent intercept and slope variables even though we only have a single group (previously, multigroup analyses were needed to estimate latent means). The data required are repeated measures (generally three or more) of the same individuals on some variable capable of showing growth (that is, for example, not measures standardized within age).

Once we estimated the “simple” growth model, we were able to test for possible influences on both the initial level of math ability and on growth in math abilities. In our simulated data, we found statistically significant effects for Parent Education and Cognitive Ability on the initial level (intercept) of math and significant effects for Parent Education, Cognitive Ability, and Age on growth in math skills (latent slope). Although not done in this example, it would also be possible to examine the effects of these latent initial level and slope variables on other variables (e.g., students’ subsequent academic self-esteem). Thus, with LGM, we are able to study the process of growth, the variables that influence it, and the results of it.

Of course it is not always growth that we are interested in; sometimes it is decay, or some other developmental process. It is thus not uncommon for those doing LGM to talk of examining the “trajectory” of some developmental process. The chapter concluded with an examination of alternative methods of specifying the LGM, and the meaning of those alternatives. We also discussed other methods of analyzing growth, including repeated measures ANOVA and MLM. Of course it is possible to use some of the other methods we have explored for SEM, including multi-group and higher-order models, with LGM.

EXERCISES

1. Conduct the analyses outlined in this chapter. The data are in the file labeled “math growth final.sav.” See the website (www.tzkeith.com) for initial setup for these models for Amos and Mplus.
2. Does being the child of an alcoholic influence adolescent drinking behavior? Curran, Stice, and Chassin examined the growth in adolescents’ alcohol, along with that of their peers, over a three-year period (Curran, Stice, & Chassin, 1997). Here we will use a portion of the data to examine the effect of parents’ alcoholism and adolescents’ rebelliousness on the developmental trajectory of adolescents’ drinking behavior. The data (matrix) are in the file are in the file “curran et al alcohol.xls,” with the matrix derived from reports of 363 adolescents age 10 to 15 and their parents. Variables include self reports of students’ drinking behavior yearly for three years (Adol1 through Adol3), composite scores derived from items addressing frequency of use, frequency of excessive drinking, and frequency of getting drunk. Possible explanatory variables include parent alcoholism (Parent, 1 for yes, 0 for no), Age at time 1 (in years), and self-reported rebelliousness (Rebel1), a composite in which adolescents rated agreement with eight items concerning rule breaking a getting away with things. The file includes a Sex variable (0=girl, 1=boy), but it is not used for this exercise. See if you can set up these analyses from this description. For more information about these variables and the study (and for an illustration of a LGM with more than one set of developmental trajectories) see the original article (Curran et al., 1997).
 - a. Develop an unconditional LGM to explain the developmental trajectory among these youth in drinking behavior. Start with a model in which the error variances for Adol1 through Adol3 are constrained to be equal. Does the fit improve when you free this constraint in a second model?
 - b. You may need to constrain the error variances for Adol1 and Adol3 to zero (to avoid negative values). What happens to the fit of this model? (These constraints will not be needed for the conditional model with explanatory variables.)

- c. Interpret your final unconditional model. What do the means and variances of the Drinking intercept and Drinking slope tell you? Are these two latent variables correlated? What does that correlation tell you?
- d. Add the variables Age, Parent Alcoholism, and Rebelliousness as possible explanatory variables in a conditional LGM. Does parent alcoholism affect adolescents' drinking behavior? Does adolescent rebelliousness affect drinking? What do those effects mean? Briefly interpret any statistically significant effects. Be sure to give a real-world interpretation (one that would make sense to your grandmother).
- e. Provide a table of fits of the various models. I suggest including χ^2 and df , $\Delta\chi^2$ and Δdf , RMSEA, SRMR, CFA, and AIC, although your instructor may have different preferences.

Note

- 1 What do I mean by "If we had chosen a different initial level for the model?" Here, the paths from Growth to the Math scores were set to 0, 1, 2, etc. We could have set the first score to -1, the second to 0, the third to 1, and so on. This would have made the second measurement the initial level, and would have resulted in a different value for the correlation between the slope and the intercept. Don't overinterpret this correlation.

22

Latent Variable Interactions and Multilevel Modelling in SEM

Interactions Between Continuous Variables	534
<i>Testing Curvilinear Effects</i>	539
Multilevel Modeling in SEM	544
<i>Effects of Homework on Achievement</i>	546
<i>A Non-ML SEM Alternative</i>	557
Summary	559
Exercises	560

The purpose of this chapter is to discuss from an SEM standpoint some of the topics we first introduced from a MR orientation. In Part 2, we discussed testing for interactions between categorical and continuous variables using multigroup analysis in SEM. In this chapter, we will see how to test for interactions between continuous latent variables, thus extending the topic first presented (from a MR orientation) in Chapter 8. In Chapter 11, I briefly presented multilevel modeling from a regression orientation. Here, we will see how to analyze these models in an SEM context, with both measured and latent variables.

Like Chapter 11 in Part 1, this chapter does not go into the depth of the others in Part 2. My main purpose is to alert you to the presence of these fairly advanced topics and to illustrate the kind of results you might find. In addition, this chapter is also more program specific than the rest of Part 2. In most of this part of the book I have endeavored to make the directions and output generalizable to any SEM program (although I have used Amos to draw figures and illustrate output, and the web page includes Mplus illustrations). In this chapter I will still use Amos to draw many of the models, but all analyses were conducted with Mplus. Mplus has made such analyses more widely and easily available, and thus is a good choice for illustration. I will also discuss resources for using other programs to analyze latent variable interaction analyses and multilevel modeling.

INTERACTIONS BETWEEN CONTINUOUS VARIABLES

In Chapter 8, we discussed and illustrated the use of cross products to test for interactions between two continuous variables in multiple regression. We can use the same method to test for interactions in path models (measured variable SEMs): by including a cross product of the two variables that interact in the path model. But how, you may wonder, can such testing be done when the variables that we are interested in are latent variables?

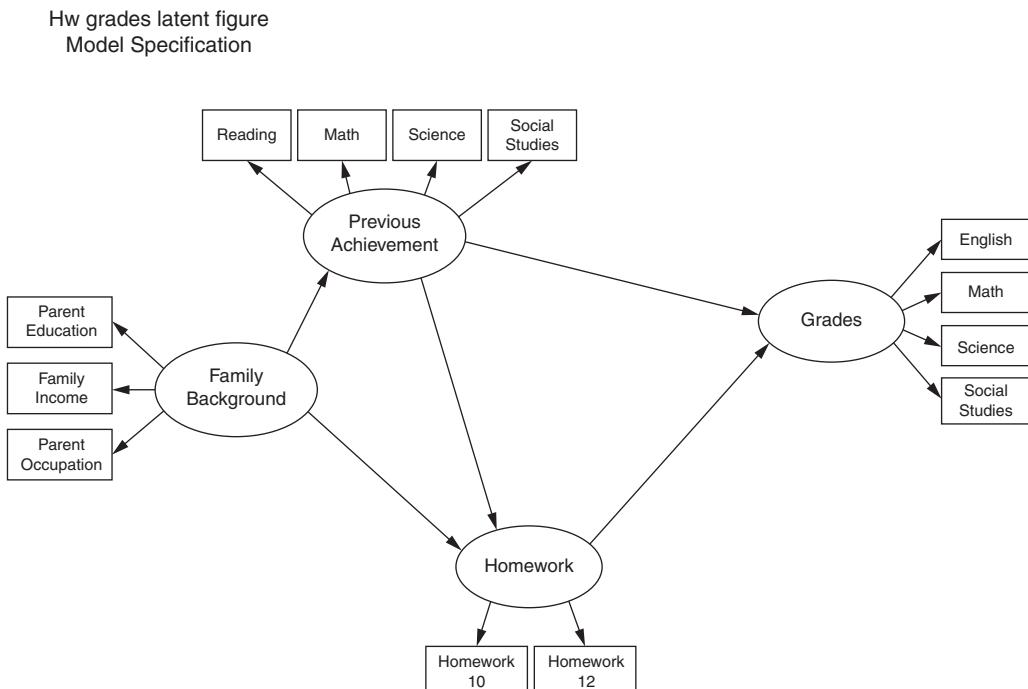


Figure 22.1 Model testing the effect of out-of-school Homework on high school grades.

We saw in Chapters 18 and 19 how to use multigroup analysis to test for interactions in SEM when one of the variables of interest is a categorical or grouping variable. But that method does not generalize well to the situation in which both variables are continuous latent variables, because such grouping variables are generally measured (see, however, latent mixture modeling for an exception), and because it is generally poor practice to turn a continuous variable into categorical one.

The method used in MR—creating cross products and using those in the analysis—continues to be a viable approach in latent variable SEM. The difference is that in latent variable SEM these cross products are used as indicators of a *latent interaction* term. I will illustrate this conceptually first, and then through analysis.

Figure 22.1 shows a variation of the Homework model we have used off and on in this book. Errors, disturbances, and other details are not included to make it clear that this is primarily a conceptual model. In this model, time spent on homework is assumed to influence students' academic performance in high school. The model shown is tied to data from 8th through 12th grade from the NELS dataset. The grade measured variables are transcript final GPAs in English, math, science, and social studies in high school on a scale that ranged from a low of zero to a high of 12. The homework measured variables are student reports of the average amount of time spent on homework per week, out of school. Family Background and Previous Achievement (8th grade test scores) are controlled.

Suppose that you wondered about possible differential effects of homework depending on students' prior levels of achievement. You wonder, perhaps, if time spent working on homework might be more effective for students with lower levels of achievement. We have seen (Chapter 8) that homework time appears to have diminishing returns on its effect on grades, with additional hours showing smaller effects at some point. Perhaps homework also has diminishing returns based on prior achievement, with homework being more effective for

Hw grades interaction figure
Model Specification

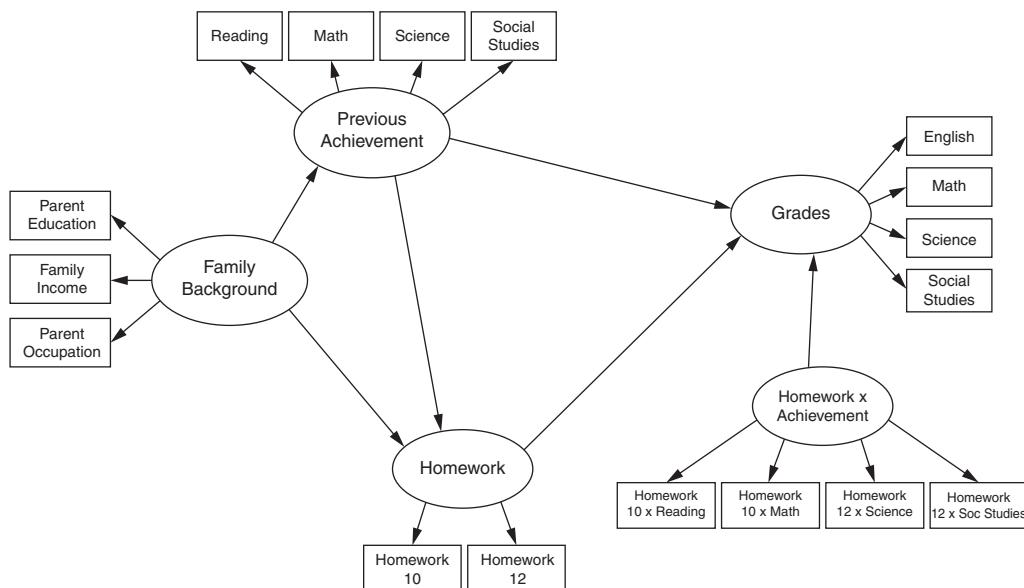


Figure 22.2 Conceptual model with the addition of latent variable testing the possible interaction of Previous Achievement and Homework in their effect on Grades.

low-achieving as opposed to already high-achieving students. Here we are speculating that the effects of homework on subsequent grades depend on the students' prior achievement; we are using the "it depends" lingo of interactions (moderation).

The model shown in Figure 22.2 includes an interaction term as a latent variable, with cross products as indicators of that latent variable. This is the general idea behind "product indicator" approaches to latent interaction analysis (Marsh, Wen, Hau, & Nagengast, 2013; I will adopt the terminology from this chapter here), of which there are several variations. These approaches differ on the number and type of measured cross products used, the constraints required by the model, how the latent interaction term relates to the other latent variables, and various other details. The original "constrained approach" developed by Kenny and Judd (1984), for example, uses all possible cross products of the measured indicators for the two variables tested for interaction. Thus, in this example, the latent interaction term would include eight indicators, including Reading \times Homework10, Math \times Homework10, and so on. The method also requires considerable non-linear parameter constraints. The unconstrained approach (Marsh, Wen, & Hau, 2004) uses fewer cross products, does not require the non-linear constraints of the constrained approach, and allows correlations between the latent interaction term and the constituent latent variables. This approach works best when the two variables thought to interact have the same number of indicators; Marsh and colleagues present several alternatives when this is not the case (2013). A third variation uses residualized cross-product variables as indicators (Little, Bovaird, & Widaman, 2006). In this approach, each of the measured cross products is regressed on the other indicators, with the residuals from these analyses used as the indicators of the latent interaction variable. There are other variations of the product-indicator approach, and this continues to be an area of research. Note that all models also test the effects on the outcome of the variables thought to interact. For this example, Previous Achievement and Homework also affect Grades.

```

ANALYSIS:
  TYPE=RANDOM;
  ALGORITHM=INTEGRATION;

MODEL:
  FAMBACK BY
    PAROCC
    BYPARED
    BYFAMINC;

  PREVACH BY
    BYTXRSTD
    BYTXMSTD
    BYTXSSTD
    BYTXHSTD;

  HW BY
    HW_10
    HW_12;

  GRADES BY
    ENG_12
    MATH_12
    SCI_12
    SS_12;

  BYTXRSTD    WITH ENG_12;
  BYTXMSTD    WITH MATH_12;
  BYTXSSTD    WITH SCI_12;
  BYTXHSTD    WITH SS_12;

  PREVxHW | PREVACH XWITH HW;

  PREVACH ON FAMBACK;
  HW ON PREVACH FAMBACK;
  GRADES ON PREVACH HW PREVxHW;

```

Figure 22.3 Mplus syntax to test for a possible interaction between Homework and Previous Achievement in their effects on Grades.

The Mplus program implements an approach that is considerably easier for the applied researcher. This approach is classified as a “distribution analytic” approach by Marsh and colleagues (2013), and is based on the latent moderated structural equations approach of Klein and Moosbrugger (2000). “Distribution-analytic approaches estimate latent nonlinear effects by directly modeling the nonnormality in the indicators of the outcome variables that is implied by latent interaction effects” (Marsh et al., 2013, p. 290). For more detail, see Muthén and Asparouhov (2015), available on the Mplus website (www.statmodel.com).

What makes this approach easier is that it is not necessary to create cross-products and use them as indicators of latent variables. Instead, an “XWITH” command is added to the Mplus syntax to specify the latent variables that are being tested for interaction and to name that interaction. The outcome latent variable is regressed on the normal variables in the model plus interaction term (there are also two changes needed to the ANALYSIS command). Figure 22.3 shows the Mplus ANALYSIS and MODEL commands needed to test for the interaction of Previous Achievement and Homework used as an example here, with the unusual portions of the commands highlighted.

Figure 22.4 shows a portion of the unstandardized output from the analysis. It shows the unstandardized factor loadings and unstandardized paths. Note the highlighted line that shows that the effect of the latent interaction is not statistically significant. Figure 22.5 shows

MODEL RESULTS

				Two-Tailed P-Value
		Estimate	S.E.	Est./S.E.
FAMBACK	BY			
PAROCC		1.000	0.000	999.000
BYPARED		0.069	0.003	22.067
BYFAMINC		0.124	0.006	21.763
PREVACH	BY			
BYTXRSTD		1.000	0.000	999.000
BYTXMSTD		0.976	0.028	34.829
BYTXSSTD		0.950	0.031	30.908
BYTXHSTD		0.931	0.028	33.100
HW	BY			
HW_10		1.000	0.000	999.000
HW_12		1.004	0.105	9.561
GRADES	BY			
ENG_12		1.000	0.000	999.000
MATH_12		0.875	0.023	37.663
SCI_12		0.961	0.022	44.127
SS_12		1.049	0.021	50.540
PREVACH	ON			
FAMBACK		0.327	0.022	15.185
HW	ON			
PREVACH		0.045	0.008	5.670
FAMBACK		0.018	0.005	4.002
GRADES	ON			
PREVACH		0.138	0.012	11.655
HW		0.673	0.128	5.278
PREVXHW		-0.004	0.007	-0.506
				0.613

Figure 22.4 Edited Mplus output (unstandardized coefficients) from the interaction analysis.

				Two-Tailed P-Value
		Estimate	S.E.	Est./S.E.
FAMBACK	BY			
PAROCC		0.710	0.020	35.951
BYPARED		0.834	0.018	47.186
BYFAMINC		0.728	0.021	34.803
PREVACH	BY			
BYTXRSTD		0.855	0.011	78.047
BYTXMSTD		0.849	0.010	82.847
BYTXSSTD		0.818	0.012	66.243
BYTXHSTD		0.828	0.012	69.031
HW	BY			
HW_10		0.688	0.042	16.463
HW_12		0.591	0.037	15.806
GRADES	BY			
ENG_12		0.915	0.007	124.317
MATH_12		0.812	0.013	62.050
SCI_12		0.884	0.009	97.651
SS_12		0.909	0.008	109.291
PREVACH	ON			
FAMBACK		0.578	0.028	20.473
HW	ON			
PREVACH		0.332	0.054	6.109
FAMBACK		0.237	0.059	4.009
GRADES	ON			
PREVACH		0.486	0.040	12.117
HW		0.321	0.051	6.303
PREVXHW		-0.015	0.029	-0.508
				0.611

Figure 22.5 Standardized output from the interaction analysis.

Hw ach interaction output

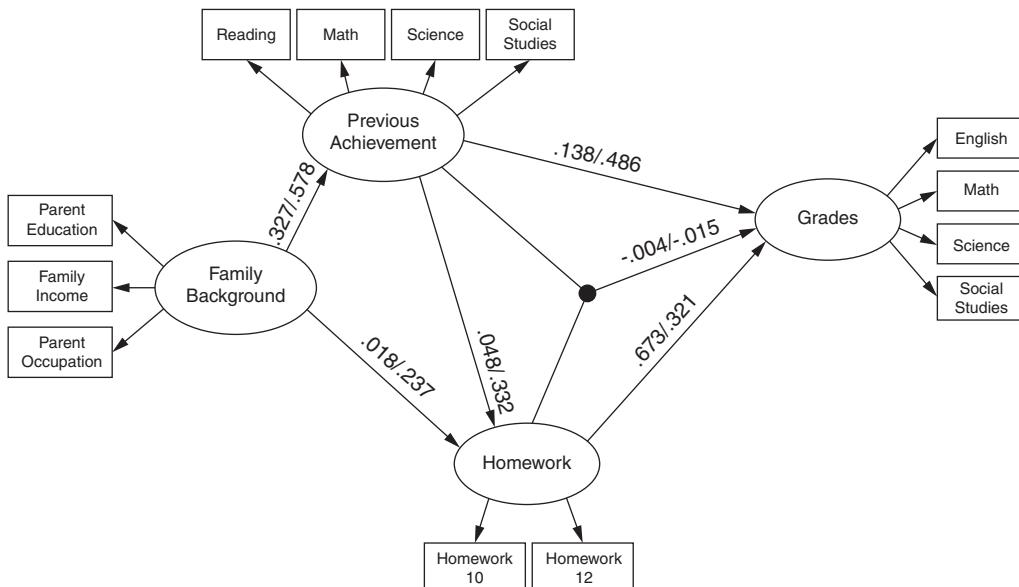


Figure 22.6 Unstandardized and standardized paths from the test of whether previous achievement moderates the effect of homework on grades. The solid dot is used to signify the interaction between these two latent variables.

the corresponding standardized output. Again, the effect of the latent interaction is small and not statistically significant. Figure 22.6 shows the unstandardized and standardized paths (unstandardized/standardized) from the output in figural form. This is how the Mplus manual illustrates interactions: lines leading to a solid dot that represents the interaction of the two latent variables. (The display in the Mplus diagrammer is not the same.) Note that I did not include a lot of the detail that could have been included, such as disturbances, correlated errors, factor loadings, etc.

These findings suggest that my speculation concerning possible interaction between Previous Achievement and Homework was not correct; homework does not appear to have differential effects on subsequent grades depending on students' previous levels of achievement. This should not be surprising if you recall our initial discussion of testing for interactions in multiple regression in Part 1 of this text. There I argued that interactions are relatively rare in nonexperimental research, and that we should usually only test for these if we have theoretical or research support for doing so. Here, I just made up my speculation about a possible Achievement-by-Homework interaction, and it appears that speculation was unfounded.

Testing Curvilinear Effects

These findings suggest that Previous Achievement and Homework do not interact in their effect on subsequent Grades. But recall that we *did* find evidence (Chapter 8) for a curvilinear effect for Homework on Grades. Recall also that in multiple regression we tested for a curve in the regression plane by including in the analysis a squared version of the variable of interest (e.g., Homework^2). We likened this approach to testing whether a variable interacts with itself in its effect on an outcome of interest. The same process generalizes to latent variable models. We could use this approach in a product-indicator model, and we can use this

```

TITLE: Homework Achievement Latent curve in regression line?

ANALYSIS:
  TYPE=RANDOM;
  ALGORITHM=INTEGRATION;

MODEL:

  FAMBACK BY
    PAROCC
    BYPARED
    BYFAMINC;

  PREVACH BY
    BYTXRSTD
    BYTXMSTD
    BYTXSSTD
    BYTXHSTD;

  HW BY
    HW_10
    HW_12;

  GRADES BY
    ENG_12
    MATH_12
    SCI_12
    SS_12;

  BYTXRSTD      WITH ENG_12;
  BYTXMSTD      WITH MATH_12;
  BYTXSSTD      WITH SCI_12;
  BYTXHSTD      WITH SS_12;

  HW_Sq | HW XWITH HW;

  PREVACH ON FAMBACK;
  GRADES ON PREVACH HW HW_Sq;
  HW ON PREVACH FAMBACK;

OUTPUT:  SAMPSTAT STDYX RESIDUAL TECH1 TECH8;

```

Figure 22.7 Mplus syntax (edited for display) to test for a curvilinear effect of homework on grades.

approach in a distribution-analytic approach. I will illustrate the latter with an Mplus analysis of a possible curvilinear effect for Homework on Grades.

Figure 22.7 shows a portion of the input file for Mplus to test for a non-linear (quadratic) effect for the latent Homework variable on Grades (the complete input file and the data are on www.tzkeith.com). Figures 22.8 and 22.9 show the unstandardized and standardized output from the analysis. As shown by the highlighted line in these figures, the Homework² variable (HW_SQ) had a statistically significant effect on Grades (z -test = 4.047, $p < .001$). Figure 22.10 shows the output (standardized) from the Mplus diagrammer.

You may wonder why I have not referred to the fit statistics to evaluate these models. The primary reason is that given the estimation method used Mplus produces only a very restricted set of fit statistics; these are shown in Figure 22.11. I think it is possible to augment these in several ways. First, we can estimate a model without Homework² (or, in the earlier example, the Previous Achievement-by-Homework interaction term) using maximum likelihood estimation to make sure the model provides a reasonable fit to the data. The fit statistics for this model are shown in Figure 22.12; yes, it appears the basic Homework model, prior to inclusion of a Homework² variable, fits the data well (e.g., RMSEA = .032, SRMR = .023).

MODEL RESULTS (Unstandardized)

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FAMBACK	BY				
PAROCC		1.000	0.000	999.000	999.000
BYPARED		0.069	0.003	22.117	0.000
BYFAMINC		0.124	0.006	21.762	0.000
PREVACH	BY				
BYTXRSTD		1.000	0.000	999.000	999.000
BYTXMSTD		0.988	0.028	35.885	0.000
BYTXSSTD		0.960	0.029	32.713	0.000
BYTXHSTD		0.943	0.028	34.204	0.000
HW	BY				
HW_10		1.000	0.000	999.000	999.000
HW_12		0.953	0.130	7.314	0.000
GRADES	BY				
ENG_12		1.000	0.000	999.000	999.000
MATH_12		0.876	0.023	37.732	0.000
SCI_12		0.961	0.022	44.256	0.000
SS_12		1.050	0.021	50.437	0.000
PREVACH	ON				
FAMBACK		0.325	0.021	15.396	0.000
GRADES	ON				
PREVACH		0.137	0.012	11.716	0.000
HW		0.771	0.130	5.949	0.000
HW_SQ		-0.181	0.045	-4.047	0.000
HW	ON				
PREVACH		0.044	0.008	5.268	0.000
FAMBACK		0.019	0.005	4.178	0.000

Figure 22.8 Unstandardized results, test for a curvilinear effect for Homework on Grades.

With the addition of the Homework² variable to the model, its statistical significance is shown by the lines already highlighted in Figures 22.8 and 22.9. Another possible method would be to rerun the input shown in Figure 22.7 but to constrain the Homework² to Grades path to zero (GRADES ON HW_SQ@0). It would then be possible to use the AIC (or BIC or aBIC) values to compare these two models. The model with a path from Homework² to Grades produced an aBIC value of 64378.291, versus an abIC value of 64388.790 when this path was constrained to zero. Given that smaller AIC and aBIC values are preferred, this approach would have favored the model allowing a quadratic (HW_SQ) effect. It would also be possible to estimate a $\Delta\chi^2$ from the log-likelihood values, but I am not sure how appropriate this is.

Recall also from Chapters 7 and 8 that my advice was to graph interactions and curves to better understand their nature. It is also possible to examine the sign of the coefficients. The linear aspect of the Homework effect was positive, whereas the Homework² path was negative. According to the information presented in Table 8.1, the resulting regression line should thus be sloping upward with a convex shape.

There are several possible ways to graph the regression line. One possibility is to output factor scores and use these to produce a scatterplot in a program like SPSS (see Caemmerer, Maddocks, Keith, & Reynolds, 2018, for an example of this approach). For this procedure you should use the measurement model only, that is, a CFA of the latent variables used in the SEM models; this CFA model does not include the Homework² variable. This has been done in Figure 22.13. The figure shows a quadratic regression line produced in SPSS, with

STDX Standardization

		Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
FAMBACK	BY PAROCC	0.710	0.020	35.928	0.000
	BYPARED	0.833	0.018	47.273	0.000
	BYFAMINC	0.729	0.021	34.922	0.000
PREVACH	BY BYTXRSTD	0.852	0.011	79.185	0.000
	BYTXMSTD	0.852	0.010	81.492	0.000
	BYTXSSTD	0.819	0.012	67.191	0.000
	BYTXHSTD	0.830	0.012	68.561	0.000
HW	BY HW_10	0.714	0.055	13.015	0.000
	HW_12	0.582	0.043	13.617	0.000
GRADES	BY ENG_12	0.918	0.007	126.070	0.000
	MATH_12	0.817	0.013	63.343	0.000
	SCI_12	0.887	0.009	99.226	0.000
	SS_12	0.912	0.008	111.206	0.000
PREVACH	ON FAMBACK	0.579	0.028	20.562	0.000
GRADES	ON PREVACH	0.472	0.038	12.325	0.000
	HW	0.376	0.049	7.674	0.000
	HW_SQ	-0.107	0.027	-3.947	0.000
HW	ON PREVACH	0.307	0.057	5.415	0.000
	FAMBACK	0.241	0.058	4.172	0.000

Figure 22.9 Standardized results, effects of Homework and Homework² on Grades.

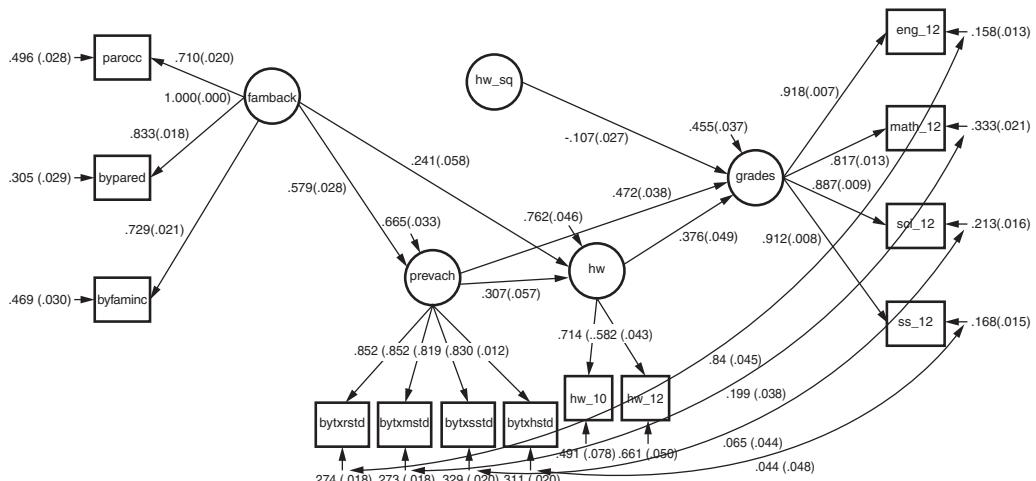


Figure 22.10 Mplus diagrammer output, testing for a curvilinear homework effect on grades, standardized coefficients

MODEL FIT INFORMATION	
Number of Free Parameters	49
Loglikelihood	
H0 Value	-32097.719
H0 Scaling Correction Factor for MLR	1.0282
Information Criteria	
Akaike (AIC)	64293.438
Bayesian (BIC)	64533.918
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	64378.291

Figure 22.11 Fit information for the Homework curvilinear example. Note that only the log-likelihood and information criteria are supplied for this type of analysis.

MODEL FIT INFORMATION	
Number of Free Parameters	48
Loglikelihood	
H0 Value	-32104.831
H1 Value	-32048.152
Information Criteria	
Akaike (AIC)	64305.662
Bayesian (BIC)	64541.235
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	64388.784
Chi-Square Test of Model Fit	
Value	113.358
Degrees of Freedom	56
P-Value	0.0000
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.032
90 Percent C.I.	0.023 0.040
Probability RMSEA <= .05	1.000
CFI/TLI	
CFI	0.993
TLI	0.990
Chi-Square Test of Model Fit for the Baseline Model	
Value	7792.886
Degrees of Freedom	78
P-Value	0.0000
SRMR (Standardized Root Mean Square Residual)	
Value	0.023

Figure 22.12 Model fit without Homework² in the model.

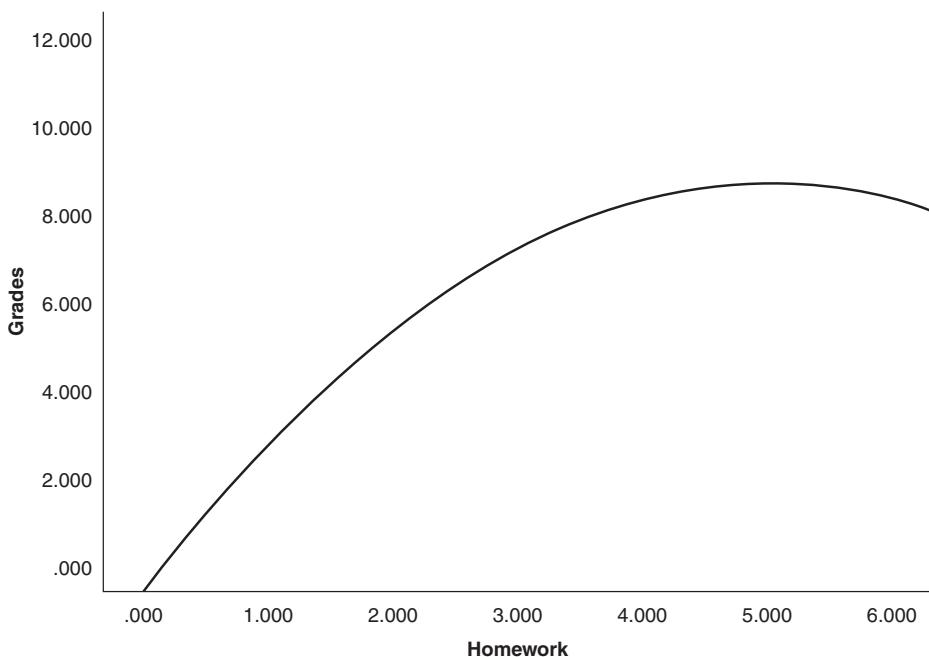


Figure 22.13 Regression line for Grades on Homework, allowing for a single curve in that line (a quadratic effect). The graph is from a scatterplot of factor scores output by Mplus.

the data points removed. Because the factor scores (like the latent variables) have means of zero, I added in the means associated with HW_10 and ENG_12 to the Homework and Grades factor scores, respectively (because these were the reference variables in the analyses). Note that you can do this either in the general statistical program (e.g., SPSS) or in Mplus by constraining the means of the latent variables to the appropriate values. For the factor scores shown in Figure 22.13, I constrained the means for the Homework and Grades variables to 2.526 and 6.123, respectively. The quadratic regression line nicely illustrates the nature of the curvilinear effect of homework. For students who do relatively little homework, an increase in time spent should likely produce noticeable improvements in grades, but when students already complete a lot of homework, increases will produce little effect in grades.

Another method for graphing the effect would be to use the regression equation derived from the Mplus output: $\text{Grades} = .137\text{PrevAch} + .771\text{Homework} - .181\text{Homework}^2$. When I did this using Excel with the addition of an appropriate constant to put the graph back into the original scale, the line looked very similar to that shown in Figure 22.13. It is also possible to produce plots in Mplus, but these are not as flexible as the ones produced here in SPSS or those produced in Excel.

I hope this section has given you a brief but understandable introduction to how it is possible to test for interactions and quadratic effects for latent variables. It is certainly not exhaustive, and I have illustrated only a few of the possible methods that have been proposed. For more depth, see the Marsh and colleagues reference already mentioned (Marsh et al., 2013). Also see Klein and Moosbrugger (2000), Kline (2016), Muthén and Asparouhov (2015), and Schumacker and Marcoulides (1998). Input and output for different models are shown on <http://tzkeith.com>.

MULTILEVEL MODELING IN SEM

In the second half of Chapter 11 we briefly explored the topic of multilevel modeling from a multiple regression orientation. Here, we will explore the same topic from a structural equation modeling orientation. It may be useful to review that section before reading this

one. Like the previous discussion, this one will be relatively brief, with a focus on explaining a single illustration, and with a focus on assimilating that illustration into our current SEM mental framework. I hope you will find that the visual and conceptual nature of SEM will help you understand multilevel SEM (MLSEM). Keep in mind, however, that we are just scratching the surface of this topic, with lots of issues unexplored. I'll also caution you that I am a novice on this topic myself, so I will no doubt miss some of the finer points of MLSEM.

It would be possible, of course, to analyze the example used to illustrate MLM in Chapter 11 using SEM software. I will not do so here, but see the website (<http://tzkeith.com>). It is worth illustrating that model using our SEM/figural displays, however. A path model version of such an analysis is illustrated in Figure 22.14. Note that I have not included disturbances in the model. Recall for this analysis that we found that there were both school-level (between-school) and individual-level (within school) effects for SES on Achievement. There were also differences in the intercepts of the school-level regressions (Achievement on SES), but the within-school slopes did not vary as a result of school-level SES (we tested for but did not find an interaction of school-level and individual-level SES). This information is conveyed in the figure. First, the figure shows two path models, one representing the individual-level (within-school) effects of SES on Achievement (lower half), and one representing the effects of school-level SES on school-level Achievement (between-school, or between model, upper half). Second, the solid dot at the end of the path from SES to Achievement for the within model signifies that there were differences in intercepts across the schools. Finally, the solid circle in the middle of the SES to Achievement path, designated by *s*, signifies that we tested for random slopes across schools.

This is one method of conveying such MLSEM analyses, and is similar to the type of display used in the Mplus manual (although Achievement would likely be symbolized as a latent variable at the between level, given the random intercept). This method of display is fairly common, although you will see variations of it.

Figure 22.15 shows an alternative figural display for this model (similar to Stapleton, 2013). This display shows that each measured variable is used to estimate both a within-school and a between-school variable. These variables are symbolized as latent because they are not measured, but are, instead, estimated by the measured SES and Achievement variables. These latent variables are also shown with a lighter line for the ovals, to suggest that

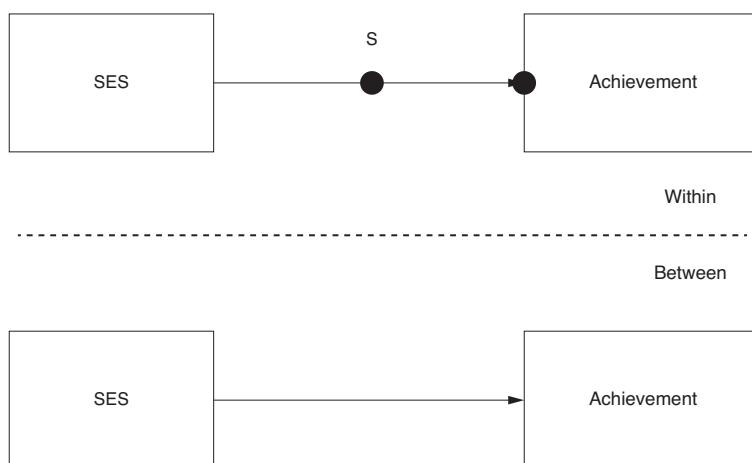


Figure 22.14 Illustration of the MLM model from Chapter 11 in SEM (Mplus) format, variation one.

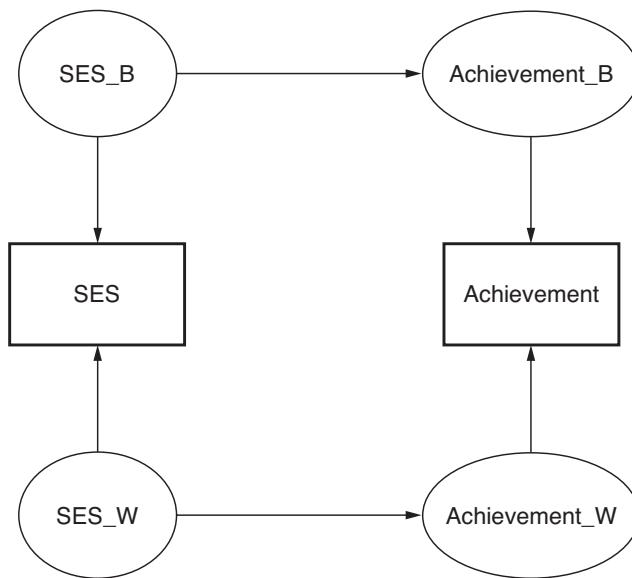


Figure 22.15 Illustration of the MLM example from Chapter 11 in SEM (Stapleton) format, second variation.

these latent variables do not actually appear in the analysis, but are a good way of thinking about the analysis. At each level, SES is assumed to affect Achievement. You may wonder how this model could work without being badly underidentified. First, consider that this is a conceptual model. Second, consider that an actual MLM model would be estimated by the individual-level covariance matrix (for the within portion of the model) and the between-school covariance model (for the between model). This type of display appeals to my latent variable orientation.

Effects of Homework on Achievement

Let's illustrate, briefly, a latent variable ML analysis. The example is fairly simple, with a single latent Homework variable affecting a latent Achievement variable. An SES composite is also controlled. The analysis is conducted in Mplus, and I use the steps recommended by Stapleton (2013), although I do not present the analysis in as much detail as she does. See her chapter for a more in-depth presentation, and see the website (<http://tzkeith.com>) for actual Mplus input and output.

Figure 22.16 displays the conceptual model underlying the analysis. The variables are ones you have seen before. SES is a composite of parent education, parent occupational status, and family income. The homework latent variable is indexed by student reports of out-of-school homework in 10th and 12th grades. The achievement scores are reading, math, science, and social studies test scores in 12th grade. Given that schools often have different homework policies, or at least homework cultures, it seems reasonable to seek to assess the effects of homework at both the individual (within-school) level and at the between-school level. The heavier-lined variables—both measured and latent—actually appear in the analyses. The lighter-gray lined variables do not appear in the input or output for the analysis, but are included in the figure as a heuristic device. Note that I have not included a measure of previous achievement in the model, even though all our previous analyses suggest previous achievement or ability is a likely common cause of homework and current achievement. I'm trying to keep the example fairly simple.

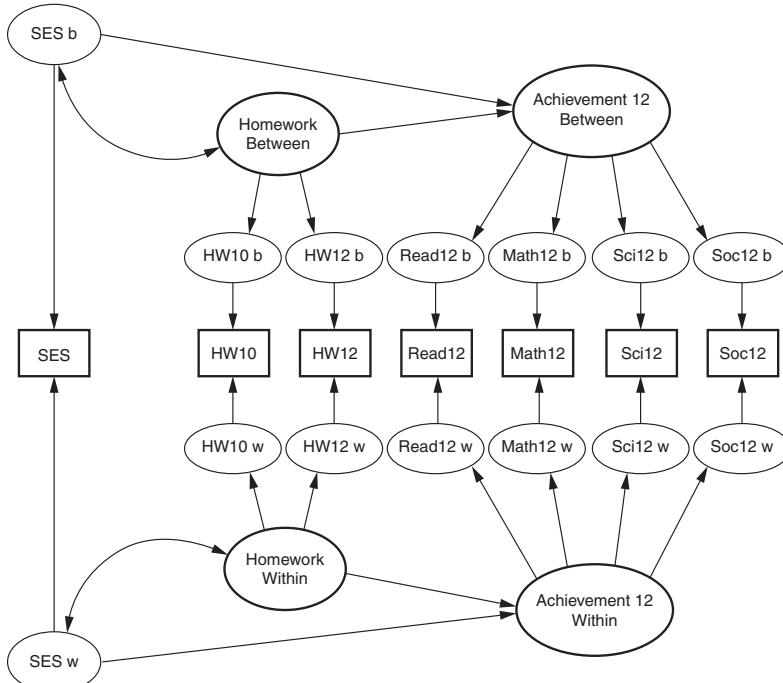


Figure 22.16 Latent multilevel homework-achievement conceptual model.

Steps in the analysis include:

- Step 1. Evaluate descriptive information of all variables.
- Step 2. Run baseline models for both the within- and between-cluster levels.
- Step 3. Run a theoretical model at the within level, saturated at the between level.
- Step 4. Run a theoretical model at the between level, saturated at the within level.
- Step 5. Run a model with theory imposed at both levels.
- Step 6. Evaluate random coefficients at the within level (Stapleton, 2013, pp. 537–538).

Step 1: Descriptive Information

Data were selected from the NELS base year through second follow-up (8th through 12th grades); I selected schools with 20 or more students. Data for 4020 students from 175 schools, with an average of 23 students sampled per school, are contained in the file “homework smaller for mplus conversion 4.sav” for SPSS and “homework smaller 4.dat” for use in Mplus. Basic descriptive information calculated via SPSS is displayed in Table 22.1: means, standard deviations, minimum, maximum, skew, and kurtosis. At this step, Stapleton also recommended analyzing a “null” or ‘unconditional’ model for each variable (2013, p. 538). For this analysis, the variance of each variable is partitioned into within- and between-cluster components based on the clustering variable (in this case the school ID variable). This analysis will produce the interclass correlation (ICC), the ratio of the between cluster to the total (between + within) variance. Recall from Chapter 11 that large interclass correlations suggest a lot of between-cluster variance and the need for multilevel analysis. The Mplus syntax and edited output for the SES variable is shown in Figure 22.17. Note that you can also calculate the ICC from the variances highlighted for within and between in the Mplus output. The interclass correlations for all variables are also shown in Table 22.1.

```

TITLE: Homework SEM Stapleton step 1a Famback

DATA:
FILE IS homework smaller 4.dat;

VARIABLE:
MISSING ARE ALL (-99);
NAMES ARE sch_id bys81a bys81b bys81c bys81d f1s36a2 f2s25f2
read12 math12 sci12 soc12 FamBack;

USEVARIABLES ARE
sch_id FamBack;

CLUSTER = sch_id;

ANALYSIS: TYPE = TWOLEVEL;
MODEL:
%WITHIN%
FamBack;
%BETWEEN%
FamBack;

OUTPUT: STANDARDIZED SAMPSTAT;

```

Variable	Intraclass Correlation
FAMBACK	0.404

MODEL RESULTS

	Estimate	S.E.	Two-Tailed P-Value	
Within Level				
Variances				
FAMBACK	0.399	0.012	32.955	0.000
Between Level				
Means				
FAMBACK	-0.031	0.041	-0.761	0.446
Variances				
FAMBACK	0.271	0.035	7.690	0.000

Figure 22.17 Step 1, Mplus input and edited output for the FamBack (SES) variable.

Table 22.1 Descriptive Information for the MLSEM Example (Step 1)

Variable Name	Variable Name (SPSS/Mplus)	N	Minimum	Maximum	M	SD	Skew	Kurtosis	Interclass correlation
SES	FamBack	4018	-1.95	2.67	-0.01	.83	.51	.35	.404
HW10	f1s36a2	3857	0	7	2.66	1.78	.75	-.14	.128
HW12	f2s25f2	3695	0	8	3.51	2.03	.46	-.42	.105
Read12	f22xrstd/read12	3434	29.29	68.09	52.22	9.67	-.42	-.83	.166
Math12	f22xmstd/math12	3433	30.14	71.37	52.91	9.85	-.24	-.92	.222
Sci12	f22xsstd/sci12	3420	30.33	70.60	52.40	9.83	-.21	-.93	.179
Soc12	f22xhstd/soc12	3408	27.08	70.26	52.31	9.76	-.22	-.80	.178

Step 2: Baseline Models for Within and Between

As noted by Stapleton (2013), fit indices produced by SEM programs are confounded by misfit at both levels of the analysis, within and between clusters. The purpose of this step is to provide baseline models for later comparisons. Recall that fit indices such as CFI and TLI compare a target model with a null (baseline) model, one in which the variables are unrelated to each other. The goal of step 2 is to provide an unconfounded null model for the between model and a different unconfounded null model for the within model. To provide the baseline model for within, we constrain the covariances among the measured variables to zero at the within level and allow the variables to covary freely at the between level. The relevant portion of the Mplus syntax is shown in Figure 22.18. For the between-baseline model, we do the reverse: constrain the between-level covariances to zero and freely estimate the within-level covariances. The within-baseline model produced a χ^2 of 9498.886 with 21 df . For the between-baseline model $\chi^2 = 842.022$, $df = 21$. We will use these values in steps 3 and 4 to recalculate the CFI.

Steps 3 and 4: Theory-Consistent Model at One Level; Unconstrained Model at the Other Level

For step 3 (theory within), we estimate the conceptual model shown in the lower portion of Figure 22.16, while allowing the measured variables to covary freely at the between level. Compared to model 2, this model allows the same unconstrained version of the model at the between level, but imposes our hypothesized measurement and structural model at the within level. Thus, at the within level, HW10 and HW12 are used as indicators of a latent within Homework variable, and the reading, math, science, and social studies measured test scores are used as indicators of a latent within-school Achievement variable. This step allows

```
VARIABLE:
MISSING ARE ALL (-99);
NAMES ARE sch_id bys81a bys81b bys81c bys81d f1s36a2 f2s25f2
read12 math12 sci12 soc12 FamBack;

USEVARIABLES ARE
sch_id FamBack f1s36a2 f2s25f2 read12 math12 sci12 soc12;

CLUSTER = sch_id;

ANALYSIS: TYPE = TWOLEVEL;
ESTIMATOR IS ML;

MODEL:
%WITHIN%
FamBack WITH F1s36a2@0 f2s25f2@0 read12@0
math12@0 sci12@0 soc12@0;
F1s36a2 WITH f2s25f2@0 read12@0
math12@0 sci12@0 soc12@0;
f2s25f2 WITH read12@0 math12@0 sci12@0 soc12@0;
read12 WITH math12@0 sci12@0 soc12@0;
math12 WITH sci12@0 soc12@0;
sci12 WITH soc12@0;

%BETWEEN%
FamBack WITH F1s36a2 f2s25f2 read12
math12 sci12 soc12;
F1s36a2 WITH f2s25f2 read12
math12 sci12 soc12;
f2s25f2 WITH read12 math12 sci12 soc12;
read12 WITH math12 sci12 soc12;
math12 WITH sci12 soc12;
sci12 WITH soc12;
```

Figure 22.18 Step 2a: Generating a baseline model for the within portion of the model.

the evaluation of the model fit and parameters at the within-school level while recognizing that there is also a between-school level (but minimizing any misfit from that level). The relevant portion of the syntax to accomplish this analysis is shown in Figure 22.19. The fit information produced for this step is shown in Figure 22.20. These generally support this

```

MODEL:
%WITHIN%
HWork_w BY f1s36a2 f2s25f2;
Ach_w BY read12 math12 sci12 soc12;
Ach_w ON HWork_w FamBack;
HWork_w WITH FamBack;

%BETWEEN%
FamBack WITH F1s36a2 f2s25f2 read12
math12 sci12 soc12;
F1s36a2 WITH f2s25f2 read12
math12 sci12 soc12;
f2s25f2 WITH read12 math12 sci12 soc12;
read12 WITH math12 sci12 soc12;
math12 WITH sci12 soc12;
sci12 WITH soc12;

```

Figure 22.19 Syntax for step 3. The theoretical model is estimated at the within level, while an unconstrained model (all measured variables freely correlated) is estimated at the between level.

MODEL FIT INFORMATION	
Number of Free Parameters	51
Loglikelihood	
H0 Value	-63876.837
H1 Value	-63771.562
Information Criteria	
Akaike (AIC)	127855.674
Bayesian (BIC)	128176.925
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	128014.870
Chi-Square Test of Model Fit	
Value	210.551
Degrees of Freedom	12
P-Value	0.0000
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.064
CFI/TLI	
CFI	0.982
TLI	0.936
Chi-Square Test of Model Fit for the Baseline Model	
Value	10878.231
Degrees of Freedom	42
P-Value	0.0000
SRMR (Standardized Root Mean Square Residual)	
Value for Within	0.023
Value for Between	0.005

Figure 22.20 Model fit information for Step 3 (Within theoretical model).

phase of model testing, but, as noted, they confound misfit at both levels. The SRMR is calculated for both the within and between portions of the model, but even there we can see that the value for the between portion of the model is not perfect as we would think, given the unconstrained nature of the between model; the misfit in the within portion of the model bleeds over into the between portion. This is so because SEM seeks to minimize discrepancies between the model and the data for the entire model, not each portion separately.

Stapleton recommended recalculating the CFI using the χ^2 values from model 3 versus model 2a. The formula for CFI is $1 - \frac{\chi^2_{(3:\text{target})} - df}{\chi^2_{(2a:\text{null})} - df}$, with the added condition that the numerator and the denominator can only have a minimum value of zero (see Arbuckle, 2017, Appendix C for formulae for various fit indices). For the within theoretical model (between level unconstrained), $\text{CFI} = \frac{210.551 - 12}{9498.886 - 21} = .979$. The imposition of the theoretical model

at the within level resulted in a good fit of the model to the data. The model appears to do a good job in explaining the effects of Homework and SES on Achievement at the within-school level. It may also make sense to compare the AIC or other information criteria fit indices for these two models. The AIC for the step 3 model was 127855.674, versus 128469.145 for the model from step 2a. The AIC also supports the within-level theoretical model.

For step 4 we go through these same steps with the syntax for the within and between portions of the model reversed (assuming we specify the same model at both levels, as we have in Figure 22.16). That is, for this step the theoretical model is specified at the between level, whereas the measured variables are allowed to covary freely at the within level.

```

TITLE: Homework Ach SEM; Step 5

DATA:
  FILE IS homework smaller 4.dat;
VARIABLE:
  MISSING ARE ALL (-99);
  NAMES ARE sch_id bys81a bys81b bys81c bys81d f1s36a2 f2s25f2
  read12 math12 sci12 soc12 FamBack;

USEVARIABLES ARE
  sch_id FamBack f1s36a2 f2s25f2 read12 math12 sci12 soc12;

CLUSTER = sch_id;

ANALYSIS:  TYPE = TWOLEVEL;
            ESTIMATOR IS ML;
MODEL:
  %WITHIN%
    HWork_w BY f1s36a2 f2s25f2;
    Ach_w BY read12 math12 sci12 soc12;
    Ach_w ON HWork_w FamBack;
    Hwork_w WITH FamBack;
  %BETWEEN%
    HWork_b BY f1s36a2 f2s25f2;
    Ach_b BY read12 math12 sci12 soc12;
    Ach_b ON HWork_b FamBack;
    Hwork_b WITH FamBack;
OUTPUT:  STANDARDIZED SAMPSTAT;
```

Figure 22.21 Syntax for Step 5 (Theoretical model at both levels).

For step 4, $\chi^2 = 16.928$ (12 df) and AIC = 127662.051. For step 2b, $\chi^2 = 842.022$ (21 df) and AIC = 128469.145. The AIC is lower for the step 4 model, and the calculated CFI = .994. This step supports the imposition of the theoretical model at the between level.

All measurement and structural paths were statistically significant, meaningful, and reasonable for the within portion of the model at step 3. For the between theoretical model, all paths and factor loadings were statistically significant, with the exception of the path from Homework (between) to Achievement 12 (between). Although substantial, this path was not statistically significant ($\beta = .248, p = .073$). Keep in mind that the sample size for the between portion of the model (175 schools) is much smaller than for the within portion of the model. We will see whether this pattern is repeated in the next model.

Step 5: Theoretical Model at Both Levels

Step 5 imposes the theoretical model at both the within and the between level. As shown in Figure 22.21, this is accomplished by including the within syntax from step 3 with the between syntax from step 4. The fit indices for this model are shown in Figure 22.22. The

MODEL FIT INFORMATION	
Number of Free Parameters	39
Loglikelihood	
H0 Value	-63888.250
H1 Value	-63771.562
Information Criteria	
Akaike (AIC)	127854.501
Bayesian (BIC)	128100.163
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	127976.238
Chi-Square Test of Model Fit	
Value	233.377
Degrees of Freedom	24
P-Value	0.0000
RMSEA (Root Mean Square Error Of Approximation)	
Estimate	0.047
CFI/TLI	
CFI	0.981
TLI	0.966
Chi-Square Test of Model Fit for the Baseline Model	
Value	10878.231
Degrees of Freedom	42
P-Value	0.0000
SRMR (Standardized Root Mean Square Residual)	
Value for Within	0.023
Value for Between	0.019

Figure 22.22 Fit indexes for step 5.

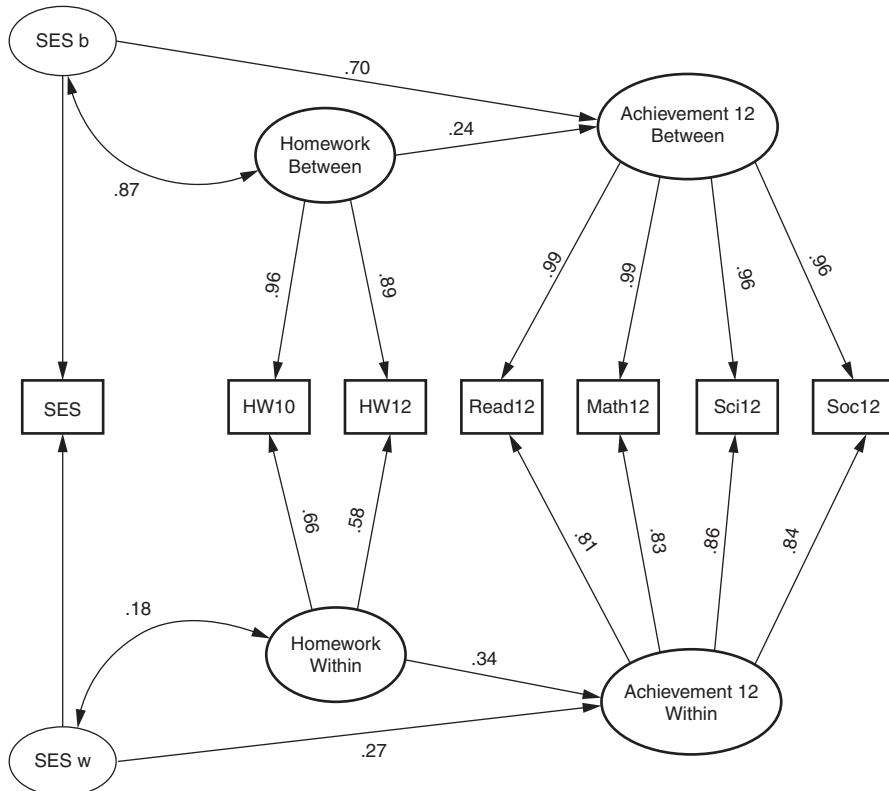


Figure 22.23 Standardized model results for step 5, with the theoretical model estimated both within and between schools. The model is simplified from the initial display (Figure 22.16)

standardized results from this model are displayed in Figure 22.23, and portions of the detailed unstandardized results are shown in Figure 22.24. Note that the model shown in Figure 22.23 is simplified from that shown in Figure 22.16 (it only includes the variables that have actual path coefficients in the output).

The overall model fit well according to the criteria we have generally used, and, according to the SRMR, both the within and between portions of the model reproduced their respective correlation matrices well. The coefficients for the within portion of the model were quite close to those from step 3, and the coefficients for the between portion of the model were very similar to those from step 4. Substantially different coefficients would suggest instability of the model at one or both levels (Stapleton, 2013).

The findings suggest that within schools, Homework has a large effect on Achievement ($\beta = .34$). For each one unit increase in time spent on homework per week out of school, Achievement will increase by 2.23 points. (Given the coding of the HW10 variable, these are not one-hour increments; Homework is indexed by HW10 and Achievement 12 by Read12). SES also had a substantial effect on Achievement at the within-school level.

At the between-school level, and as in step 4, the effect of Homework (between) on Achievement 12 (between) was not statistically significant. As much sense as it made to think that school-level homework should affect school-level achievement, this expectation was not supported. SES, in contrast, had a large and statistically significant school-level effect ($\beta = .70$). The average achievement level of a school is very much related to the SES level of that school. Given the group-mean centering inherent in these analyses, these between coefficients are

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Within Level				
HWORK_W BY				
F1S36A2	1.000	0.000	999.000	999.000
F2S25F2	1.018	0.093	10.994	0.000
ACH_W BY				
READ12	1.000	0.000	999.000	999.000
MATH12	1.004	0.019	53.480	0.000
SCI12	1.073	0.019	55.187	0.000
SOC12	1.037	0.019	55.113	0.000
ACH_W ON				
HWORK_W	2.229	0.223	10.007	0.000
ACH_W ON				
FAMBACK	3.028	0.208	14.556	0.000
HWORK_W WITH				
FAMBACK	0.121	0.017	7.001	0.000
Between Level				
HWORK_B BY				
F1S36A2	1.000	0.000	999.000	999.000
F2S25F2	0.955	0.079	12.158	0.000
ACH_B BY				
READ12	1.000	0.000	999.000	999.000
MATH12	1.169	0.041	28.384	0.000
SCI12	0.999	0.041	24.466	0.000
SOC12	1.011	0.041	24.628	0.000
ACH_B ON				
HWORK_B	1.533	0.902	1.700	0.089
ACH_B ON				
FAMBACK	5.249	1.003	5.235	0.000
HWORK_B WITH				
FAMBACK	0.279	0.037	7.609	0.000

Figure 22.24 Selected unstandardized output from step 5, with the theoretical model imposed at both the within and between levels.

confounded with the within model effects. They can be corrected by subtraction or statistically via a model constraint in Mplus (see Stapleton, 2013, for more information). For SES, the corrected (contextual) between-school effect was $5.249 - 3.028 = 2.221$ (the unstandardized between coefficient minus the unstandardized within coefficient). This value is statistically significant (output not shown), and suggests that two students who have the same levels of SES, but who attend schools that are one unit different on SES (recall SES is a mean of z-scores, so this is close to a SD unit difference) will differ in their achievement by a little more than 2 points. The SES level of schools really does matter (as we also found in Chapter 11)! Note also the high correlation between school-level SES and school-level Homework. Higher SES schools appear to have higher average levels of homework demand, as well.

Step 6: Random Coefficients

The final step allows the regression coefficients from SES to Achievement or Homework to Achievement (or both) to vary across schools. That is, perhaps SES has a stronger effect on Achievement in school A as opposed to school B, or Homework has a stronger effect on Achievement in one school as opposed to another. I was unsuccessful in running a model with either (or both) of these specifications, however. I think this was likely a result of the relatively low ICC level for the homework measured variables and empirical underidentification at the between level (see Chapter 16, note 1 for an explanation of empirical underidentification). See the website (<http://tzkeith.com>) for some possible ways of dealing with this and estimating the original model. For our purposes here, however, I will analyze the model without the homework variables. Figure 22.25 shows selected syntax for a model that included only the Achievement variables and SES. The distinct portions of this syntax are highlighted, and include the addition of a RANDOM command in the ANALYSIS line and the naming (s1) and defining of random paths/slopes by school for the regression of the latent Achievement variable on SES as a part of the MODEL %WITHIN% statement.

Edited output from the analysis is shown in Figure 22.26. Not shown are the fit indices. These are limited to the Log-likelihood and the Information Criteria when random coefficients are allowed. The AIC for this model (97049.741) was very slightly lower than that for a step 5 version of this same model (97050.041, not including Homework), thus supporting the addition of the random slopes. Interestingly, if the abIC or the BIC were used for model

```

TITLE: Homework Ach SEM; Step 6 SES only

DATA:
FILE IS homework smaller 4.dat;
VARIABLE:
MISSING ARE ALL (-99);
NAMES ARE sch_id bys81a bys81b bys81c bys81d f1s36a2 f2s25f2
read12 math12 sci12 soc12 FamBack;

USEVARIABLES ARE
sch_id read12 math12 sci12 soc12 FamBack;

CLUSTER = sch_id;

ANALYSIS:   TYPE = TWOLEVEL RANDOM;
ESTIMATOR IS ML;

MODEL:
%WITHIN%
Ach_w BY read12 math12 sci12 soc12;
s1 | Ach_w ON FamBack;

%BETWEEN%
Ach_b BY read12 math12 sci12 soc12;
Ach_b ON FamBack;

OUTPUT: SAMPSTAT;

```

Figure 22.25 Selected Mplus syntax for a revised step 6 model. The model only examines the effect of SES on Achievement, and tests for the possibility of random slopes.

comparisons, with their greater reward for parsimony, the step 5 version of this model would have been supported instead ($aBIC = 97115$ versus 97117 and $BIC = 97185$ versus 97191 for models 5 and 6, respectively).

Of primary interest in the output shown in Figure 22.26 is the variance for S1. The fact that the value is statistically significant means that there is indeed variation in the paths from SES to the latent Achievement variable across schools (assuming $\alpha = .05$). The value is not large in relation to the standard error, however, and if we had chosen an alpha of .01 or $aBIC$ as our fit index, we would have concluded that these paths did not vary across groups after all.

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Within Level				
ACH_W BY				
READ12	1.000	0.000	999.000	999.000
MATH12	1.002	0.019	52.885	0.000
SCI12	1.074	0.020	54.737	0.000
SOC12	1.038	0.019	54.943	0.000
Residual Variances				
READ12	26.356	0.840	31.377	0.000
MATH12	23.725	0.780	30.427	0.000
SCI12	20.244	0.745	27.175	0.000
SOC12	22.717	0.779	29.177	0.000
ACH_W	45.397	1.700	26.708	0.000
Between Level				
ACH_B BY				
READ12	1.000	0.000	999.000	999.000
MATH12	1.296	0.087	14.871	0.000
SCI12	0.982	0.077	12.734	0.000
SOC12	0.997	0.076	13.135	0.000
ACH_B ON FAMBACK				
	3.607	0.443	8.135	0.000
Means				
S1	3.727	0.231	16.113	0.000
Intercepts				
READ12	52.259	0.197	265.264	0.000
MATH12	52.919	0.221	239.550	0.000
SCI12	52.425	0.215	243.597	0.000
SOC12	52.324	0.214	244.405	0.000
Variances				
S1	1.702	0.761	2.236	0.025
Residual Variances				
READ12	0.204	0.243	0.841	0.401
MATH12	0.578	0.320	1.805	0.071
SCI12	1.412	0.331	4.265	0.000
SOC12	1.354	0.341	3.970	0.000
ACH_B	1.905	0.561	3.394	0.001

Figure 22.26 Selected Mplus output for a revised step 6 model. The model only examined the effect of SES on Achievement, and tested for the possibility of random slopes.

I hope this abbreviated example gives you a taste of how multilevel analysis can work in an SEM framework. Obviously, this is a complex topic and we have just scratched the surface, and have illustrated the method using a single program, Mplus. I like Stapleton's (2013) approach and series of steps because it makes a lot of sense from an SEM orientation. There are other orientations and other excellent discussions, as well (e.g., Heck & Thomas, 2015, chap. 6; Hox, 2018, chap. 15; Muthén & Asparouhov, 2011, and, of course, the Mplus manual, Muthén & Muthén, 2017). Other SEM programs, including LISREL and EQS (but not Amos), perform MLSEM.

A Non-ML SEM Alternative

One of our reasons for exploring multilevel models is that given complex sampling designs such as NELS, with students clustered and selected within schools, the standard errors of effects are likely inflated. The larger the ICC, the more these values are underestimated. MLSEM solved this problem by examining separately the within- and between-school effects in our example from NELS, with the added benefit of allowing us to compare the importance of effects at each level. We found, for example, that individual-level (within-school) homework effects continue to appear important as in previous chapters, whereas school-level effects are not. In contrast, the SES level of a school appears quite important in influencing achievement.

Suppose, however, that all we really care about are effects of these variables at an individual level, but we want to make sure we have the correct standard errors so that we do not conclude that an effect is statistically significant when it is not. Another possible approach is a “designed-based analysis,” which “adjusts estimates of standard errors given a sampling design but does not explicitly model the sampling design” (Stapleton, 2013, p. 528).

A design-based Mplus analysis similar the Homework model used for MLSEM is shown in Figures 22.27 (syntax) and 22.28 (selected output). As can be seen, individual-level student Homework and SES both had statistically significant and large effects on Achievement

```

TITLE: Homework Ach SEM; designed based analysis, like step 5

DATA:
  FILE IS homework smaller 4.dat;
VARIABLE:
  MISSING ARE ALL (-99);
  NAMES ARE sch_id bys81a bys81b bys81c bys81d f1s36a2 f2s25f2
  read12 math12 sci12 soc12 FamBack;

USEVARIABLES ARE
  sch_id FamBack f1s36a2 f2s25f2 read12 math12 sci12 soc12;

STRATIFICATION IS sch_id;

ANALYSIS: TYPE = COMPLEX;

MODEL:
  HWork_w BY f1s36a2 f2s25f2;
  Ach_w BY read12 math12 sci12 soc12;
  Ach_w ON HWork_w FamBack;
  Hwork_w WITH FamBack;

OUTPUT: STANDARDIZED SAMPSTAT;

```

Figure 22.27 Syntax for a MLSEM alternative. This analysis takes the sample stratification into account, but does not analyze a multilevel structure.

in grade 12. By way of comparison, if we had not taken the stratified sample design into account using the TYPE=COMPLEX command, we would have still estimated the unstandardized effect of Homework on Achievement as 2.212, but with an *SE* of .179. Design-based analyses such as these can also take into account sampling weights, often used in large datasets like NELS.

Two final points relate to the variables at the second (or higher) levels of a ML analysis. The first concerns the interpretation of the between-level variables in our example analysis. Reread my previous statement, “We found, for example, that individual-level (within-school) homework effects continue to appear important as in previous chapters, whereas

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HWORK_W BY				
F1S36A2	1.000	0.000	999.000	999.000
F2S25F2	0.954	0.053	17.840	0.000
ACH_W BY				
READ12	1.000	0.000	999.000	999.000
MATH12	1.044	0.015	70.950	0.000
SCI12	1.052	0.015	70.361	0.000
SOC12	1.028	0.014	75.279	0.000
ACH_W ON				
HWORK_W	2.212	0.196	11.304	0.000
ACH_W ON				
FAMBACK	3.671	0.180	20.342	0.000
HWORK_W WITH				
FAMBACK	0.418	0.025	16.905	0.000

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
HWORK_W BY				
F1S36A2	0.725	0.022	32.734	0.000
F2S25F2	0.606	0.020	30.133	0.000
ACH_W BY				
READ12	0.848	0.006	131.665	0.000
MATH12	0.868	0.006	152.130	0.000
SCI12	0.877	0.005	168.245	0.000
SOC12	0.863	0.006	144.349	0.000
ACH_W ON				
HWORK_W	0.348	0.024	14.450	0.000
ACH_W ON				
FAMBACK	0.370	0.018	21.079	0.000
HWORK_W WITH				
FAMBACK	0.391	0.018	21.713	0.000

Figure 22.28 Results from the design-based analysis of the effect of Homework on Achievement. Standard errors are corrected based on the stratification by schools.

school-level effects do not. In contrast, the SES level of a school appears quite important in influencing achievement.” It is probably pretty easy to understand school-level effects of SES, and the finding of school-level effects shown on achievement likely makes sense (and is music to the ears of realtors). But the meaning of school-level Homework is a little less clear-cut. Variables at the between level that are means of individual-level variables are not always easy to understand. Second, it is of course possible to have variables that appear only at the between level, such as variables that are true school-level variables. For example, a school-type variable (public versus private) would properly appear at the between level in our model, not the within level.

SUMMARY

This chapter has presented a very brief introduction to two advanced SEM topics: testing for interactions among latent variables and multilevel SEM. In Part 1 we discussed methods for testing for interactions (moderation) among continuous variables in multiple regression, and in Chapters 18 and 19 we have discussed testing for interactions using multigroup SEM when one of the variables is a grouping variable (experimental/control, female/male, etc.). In this chapter, we see that it is possible to test for interactions among continuous latent variables as well. Conceptually, this method is an extension of the cross-products approach we used in Part 1. We create cross products of the measured indicators of latent variables we expect to interact, and we use those cross products as indicators of a latent interaction variable. In our example, we tested whether Previous Achievement and Homework, both latent variables with multiple continuous indicators, interacted in their effect on subsequent student Grades. Our results suggested that these variables did not interact; that is, that Homework had similar effects on Grades for all levels of Achievement.

In a second analysis, we tested whether the latent Homework variable had a non-linear effect on Grades. We saw in Chapter 8 that we can test for a curve in a regression plane by entering a cross product of a variable with itself in the regression. A curve in a regression may be thought of as a variable interacting with itself on some outcome. Thus, likewise in SEM, we can test for curvilinear effects in the same way we test for interactions. Allowing a latent Homework² variable to affect Grades improved the fit of our model, and the variable had a statistically significant effect on Grades; Homework, it appears, has differential effects on Grades depending on how much homework students do. That is, the effect of homework depends on the amount of homework. We were able to generate a regression equation and plot the nature of the effect using factor scores, both of which showed that each additional hour increase in homework time has a smaller effect on Grades than did the previous hour increase.

We used another Homework model to illustrate multilevel SEM. In Part 1 we illustrated a multilevel regression of Achievement on SES. Here, we developed a multilevel SEM model in which SES affected a latent Achievement variable at both the within-school and between-school level. The model also incorporated a latent Homework variable as both a within-school and between-school effect. We illustrated testing a series of multilevel models designed to test both within-school effects and school-level effects. The primary finding was that SES had both individual/within-school effects, and quite large between-school effects, but that Homework’s effects were primarily at the student/within level. We also found evidence for possible differences in the effects of SES across schools.

For both topics presented in this chapter, my intent was not to provide a detailed explanation of the methodology. Instead, the hope was to provide a way of understanding these advanced topics using something that you already understand, multiple regression and SEM. References were provided for both methods for further study.

EXERCISES

1. Find a research article in an area of interest to you that tested for interactions between latent variables in the analysis or tested for a curvilinear effect for a latent variable. What method did the authors use to create the interaction/product term? Did they use some of the same jargon used here? Was the interaction/curve statistically significant? How did the authors describe the results? Did they graph the interaction or curve? How did they do so? Are you able to understand the findings using the information from this chapter? What aspects of the reported findings are still puzzling to you?
2. Find a research article in an area of interest to you that used multilevel SEM. Did the research focus on measured variables (a path analysis) or on latent variables (a latent variable structural equation model)? What program was used to conduct the MLM? What were levels of analysis? Which variables were analyzed at each level? What models were tested? Were they similar to the models tested here? Were you able to interpret the results of the MLSEM using the suggestions in this chapter? What aspects of the reported findings are still puzzling to you?

23

Summary

Path Analysis, CFA, SEM, Mean Structures, and Latent Growth Models

Summary	562
<i>Path Analysis</i>	562
<i>Error</i>	565
<i>Confirmatory Factor Analysis</i>	566
<i>Latent Variable SEM</i>	567
<i>SEM With Mean Structures</i>	570
<i>CFA With Latent Means and Invariance</i>	572
<i>Latent Growth Modeling</i>	573
Issues Incompletely or Not Covered	574
<i>Maximum Likelihood Estimation</i>	574
<i>Missing Values</i>	574
<i>Sample Size, Number of Parameters, and Power</i>	578
<i>Longitudinal Models</i>	579
<i>Formative Measures</i>	580
<i>Categorical Variables</i>	580
<i>Differences Across Programs</i>	580
<i>Causality and the Veracity of Models</i>	580
Additional Resources	581
<i>Introductory Texts</i>	581
<i>More Advanced Resources</i>	581
<i>Books About Specific SEM Programs</i>	582
<i>Reporting SEM Results</i>	582
<i>Cautions</i>	583

Part 1 discussed multiple regression as a research tool. Part 2 has been concerned with the “And Beyond” portion of the title of the book and has focused on path analysis, confirmatory factor analysis, structural equation modelling, and latent growth modeling. For SEM, we went from basic models to those with means, multiple groups, interaction terms, and multiple levels. This final chapter will begin with a review and summary of Part 2. I will then briefly discuss several topics about which you should be aware but have not yet been covered in this text.

SUMMARY

Path Analysis

Basics. Throughout this book I have assumed that we are primarily interested in estimating the effects of one variable on another. We became even more explicit in this assumption in Part 2, where we focused on variations of structural equation modeling. The journey of SEM discovery started with path analysis, the simplest form of SEM.

If, through previous research, relevant theory, and logic, you can specify the likely causal relations among a set of variables, you can (given a few other conditions) estimate these effects using the correlations among the variables and simple algebra. Figure 23.1 shows such a model with the likely causal relations among the variables represented by paths. The paths represent a weak causal ordering, meaning that they do not assert that one variable directly affects another but rather that if the two variables are causally related the influence is in the direction shown, rather than the reverse. If this model includes a one-way causal flow, we can forgo the algebra and use multiple regression to estimate the effects of one variable on another. These estimates, or paths, are estimated by the standardized and unstandardized coefficients in multiple regression. The paths to Achievement are estimated by the simultaneous regression of Achievement on Family Background, Ability, Motivation, and Coursework; the paths to Coursework are estimated using the regression coefficients from the regression of Coursework on Family Background, Ability, and Motivation, and so on. The standardized paths from disturbances, represented by the variables in ovals labeled d1 through d4, are estimated as $\sqrt{1 - R^2}$ from each regression equation. The disturbances represent all other influences on these variables beyond the variables in the model; many writers use the term residuals (consistent with MR) or errors instead of disturbances. We interpret the paths in much the same way as we did explanatory regressions in Part 1. The standardized paths (β 's) represent the change in standard deviation units in the outcome for each standard deviation change in the influence, and the unstandardized paths (b 's) represent the amount of change in the outcome for each 1-unit change in the influence.

We dealt with some of the jargon and symbols you are likely to encounter in structural equation modeling. Measured variables, the variables actually measured in your research,

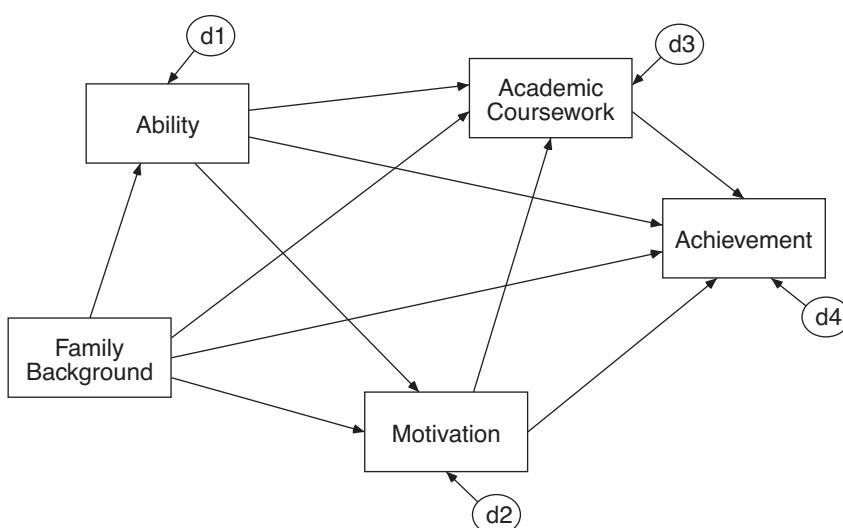


Figure 23.1 Path model; the paths represent the presumed effects of one variable on another.

are symbolized by rectangles. Unmeasured, or latent, variables are symbolized by circles or ovals. Disturbances/residuals represent unmeasured variables not considered in the model. Recursive models have arrows pointing in one direction only, whereas nonrecursive models have feedback loops, or arrows pointing in two directions. Just-identified models are those for which we have just enough information to solve for the paths, and overidentified models are those for which we have more information than we need and can thus estimate some of the paths in more than one way. Underidentified models are those for which we have more paths than we have information to estimate the paths; they are, therefore, not solvable without additional constraints. The causes of exogenous variables come from outside the model; exogenous variables have no paths pointing toward them. Endogenous variables are effects; they have paths pointing to them in the model. Most of this jargon is summarized in Figure 12.17.

The paths provide estimates of the direct effects of one variable on another. It is also possible to estimate indirect effects, such as the effect of Motivation on Achievement through Coursework in Figure 23.1. We can estimate the indirect effects by multiplying the paths involved. Indirect effects are also referred to as mediation (especially in longitudinal models): we may be interested in the extent to which Coursework *mediates* the effect of Motivation on Achievement. When the indirect and direct effects of one variable on another are added together, they provide an estimate of the total effect of one variable on the other. We can also calculate total effects directly using the regression coefficients from a series of sequential regressions. We finished Part 1 with questions about which type of MR to use. Although we had discussed direct versus total effects and mediation, these distinctions became much clearer with the development of path analysis: simultaneous regression focuses on direct effects, whereas sequential regression focuses on total effects. If nothing else, path analysis provides a valuable heuristic device for understanding and organizing the results of multiple regressions. I argued that path analysis should be the method of choice for those interested in MR for explanatory, nonexperimental research.

One noteworthy aspect of this process is how we made decisions concerning the influence of one variable on another: through logic, theory, and previous research. The correlations did not inform these decisions, they merely provided fuel for the calculations once we developed the causal model. To make a valid inference of cause and effect, there must be a functional relation between the variables, the cause must precede the effect in time, and the relation must not be spurious. For multiple regression to provide valid estimates of paths, we must be able to assume that there are no omitted common causes of the presumed cause and presumed effect, that there is no reverse causation, and that the exogenous variables are perfectly measured.

Dangers

The biggest danger of path analysis is that of omitted common causes. When a common cause (a variable that affects both the presumed cause and the presumed effect) is omitted from the model, we get inaccurate estimates of the effects of one variable on another. We showed that the problem of omitted common causes is at the heart of the dreaded spurious correlation, which is, in turn, the reason for the admonition that we should not infer causation from correlations. When common causes are accounted for, paths provide accurate estimates of the effects of one variable on another. True experiments provide powerful evidence of cause and effect because the process of random assignment to groups rules out the possibility of common causes. The problem of common causes is not unique to path analysis, but is paramount in any nonexperimental (and most quasi-experimental) research. Omitted common causes are one likely reason for variability in findings from such research. If you

disagree with the results of nonexperimental research, focus on the possibility of the research having omitted a common cause of the presumed cause and effect. You need to go beyond mere armchair analysis, however, and provide *evidence* of an omitted common cause.

The danger of common causes does not mean that *all* possible causes of every variable must be included in a model. If a variable, for example, affects an endogenous variable in research, but not an exogenous variable, it does not necessarily need to be included. Likewise, it is not necessary to include intervening or mediating variables in models for the results to be valid. Intervening variables are valuable, however, in that they can help us understand *how* one variable goes about influencing another. Noncommon causes and intervening variables may both be valuable in helping devise nonequivalent overidentified models, however.

Another danger in path analysis occurs when you draw a path in the wrong direction, although the extent to which this is a problem depends on the paths involved. You should not use reciprocal paths (nonrecursive models) to avoid making decisions concerning the direction of causation. Nonrecursive models are much more complex than recursive models and cannot be estimated through ordinary multiple regression. Even worse is to estimate a model via MR with a path drawn in one direction, and then the other direction; the results will *not* tell you which direction is correct.

The solution to both of these dangers is to have a good understanding of relevant theory and previous research. Think about the variables in your model, how they are related to one another. If necessary, bolster causal assumptions (e.g., *a* affects *b* rather than *b* affects *a*) through the use of longitudinal data. Think about possible common causes, and investigate them in the research literature. If necessary, test common causes in the research itself. In fact, most of what you should do to ensure the adequacy of your model boils down to the same advice for developing a model in the first place: theory, previous research, and logic. One advantage of path analysis over ordinary MR is that the figural display of the model makes your assumptions, and also any errors, very obvious.

Path Analysis Using SEM Programs

There are special computer programs for analyzing structural equation models, including path models. In Chapter 14 I illustrated their use for path analysis. Although the example used the computer program Amos, the concepts generalize to other SEM programs, and the web site illustrates the use of several such programs.

Your knowledge of MR and path analysis translates directly into SEM programs. Although there are differences in the look and labeling of output, the output from SEM programs will list unstandardized paths, standard errors, and statistical significance, along with standardized paths, correlations, covariances, and variances. Most programs will also provide tables of direct, indirect, and total effects (both standardized and unstandardized), along with their standard errors.

SEM programs become even more valuable in the analysis of overidentified models. When models are overidentified (when we have more information than we need to estimate the paths), they have positive degrees of freedom. The covariance matrix implied by the solved model will also differ to some extent from the covariance matrix that was used to solve the model when models are overidentified, and the extent of the similarity or dissimilarity of these two matrices can be used to assess the fit of the model to the data. There are a plethora of fit indexes for SEM, all of which are designed to assess the fit of the model to the data, or the likelihood that the solved model could have produced the data. We focused on RMSEA, SRMR, CFI, and TLI as measures of the fit of a single model to the data. Although we talk of the fit of the model, strictly speaking, what is really assessed is the veracity of the overidentifying restrictions (e.g., paths constrained to zero or some other value) in the model.

A major advantage of SEM programs is that they may be used to compare the fit of competing theoretical models. When two models are nested (one is a more constrained version of the other), the change in χ^2 between the two models can be used to determine which model better explains the data. When $\Delta\chi^2$ is statistically significant (when compared to Δdf), we favor the better fitting, but less parsimonious, model. When $\Delta\chi^2$ is not statistically significant, we favor the more parsimonious model (the model with the larger df). The AIC and related indexes (e.g., BIC, abIC) can be used to compare nonnested, competing models. I provided some tentative advice on rules of thumb for fit indexes (Chapter 14) but noted that others may well give different advice, that thinking about fit indexes will likely change over time, and that different fields of study may have different conventions.

Any overidentified model will likely have a number of models that are equivalent to it, models that cannot be differentiated from it based on fit. Such models may have paths reversed or replaced by correlations. We discussed rules for developing equivalent models; these rules are also useful for developing nonequivalent models, models that *can* be differentiated based on their fit. We saw that carefully designed nonequivalent models may be able to obviate one of the threats we encountered with models estimated through MR: a path drawn in the wrong direction. SEM programs can also analyze nonrecursive models.

If you can develop overidentified models, there are advantages to using an SEM program instead of a MR program. If you are using MR to estimate path models, there are few reasons to strive for overidentified models. If, however, you are using an SEM program, you should see if you can develop an overidentified model prior to estimation. Whichever method is used, be aware of the threat of equivalent models.

Error

One assumption required to interpret regression (path) coefficients in a causal fashion is that the exogenous variables be measured without error. We rarely satisfy this assumption, and thus we need to know the effect of this violation on our estimates of the effects of one variable on another. To expand this discussion, I noted that unreliability and invalidity affect *all* types of research, not just path analysis and multiple regression. Problems in measurement in both the independent and dependent variables affect our research results.

Reliability is the converse of error. Error-laden measurements are unreliable, and reliable measurements contain little error. We can consider reliability from the standpoint of variance, by thinking of true score variance as the total variance in a set of scores minus the error variance. In path analytic form, we can think of a person's score on a measurement as being affected by two influences: their true score on the measure and errors of measurement. The true score and error influences are *latent variables*, whereas the actual score the person earns on the measurement is a *measured variable*. These concepts are important for research purposes because other variables generally correlate with the true score but not the error. For this reason, the reliability of a measurement generally places an upper limit on the correlation a variable can have with any other variable. Unreliable measurements can make large effects look small and statistically significant effects look nonsignificant.

MR and path models assume that the variables in our models, and especially the exogenous variables, are measured with perfect reliability. We demonstrated that if the variables in our models were unreliable (but we assumed perfect reliability) our estimates of the effects of one variable on another were inaccurate and were often underestimates of true effects. Given the complexity of path models, unreliability can also result in the overestimation of true effects.

Reliability is not the only aspect of measurement that needs to be considered, however; there is also validity. As it turns out, validity is a subset of reliability. We can get closer to valid measurement, closer to the constructs of interest in our research, by using multiple measures of constructs.

Latent variable structural equation modeling seeks to move closer to the constructs of interest in our research by using such multiple measures. With latent variable SEM, we simultaneously perform a confirmatory factor analysis of the measured variables in our research to get at the latent variables of true interest, along with a path analysis of the effects of these latent variables on each other. In the process, latent variable SEM removes the effects of unreliability and invalidity from consideration of the effects of one variable on another and avoids the problem of imperfect measurement. In the process, latent variable SEM gets closer to the primary questions of interest: the effect of one *construct* on another.

Although our discussion focused on the effects of imperfect measurement in multiple regression and path analysis, it is worth remembering that measurement affects every type of research, however that research is analyzed. With the addition of latent variables to SEM, we are able to take measurement problems into account—to model them—and thus control for them.

Confirmatory Factor Analysis

We spent two chapters focused on confirmatory factor analysis, the measurement portion of the latent variable SEM model. CFA focuses on and tests hypotheses about the constructs measured in our research. For example, the CFA model in Figure 23.2 asserts that the 12 *measured*

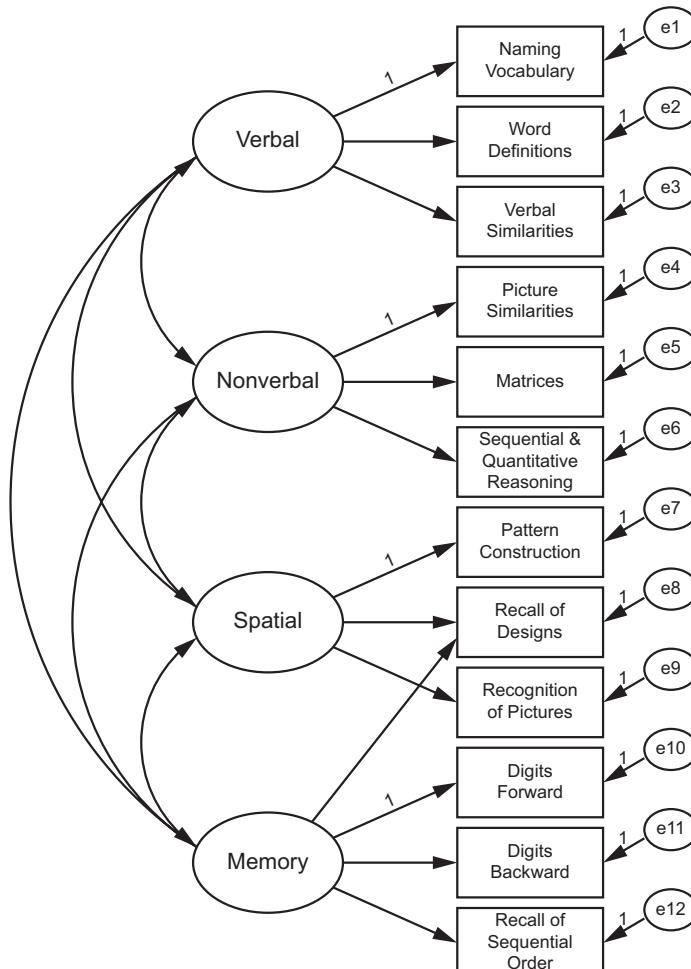


Figure 23.2 Confirmatory factor analysis model.

variables (subtests from the Differential Ability Scales Second edition) are really reflections of four broader abilities or constructs: Verbal, Nonverbal, Spatial, and Memory abilities. The fit statistics associated with the estimation of this model will tell us whether it is indeed reasonable to assume that these 12 measured variables are indicators of four such general latent abilities. We can interpret the factor loadings (the paths from the latent to the measured variables) in two ways. First, we can compare the effects as evidence of the relative validity of each test in measuring the corresponding factor (a CFA-type interpretation). We can also consider these paths as the effect of the latent variables on the measured variables (an SEM-type interpretation).

With path models, we added disturbances to account for all other influences on endogenous variables besides the variables in the model. We do something similar with the CFA-measurement models and add latent variables reflecting all other influences on each measured variable beyond its corresponding latent variable. These “all other influences” are, in fact, unreliability and invalidity, or errors of measurement. The latent variable $e1$ (for error), for example, symbolizes all other influences on the Naming Vocabulary test beyond Verbal Ability. Such influences include measurement error and specific/unique influences, such as specific vocabulary knowledge.

As in other types of SEM, with CFA we can use fit indexes to compare competing models, models that hypothesize different constructs or different compositions of these constructs. The SEM programs also provide more detailed fit statistics that may be useful for modifying poorly fitting models. Hierarchical models are also possible; for example, we tested a model in which we hypothesized that the four latent factors in Figure 23.2 were, in turn, reflections of a single general intellectual ability factor.

Latent Variable SEM

In Chapters 17 through 19 and 22, we combined path analysis and CFA into latent variable SEM. With multiple measures of the constructs of interest, SEM performs simultaneous confirmatory factor analysis of the constructs in a model and path analysis of the effects of these constructs on each other. Figure 23.3 shows such a model, designed to determine the effects of peer rejection on kindergarten students’ academic and emotional adjustment. The model included eight measured variables (in rectangles) designed to measure four constructs (in large ovals). We hypothesized that the constructs affected each other as shown by the paths connecting each latent variable. The model tested whether peer rejection affects academic and emotional adjustment and whether this effect is partially mediated (indirect effect) by children’s classroom participation. We found that all three types of effects—direct, indirect, and total effects—were meaningful and interesting.

It is generally preferable in all four variations of SEM (path analysis, CFA, latent variable SEM, and latent growth models) to test hypotheses about models by comparing competing models. “The fact that one model fits the data reasonably well does not mean that there could not be alternative models that fit better. At best, a given model represents one tentative explanation of the data. The confidence in accepting such an explanation depends, in part, on whether other, rival explanations have been tested and found wanting” (Loehlin & Beaujean, 2017, p. 63). We did so with this example and found another model that both made sense and had a better fit than the initial model. We also discussed two equivalent alternatives to this model that had different interpretations, but which were statistically indistinguishable from our accepted model.

We can easily build more complex models than that shown in Figure 23.3. Figure 23.4, for example shows a model in which one latent variable is indexed by a single measured variable. The model also includes correlated errors, the specification that the unique aspects of the measures of one construct share something in common with those of another construct

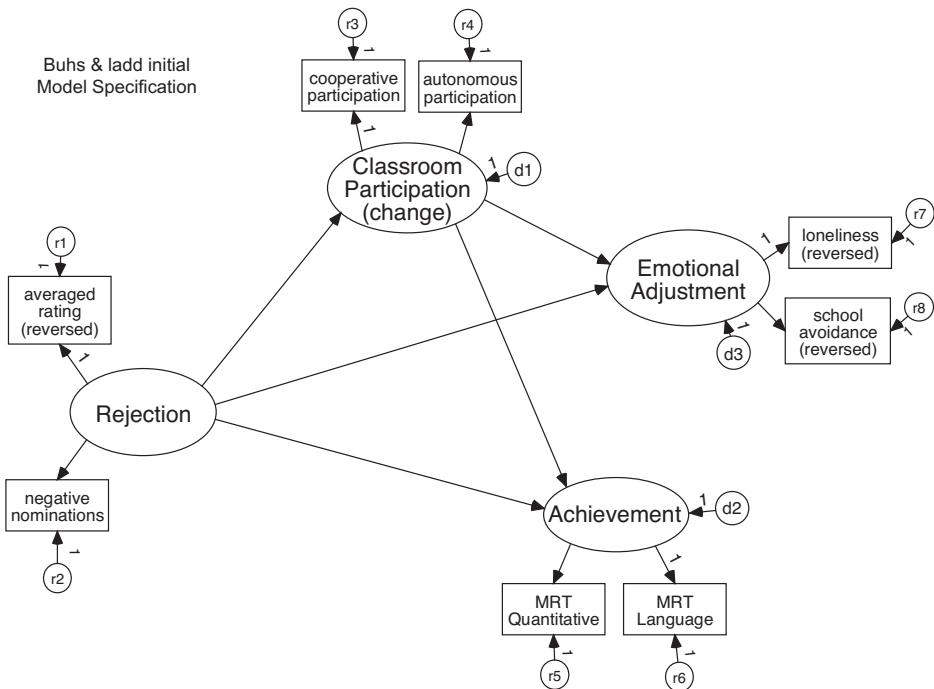


Figure 23.3 Latent variable structural equation model.

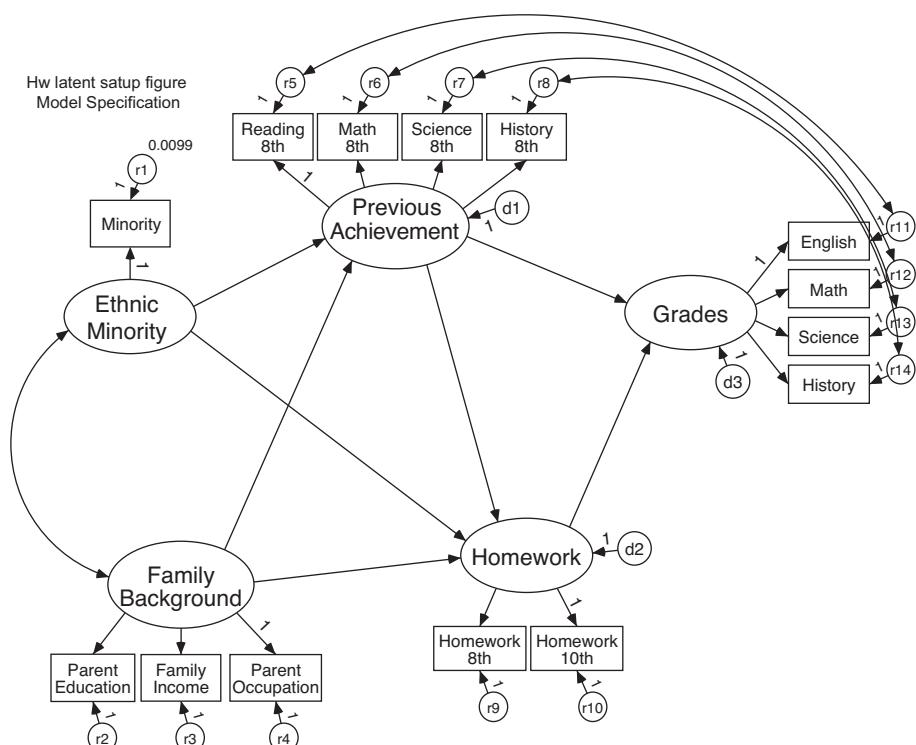


Figure 23.4 More complex latent variable SEM. This model includes a single-indicator latent variable and correlated errors.

beyond the effect of one construct on another. Such specifications are common in longitudinal research in which the same measures are obtained at several times or when different respondents are asked to provide assessments of multiple constructs. In the exercises for Chapter 18, for example, both parents and teachers provided feedback concerning multiple constructs. Correlated errors were used to control for respondent variance and remove it from consideration of the effects of one variable on another. Latent variable panel models (Chapter 18) are often used to study longitudinal developmental processes and to answer questions of which variable affects which.

It is possible to test for interactions (moderation) between categorical and other variables in SEM through multisample SEM. We analyzed the homework model separately for ethnic minority and White students, for example. By constraining various parameters to be equivalent across groups and comparing the fit of these models to models without constraints, we were able to determine that Homework (and other variables in the model) had the same effect on Achievement for one group as for the other. That is, we found that Homework and Ethnic background did not interact in their effect on Achievement, that ethnic background did not moderate the effect of Homework on Achievement.

It is also possible to test for interactions (moderation) between continuous latent variables in SEM. Conceptually, the approach is an extension of the method used in Part 1, where cross-products of the two interacting variables are included in the analysis along with the variables used to create those cross-products. With latent variable SEM, cross-products are created of the measured-variable indicators of the latent variables that interact, and those cross-products used as indicators of a latent interaction term. A conceptual model of such an analysis is shown in Figure 23.5 Given that tests for curves in a regression plane may be thought of as a test of whether a variable interacts with itself, the same method may be used to test for nonlinear effects of latent variables. We illustrated both methods in Mplus using a homework example. The first analysis suggested that Previous Achievement did not

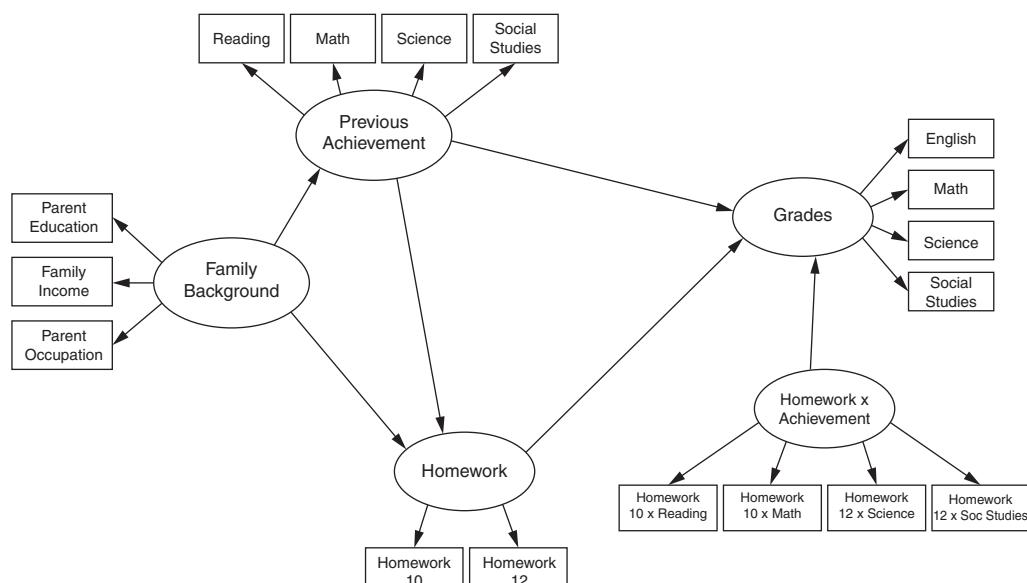


Figure 23.5 Conceptual model with the addition of latent variable testing the possible interaction of Previous Achievement and Homework in their effect on Grades.

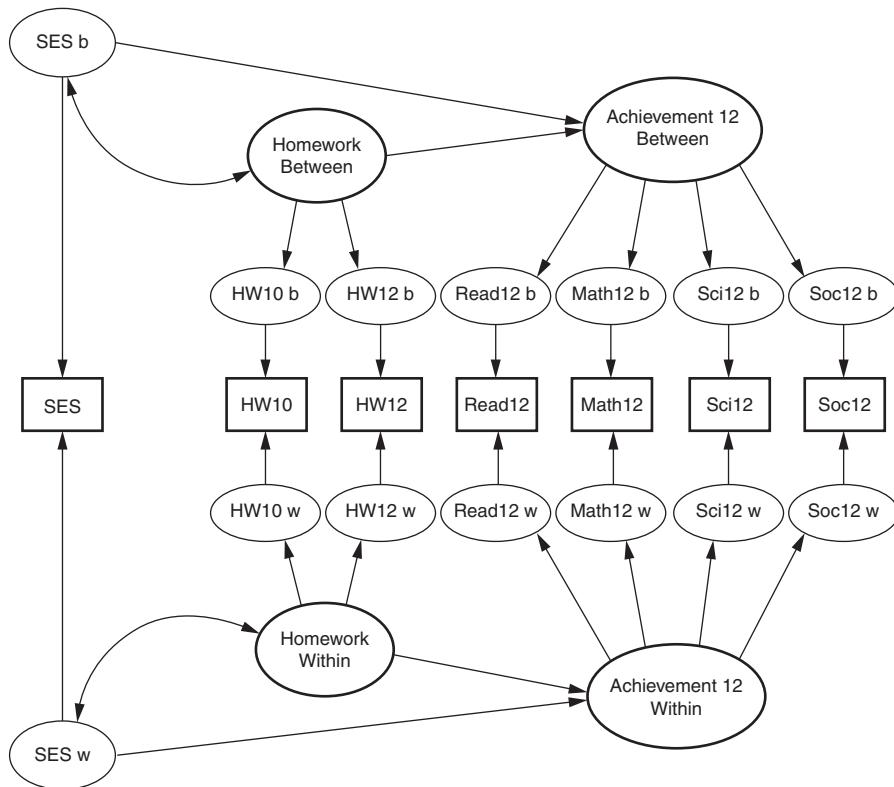


Figure 23.6 Conceptual model, multilevel effects of Homework and SES on Achievement test scores in Grade 12.

moderate the effect of Homework on Grades. The second suggested that the latent Homework time variable had diminishing returns on the Grades outcome. That is, students who completed small amounts of homework would benefit more from an hour increase in homework time than would students who were already spending a lot of time on homework.

We saw in Part 1 why to and how to analyze models where data have a nested structure, such as students nested within schools. It is also possible to analyze such multilevel models via some SEM programs. We again used a homework model to illustrate this method. The conceptual model underlying this analysis is shown in Figure 23.6. For this analysis, we found a large effect for Homework at the individual level, but no statistically significant effect at the school level. SES, however, showed a large school-level effect.

SEM With Mean Structures

Chapter 19 introduced the topic of latent means and intercepts in SEM. We had actually estimated latent means in a few previous examples via the inclusion of a dummy variable in a latent variable SEM, but in this chapter we made the issue of estimation of latent means more explicit. Not only are there advantages for the analysis of means in SEM, but the understanding of latent means is needed for subsequent topics, including invariance and latent growth models.

The model shown in Figure 23.7 tests the effect of Sex on time spent on Homework, among other things. The unstandardized path from the dummy variable Female (coded 0 for boys and 1 for girls) to Homework would tell us the difference for boys and girls on the latent Homework variable. A big advantage of SEM is that by modeling the errors of

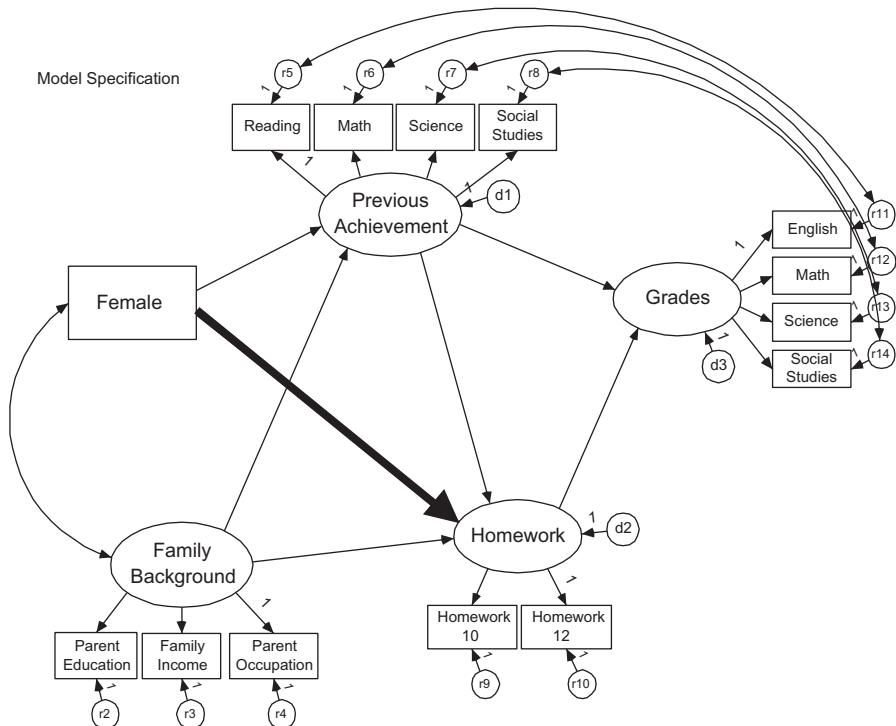


Figure 23.7 Testing for latent mean (intercept) differences using a categorical (dummy) variable approach.

measurement we are able to get closer to the true constructs of interest. This type of dummy variable model extends this advantage to the estimation of means, and gets closer to the true difference between boys and girls on time spent on homework. Note two things, however. First, the results do not tell us the actual mean level on the Homework latent variable for boys versus girls. Instead, this path tells us the *difference* on the latent variable for boys versus girls. A positive value of, for example, 2 would mean that girls score two points higher, on average, on the latent homework variable, whereas a value of -2 would mean that boys score 2 point higher (and girls score two points lower). Second, the value is actually the difference in intercepts. Recall from regression what intercepts are: they are the predicted scores on the dependent variable for those with a value of 0 on the independent variable. So we can think of these as the mean differences for boys and girls on the Homework latent variable, controlling for the other variables in the model.

If we were interested in how homework differed for boys and girls we might also wonder if homework had differential effects on Grades for girls versus boys. In Chapter 18 we learned how to test for such interactions (moderation) using a multi-group (MG) approach. This approach is also illustrated in Figure 23.8, in which we would compare *path a* in the boys model with *path a* in the girls model (via parameter constraints). With the addition of the estimation of means and intercepts to this MG approach, however, we were able to test for both the main effect (effect of Sex on Homework) and the interaction (differential effect of Homework on Grades across the sexes) in one analysis. To do so, one first tells the program to estimate means and intercepts and makes various constraints across the models. The mean or intercept of interest is set to zero in one group and freely estimated in the other group. Thus like the dummy variable approach, this method is used to estimate the *difference* in means or intercepts across groups.

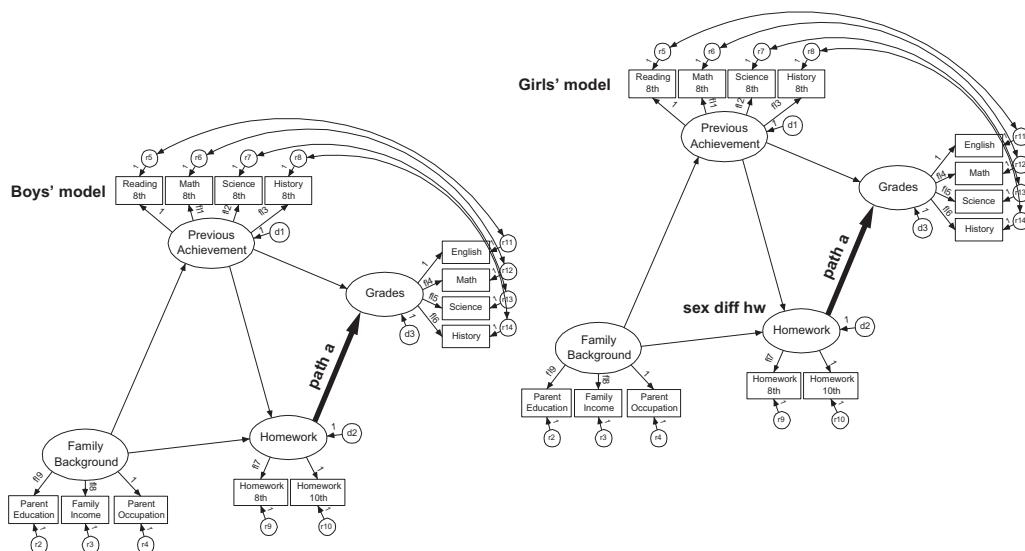


Figure 23.8 A multi-group approach for testing whether Sex moderates the effect of Homework on Grades. Differences in the magnitude of path *a* (tested via model constraints) for boys versus girls would suggest a differential effect.

The MG-MACS (multi-group mean and covariance structures analysis) approach is quite useful for the analysis of experimental data when the post-test (and pretest) are latent variables, and an example of such an analysis—the effects of hypnosis on hot flashes—was used to illustrate the method. Many programs (including Amos and Mplus) also require the analysis of mean structures when there are missing data in a raw data format.

CFA With Latent Means and Invariance

The analysis of latent means applies to confirmatory factor analysis, as well. To test for differences in factor means, factor loadings and the intercepts of the measured variables are constrained to be equal across groups. The latent mean for one group is constrained to zero, and the means for the other group (or groups) freely estimated. The latent mean obtained represents the difference across groups on the construct of interest, the variable underlying the measures.

Chapter 20 also presented the topic of measurement invariance, first broached in Chapter 18, in more depth. Metric (aka weak) invariance requires the factor loadings to be equal across groups, and is required when wish to compare variances and covariances of latent variables across groups. It is also required when we wish to compare effects, including paths from one latent variable to another, across groups. Intercept (strong) invariance requires that the intercepts of the measured variables are equivalent across groups. If strong invariance is obtained, it is then possible to compare validly the differences in latent means across groups. These and other possible steps involved in invariance testing (including both measurement and substantive comparisons) were detailed in the chapter. It is important to realize that measurement invariance is not just applicable to those interested in CFA. Measurement invariance is, in fact, assumed, but often untested, in most comparisons across groups. When we test for differences in means across two groups (e.g., in a typical ANOVA) we are assuming that intercept invariance of the dependent variable is plausible. When we test for differences in effects across groups (e.g., in an interaction analysis in MR or ANOVA) we are

assuming that metric invariance holds for the measures used. This chapter showed how these assumptions can be tested via invariance comparisons.

Latent Growth Modeling

We have been interested in change, loosely defined, throughout this text. One plausible interpretation for a regression coefficient is along the lines of “for each unit increase in X, Y will increase by so many units.” Yet for most of these regressions, no one really increased or decreased; instead we inferred such change based on comparing an individual at one level with another individual who was at another level. Later in the text we investigated longitudinal path models that examined the effects variables on some outcome controlling for previous scores on that outcome. Panel models became even more explicit in investigating change, examining the effect of time 1 variables on time 2 variables, and beyond.

With latent growth modeling (LGM) we are, for the first time, able to actually model the process of change over time. Consider if you had measures of the same variable for the same people at three or more points over time. Given the same underlying unstandardized scaling, it would be possible to derive two underlying factors from these repeated measures: one factor representing the latent starting point for the repeated measures, and one representing the latent growth in those measures. The initial level factor may also be thought of as a latent intercept, with the growth factor as a latent slope for the repeated measures. As with other latent variables, these initial level and growth latent variables would come closer to the true underlying construct than would the actual measures. This is the thinking underlying LGM. An example LGM model from Chapter 21 is shown in Figure 23.9.

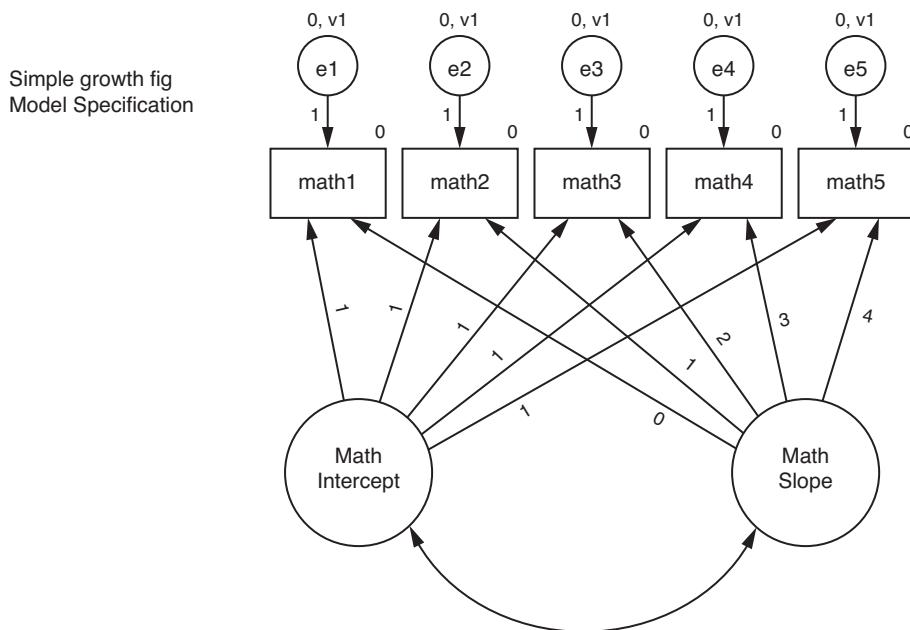


Figure 23.9 A latent growth model. The Intercept latent variable estimates the mean and variance in initial math level. The Slope variable provides an estimate of growth in math scores. As with other latent variables, these estimates are closer to the true constructs of interest than are the measured variables.

In previous chapters we have seen that multiple groups are required in order to estimate latent means and their differences across groups. With LGM we see that there is a way around this restriction: when the intercepts of the measured (repeated) variables are constrained to be zero, we can estimate the *means* of the latent initial level and growth variables. As a result with LGM we get estimates for the latent (true) mean and variance of the initial level of the repeatedly measured variable. We also get estimates for the latent (true) mean and variance of growth in that variable. In what is often then a second step in LGM analyses we can then examine variables that likely *influence* the initial level of the construct and variables that influence growth in the construct. Alternatively, we can examine the effect of initial level and growth in other variables. As a result, LGM allows the study of variables that influence change and the influences of change on other variables.

It is worth reiterating one more time that the fit statistics from SEM, while providing some feedback as to the quality of the model, are no panacea. In particular, the fit statistics do not help with the biggest dangers in nonexperimental research. They do not warn you when you have left out a common cause from your model. If you plan your overidentified model carefully, however, you can test hypotheses about whether paths are drawn in the correct direction.

ISSUES INCOMPLETELY OR NOT COVERED

Maximum Likelihood Estimation

In Part 1 we focused on least squares estimation; we showed that MR works by minimizing the errors of prediction around the regression line. SEM programs generally use a different approach by default, maximum likelihood estimation. Rather than minimizing errors, ML is designed to provide estimates that most likely would have resulted in the sample data. Simplistically, for each set of possible parameters, the probability that these estimates could have produced the data is computed. The estimates with the highest probability are used.

With simple, just-identified path models, maximum likelihood and least squares estimates are equivalent, and thus MR and SEM programs provide the same estimates. The two methods will also generally yield very similar results for overidentified path models. The interpretation of the coefficients is also the same.

Maximum likelihood estimation is the default for SEM programs, but other methods (e.g., generalized least squares) are also possible. For more information about maximum likelihood estimation, see Eliason (1993); for additional information about estimation in SEM, see Bollen (1989) or Loehlin & Beaujean (2017).

Missing Values

We broached the topic of missing values in SEM in Chapter 19 (Latent Means), but the topic is worth repeating and expanding here. In Part 1 I noted that two common methods of dealing with missing data in MR are listwise deletion of missing data (any case that has missing information on the variables used in the regression is not used in the analysis) and pairwise deletion of missing data (a case that has a missing value on a variable is not used to calculate the correlations with this variable, but the case is used to calculate other correlations). Currently, all SEM programs use a more sophisticated strategy for dealing with missing data, generally referred to as *full information maximum likelihood* estimation, or FIML (Arbuckle, 1996).

So what makes FIML (and other modern missing data methods, discussed below) better? Methodologists often differentiate possible missing data mechanisms (Rubin, 1976), which requires thinking about why the data are missing. Simply put, what causes missingness, and how is it related to the variables we are analyzing? First, data can be missing completely at

random (MCAR). This is the ideal missing data scenario in which the reason for the missing data is unrelated both to the values of variable that has missing data and to other variables in the model. Suppose, for example, you were interested in the effects of Homework on Grades, but in your survey not everyone reported their Grades. If the reason for the missing data was unrelated to participants' Grades and Homework, the data would be classified as missing completely at random (MCAR). This possibility is illustrated in the top of Figure 23.10. Here, the reason for the missing data is an unmeasured (and unknown, as it often is) variable. What matters is that the reason for missingness is unrelated to the value of Grades. (Please note that these are conceptual models designed to illustrate these missing data concepts with already familiar concepts; they are not models that you would actually analyze.) When data are MCAR, both traditional and modern methods of dealing with missing data provide

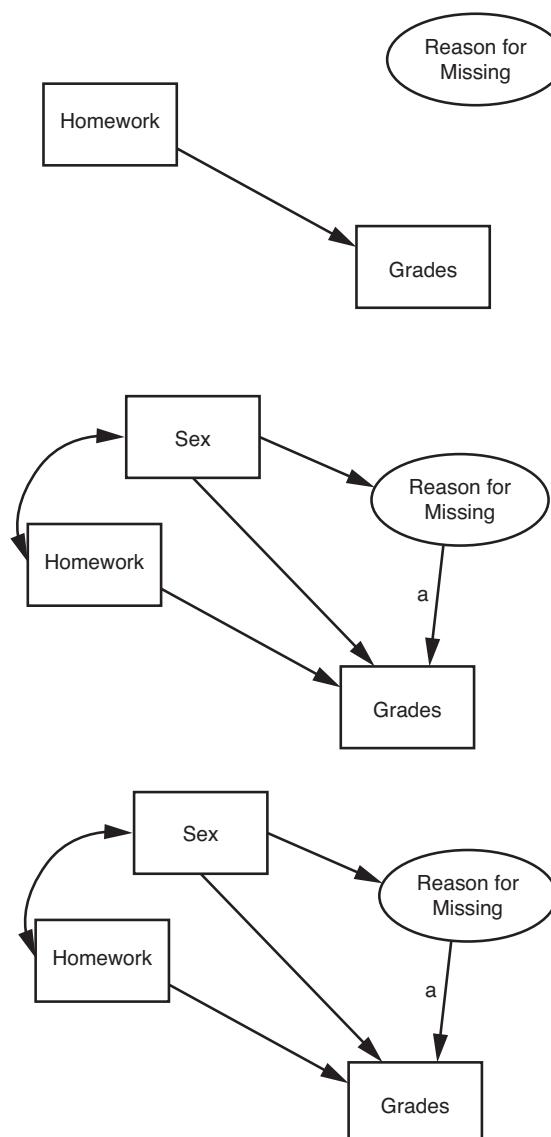


Figure 23.10 Path illustrations of MCAR, MAR, and MNAR mechanisms for missingness.

accurate estimates of means, covariances, variances, and effects in SEM models. Clearly, however, MCAR is a pretty strong assumption, and one that is likely unreasonable in much research.

Alternatively, data may be missing at random (MAR). In this situation, illustrated in the middle portion of Figure 23.10, the reason for the missing data may be related to the values of variable with missingness (e.g., those with lower Grades are less likely to report their Grades), but that relation disappears when other variables in the model are controlled (e.g., when Sex is controlled). In this scenario, perhaps boys are less likely to report their Grades than are girls, and that is the reason for the missing data on the Grades variable. However, once Sex is controlled, the reason for the missingness is no longer related to Grades; that is, path *a*, from Reason to Grades, reduces to zero. When data are MAR, FIML and other maximum-likelihood-based methods provide more accurate estimates of parameters (e.g., effects) than do traditional methods of dealing with missing data (e.g., pairwise or listwise deletion).

Note that if Sex were not in the model, if Sex were not controlled, then the path from Reason to Grades would not reduce to zero. For FIML to work with MAR, the variable that reduces the relation between the reason for missing and the outcome must be included in the model. Think of it this way: Sex is a common cause of the Reason for Missing and of Grades; common causes need to be included in the model for estimates of effects to be accurate. For data to be considered MAR, both the data and the model being estimated are important.

The lower model in Figure 23.10 illustrates data that are Missing Not at Random (MNAR). In this scenario, even after controlling for the other variables in the model, the Reason for Missingness is still related to the values on the variable of interest. In this case, perhaps (as illustrated in the Figure), the other variables in the model are not related to the reason for missing on Grades. Or (not illustrated), perhaps they are related (e.g., perhaps there is a path from Sex to Reason), but the effect of Reason on Grades is still meaningful even after the other variables are controlled. When data are MNAR, even FIML methods meant for MAR data will be inaccurate (as will other methods).

Several points are worth making here. First, we often do not understand the missing data mechanism. MCAR can be tested, but we often don't know whether we really meet the MAR assumption as opposed to MNAR. In addition, the MCAR test is affected by sample size. For these reasons alone, it is better to use the missing data routines of your SEM program as opposed to creating a matrix using listwise or pairwise deletion of missing data (or deleting all the cases with missing data in your file). In addition, because the reasons for missingness for some variables may be MAR and others MNAR, maximum likelihood approaches are generally preferred even when the MAR assumption is not clear-cut. There are other advantages in using the FIML method used in SEM programs, including the fact that FIML methods use all the data rather than a portion of the data. Second, it is important to include the variables believed to affect the missing data mechanism in the model and analysis. For this reason, missing data methodologists generally recommend the inclusion of "auxiliary variables" in SEM analyses (Enders, 2006; Graham, 2009, 2012).

Two other related methods for dealing with missing data are worth considering. The expectation-maximization (EM) algorithm is a method of obtaining maximum likelihood estimates with missing data, and it is available in general statistics software (e.g., in SPSS's Missing Values Analysis add-on). With EM, one generally estimates the variance/covariance matrix and means of a data set that has missing values and then uses this matrix in an SEM program. EM thus makes it easy to consider auxiliary variables. However, with EM, different sample sizes may need to be specified depending on missing data patterns and χ^2 estimates may need to be corrected (Enders & Peugh, 2004; Savalei & Bentler, 2009). With multiple imputation (MI), one creates multiple versions of a data set using maximum likelihood or Bayesian methods to impute the missing data. These multiple data sets are analyzed, and parameter estimates, standard errors, and fit measures summarized across the data sets. MI

requires multiple analyses for each model, however, making what is often an already complex analysis strategy even more complex (although it is often possible to automate these multiple analyses). Both Amos and Mplus, among other SEM programs, have the ability to perform MI, and it is also possible to do so in SAS and in stand-alone missing data analysis programs (e.g., NORM, Schafer, 1997, 1999, see <https://methodology.psu.edu/publications/books/missing>). Single imputation is also a possibility, and may be a good option when the amount of missing data is relatively minor (Widaman, 2006).

Unfortunately, when there are missing data, Amos, the program I have used primarily to illustrate SEM in this text, does not include many of the more detailed aspects of fit that we generally examine when first analyzing models (i.e., modification indices and standardized residuals), and other aspects of the program are also not available (e.g., bootstrapping and SRMR). One option is to use another program, such as Mplus, for the analysis; Mplus provides all of this information even when there are missing data. Another option is to use a method such as EM to generate a covariance matrix, or MI or single imputation to generate complete data sets, and analyze those in Amos. One option I have also used is to specify a model that only includes covariances (and means) among the measured variables to be used in an SEM or CFA model; the model has zero *df*. The implied covariance matrix produced by Amos using FIML can then be used as input in subsequent analyses. It seems to me that this approach can incorporate one of the advantages of EM (the ease of including auxiliary variables), but please note that I know of no studies supporting this approach. When I use one of these two-stage type approaches (estimate a matrix and then use the matrix to estimate the SEM model), I generally also go back and re-estimate the raw data and check to make sure that the results are the same.

Fortunately, the state of the art in missing data analysis methods has improved dramatically in recent years, and there are many resources for those who want to understand this topic in more depth. Enders has an excellent book on the topic (Enders, 2010), and Graham's review of the topic is also excellent (Graham, 2009; see also Graham, 2012). Widaman's 2006 article already referenced has some excellent analysis recommendations, some of which I have included and updated:

"Analysis recommendation 3: If the amount of missing data in the entire data set is very small, consider using single imputation" (Widaman, p. 61). Also consider FIML, given that simulation studies show FIML works well even with small amounts of missing data, and that the benefits increase as missingness increases (C. K. Enders, personal communication, April 11, 2014).

"Analysis recommendation 5: If the amount of missing data is moderate or large, but variables related to missingness can be included in analytic models, consider using FIML estimation" (p. 62); and

"Analysis recommendation 4: If the amount of missing data is moderate or large and the variables related to missingness cannot be included in all analyses, use multiple imputation (p. 61)." Multiple imputation may also be advised when the model includes both continuous and categorical variables (C. K. Enders, personal communication, April 11, 2014).

Whatever approach you choose, learn about and routinely use one of these modern methods (FIML, EM, MI) for dealing with missing data.

Planned Missingness

Most of what you will read on methods of dealing with missing data, including the section above, approach the topic as how to deal with a nuisance. And missing data are a nuisance, although a ubiquitous one for anyone who does research. But missing data—when planned

for—can actually improve research. Consider that if modern missing data methods allow the accurate estimation of effects when data are missing MCAR or MAR, then one can plan data collection so that data will be MCAR (or MAR) knowing that the data will be analyzed via FIML (or some other modern method). What this plan would allow is to then reduce the data demands on participants in the research. Consider a complex CFA or SEM, and the sheer amount of data that need to be collected on each participant. If the plan also calls for longitudinal data collection, the complexity and data demands multiply. Given a very complex or time consuming project, one has to wonder if those participants with complete data are unusual in some way!

It is possible to reduce data demands and improve such research by planning missing data in advance. Suppose you were interested in conducting a CFA across multiple intelligence batteries. If one set of participants took battery A and B, another set A and C, and another set A and D, with the test forms randomly assigned, the data would indeed be MCAR, and the data from all three measures could be combined into a single analysis with accurate estimation of effects (cf. Caemmerer, 2017; Reynolds, Keith, Flanagan, & Alfonso, 2012). See McArdle for an early explanation of this “reference variable” approach (McArdle, 1994). See Enders (2010) and Rhemtulla and Little (2012) for additional designs and developments.

Sample Size, Number of Parameters, and Power

In the summary for Part 1, we briefly reviewed issues of samples size and power in MR. MR, and to a greater extent SEM, are large-sample techniques, and one good rule of thumb is the more the better. But many students and researchers struggling to collect data often wonder about minimum sample sizes in SEM, just as in MR. In Part 1, we examined several of our analyses using a power analysis program to determine the sample sizes needed to have a reasonable chance of rejecting a false null hypothesis in MR.

For SEM research, MacCallum, Browne, and Sugawara (1996) showed how to calculate power for the RMSEA using the sample size and *df*, or how to calculate the sample size given the *df* and desired power (see also Hancock & French, 2013 and Kaplan, 1995, for a discussion of power). Briefly, the larger the sample size and the larger the degrees of freedom, the higher the power. Thus, complex, highly constrained models are more powerful than models with fewer *df*. As of this writing, Kris Preacher’s Quantpsy.org website (<http://quantpsy.org/rmsea/rmsea.htm>) has R utilities that can be used to compute power for RMSEA or the minimum sample size needed to achieve a desired level of power given a target level of the RMSEA and the *df* for the model. Of perhaps greater interest, because we usually want to compare various models to determine which fits better, another utility will tell you the sample size needed to detect a difference in RMSEA between two nested models (Preacher & Coffman, 2006, May).

Usually when we compare competing models, however, we use $\Delta\chi^2$, not ΔRMSEA . Loehlin & Beaujean’s (2017) text shows how to determine the sample size needed to detect, via $\Delta\chi^2$, the presence or absence of a particular parameter (e.g., a path). It is a pretty labor-intensive process, however. The Mplus website has script for conducting such analyses in Mplus (www.statmodel.com/power.shtml).

A common rule of thumb for SEM studies is that researchers should strive for a 20:1 ratio of sample size to the number of parameters to be estimated (the N:q rule, Jackson, 2007), An N:q ratio of 10:1 may be acceptable (Kline, 2016).

A related sample-size issue with SEM gets at the accuracy and stability of findings. A common rule of thumb is that SEM studies should include a minimum sample size of 100; this rule of thumb is based on simulation studies that show problems with results below this level (e.g., Boomsma, 1985; see also Loehlin & Beaujean, 2017, for a summary).

Another important consideration in SEM studies is the number of indicators per factor. Although here I have presented models that include two indicators per factor (to keep the models fairly simple), another good SEM rule of thumb is to try for three or more indicators per factor. This rule becomes more important with smaller sample sizes and when latent variables have low intercorrelations. Including more indicators should lead to more stable estimates of factors, and (because more indicators generally result in greater df) more power. For a dissenting view, however, see Hayduk, 1996.

These are a lot of rules of thumb, and some may give different answers! What's a poor graduate student to do? Try this: Draw your model and then conduct a power-analysis-derived estimate of the needed sample size. Also examine the needed sample size to detect plausible differences in RMSEA, with those plausible differences gleaned from previous studies. Does this sample size conform to the $N:q$ rule? Do these two methods together give you a do-able sample size? If the number suggested is below 200, can you obtain a sample size of 200? 150? Of course if you can obtain a sample size larger than 200, do so, because larger is better. Also, more complex models with more free parameters need larger samples. Finally, keep in mind that your model is not doomed if you cannot get as large a sample as these methods suggest, but it is better to be safe than sorry. Larger samples are safer.

Longitudinal Models

We have discussed several types of longitudinal models in this text. In addition to getting closer to studying the process of change, I have hinted that longitudinal models can help bolster your guesses about causal ordering by building in an actual time component in your analysis. If X is measured prior to Y , it is less likely that you are committing an error when you draw a path from X to Y . Another less obvious advantage of longitudinal models is that they may help control for the number one danger to causal inference, the omission of a common cause of our presumed cause and presumed effect. Consider the Homework models from Chapters 15 and 18 (See also Figure 23.7). For both, we controlled for previous achievement when examining the effect of Homework on grades. It is likely that by doing so we controlled for many likely common causes, because many influences on current grades or achievement will likely flow through previous achievement. Note that this thinking appears to be accurate with the homework models; the background variables in these models affected grades only by affecting previous achievement and homework. That does not mean that this will always be the case. Properly analyzed longitudinal models may reduce dangers inherent in nonexperimental research, but do not remove them.

It is important to realize that although longitudinal models can bolster claims of causal precedence, they are no panacea for the danger of confused time precedence. Consider, for example, if you measured Self-Concept in 2016 and (biological) Sex in 2018. Despite the longitudinal nature of the data collection, you will still be mistaken if you draw the path from Self-Concept to Sex. Sometimes logical time precedence takes precedence over actual time precedence.

Dynamic Modeling

With the introduction of latent variable panel models and latent growth models we got ever closer to modeling and testing notions about the process of change in variables over time. Dynamic modeling, also known as latent change score (LCS) modeling, takes these ideas further still by examining change scores as latent variables and can be used to test directly whether change in one variable plausibly leads to change in another. This topic, while fascinating, is beyond the scope of this book. For a brief introduction, see Ferrer and McArdle

(2010). For a cognitive development example that also includes a fascinating use of planned missingness, see Ferrer and McArdle (2004).

Formative Measures

All the latent variables we have discussed in this text have been what are known as reflective indicator models in which we have assumed that the latent variables are causes of the measured indicators. In such models, which are by far the most common latent variable models, we assume that there is an underlying factor that, in effect, partially causes the scores we get on the measured indicators of those factors. But in some instances, it makes sense to think of the arrows going in the other direction, from the measured variables to the latent variables. Is a variable such as GPA really best conceived as an underlying factor, for example? Or would it make more sense to simply think of GPA as the sum of its parts, a composite? If so, we would instead draw the measurement paths from students' GPAs in each course to a latent variable representing overall GPA. In this instance, GPA would be what is known as a formative (rather than reflective) measure. We discussed this possibility briefly concerning the Family Background latent variable. Once again, such a topic is beyond the scope of this text, but for a nice introduction, see Kline (2016). Such models can be tricky to estimate.

Categorical Variables

When we initially discussed SEM programs (Chapter 14) I noted that Mplus has sophisticated routines for analyzing categorical outcome variables. Indeed it is possible to have factors and other latent variables that are based on categorical rather than continuous variables. I believe all SEM programs have the capability of analyzing categorical as well as coarsely-ordered continuous variables (e.g., a three-choice Likert scale). Mplus seems to have the most options for such analyses. For more information on this topic and suggestions for analyzing such models (as well as very non-normal data) see Finney & DiStefano (2006).

Differences Across Programs

If you have run our examples on a software program other than Amos, you may have found minor differences in your estimates versus those presented in this book (see, for example, differences in output at www.tzkeith.com). One likely reason for this difference is that different programs calculate covariances differently. Amos, for example, uses N in the denominator (the maximum-likelihood estimate), whereas LISREL uses $N - 1$ (the unbiased estimate). The differences should be minor, however, especially with large samples. If you get substantially different results, double-check your analyses, because one of us is in error!

Causality and the Veracity of Models

It is fitting to end this section with one more discussion of causality, a fascinating topic. I have tried to find a middle ground on the issue of causality and the degree to which we can make valid inferences of causality with nonexperimental research methods. No doubt some readers will think I've gone too far, overstating the degree to which we can make such inferences. Others will think I've understated the case. This issue will continue to be debated and is certainly not settled in this text. Nevertheless, you should be aware of some fascinating developments in this realm. Pearl (2009), and colleagues (Pearl, Glymour, & Jewell, 2016; Pearl & MacKenzie, 2018) for example, have detailed advances in understanding and demonstrating causality; see also Shipley (2000) for some of these issues translated to biology. As in Part 1, my thinking is that we should use causal language (e.g., examining the effects

of this variable on that outcome), but that we should be obvious about what we mean by that language. It may be useful to add a statement like this to your research write-up (slightly modified from Chapter 8):

It is important to note that the data used in this research are nonexperimental in nature; there will be no (nor could there be) experimental manipulation of depression to determine its subsequent effect on achievement. As a result, it should be understood that all statements that discuss the “effect” of one variable on another, or that focus on variables that “explain” an outcome are dependent on the validity of the model. In other words, if the model is a reasonable representation of reality, the estimates resulting from the model indeed show the extent of the influence of one variable on another. If the model is not a reasonable representation of reality, the estimates are not accurate estimates of those effects.

At the same time, we should always be attuned the things that will help us avoid omitting an important common cause, including theory and previous research, along with, perhaps, the use of longitudinal data and models. Likewise, nonequivalent overidentified and longitudinal models can help us avoid (or even test) problems in incorrect causal ordering, and should be exploited.

ADDITIONAL RESOURCES

These last few chapters have provided an introduction to SEM, perhaps just enough to make you dangerous. To become well-versed in conducting SEM studies you should get experience conducting such studies, supplemented by further reading. I hope you have enjoyed this adventure into the fascinating world of nonexperimental analysis via SEM (and MR). I also hope you will experiment with these methods and seek to develop your initial skills more completely. The sources listed next are good starting points.

Introductory Texts

I have mentioned several introductory textbooks that are worth your review:

Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.

Loehlin, J. C., & Beaujean, A. A. (2004). *Latent variable models: An introduction to factor, path, and structural analysis* (5th ed.). New York, NY: Routledge.

See also:

Maruyama, G. M. (1998) *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner’s guide to structural equation modeling* (4th ed.). New York, NY: Routledge.

For an excellent, historically oriented annotated bibliography of path analysis and SEM literature, see:

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.

More Advanced Resources

If you want to advance your knowledge about SEM beyond the basics, I recommend the journal *Structural Equation Modeling*, published by Taylor & Francis. You may also be interested in joining the SEMnet listserve. For information, go to www.gsu.edu/~mkteer/

semnet.html or, for the archives, <https://listserv.ua.edu/archives/semnet.html>. Some worthwhile books include:

- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley. (a classic reference text)
- Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Hancock, G. R., & Mueller, R. O. (2013). *Structural equation modeling: A second course* (2nd ed.). Charlotte, NC: Information Age.
- Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. New York, NY: Guilford.
- Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Los Angeles, CA: Sage.
- Marcoulides, G. A., & Schumacker, R. E. (Eds.). (2001). *New developments and techniques in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman & Hall/CRC.
- Schumacker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interactive and nonlinear effects in structural equation modeling*. Mahwah, NJ: Erlbaum.

For more depth on the topic of CFA:

- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed). New York, NY: Guilford.

For more depth on the topic of longitudinal analysis:

- Little, T. D. (2013). Longitudinal structural equation modeling. New York, NY: Guilford.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60(1), 577–605. doi: 10.1146/annurev.psych.60.110707.163612

For more depth on the topic of LGM:

- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NY: Wiley.

- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and application* (2nd ed.). Mahwah, NJ: Erlbaum.

Books About Specific SEM Programs

Several texts are program specific and are valuable if you want to go beyond the examples presented in the user's guide to your program:

- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York, NY: Routledge.

- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.

- Byrne, B. M. (2010). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

- Byrne, B. M. (2016). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (3rd ed.). New York, NY: Routledge.

Reporting SEM Results

SEM results are obviously complex, and writing up SEM results is often a challenge. You need to provide enough detail so that other researchers can reproduce your results, but it is easy to

go overboard and report too much detail, resulting in a research report that is too long and uninteresting. How do you decide what you should report? First, model exemplary research in your area of interest. Then, turn to these references:

- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73, 3–25.
- Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, 7, 461–483.
- Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82.

Cautions

Finally, several references to remind you to be cautious in your use and reporting of SEM results:

- Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18, 115–126.
- Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics*, 12, 101–128.
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100, 107–120.
- 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.

In Chapter 18 we did a quick review of the dangers of MR, path analysis, SEM, and non-experimental research in general. Another danger of latent variable SEM is related to its complexity. Like all such methods, it is open to abuse. This section of the book includes a basic introduction to SEM and by no means make you an expert.

If you have read, understood, and worked through this section of the book, you should be fairly well equipped to be a good consumer of SEM research. You understand the primary dangers of nonexperimental methods and SEM; these dangers (e.g., omitted common causes) constitute the most likely serious problems with nonexperimental and SEM studies that you will encounter. You also have a beginning understanding of what you, and the researchers, should be looking for in SEM studies.

It is easy, however, to suspend critical judgment as you read research using complex statistical methods. The authors, after all, are the experts. Can't we assume they know what they are doing? Don't be "seduced by sophistication" (Wampold, 1987, p. 311). Yes, it may be harder to be a savvy consumer of SEM research, but it is just as necessary as with other research methods. Of course, you should also keep in mind that there is no perfect research; no study is immune to criticism, and your standard should not be unrealistically high.

These cautions are even more important for those conducting SEM research. SEM is not magic. No matter how sophisticated our analyses, they cannot turn bad data into good or a poor design into a powerful one. Even SEM cannot create a silk purse from a sow's ear. As with other research methods, "the manipulation of statistical formulas is no substitute for knowing what one is doing" (Blalock, 1972, p. 448).

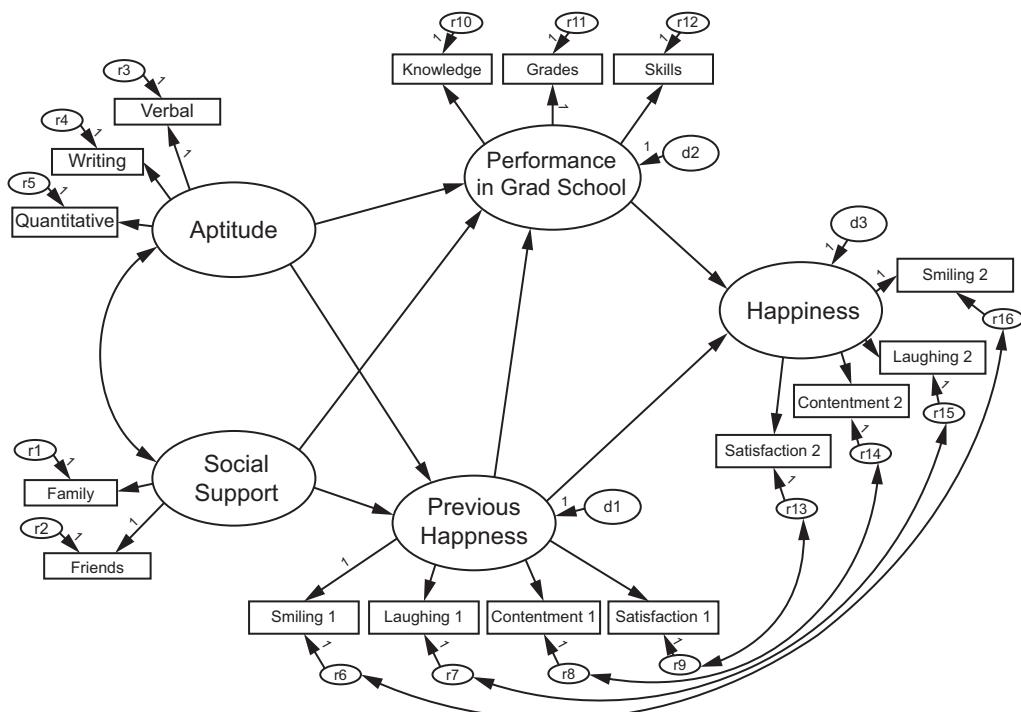


Figure 23.11 Happiness is a latent variable.

For those who wish to conduct SEM research, this section may have provided just enough information to make you dangerous. I hope these chapters have excited you about the power and possibilities of SEM and to try out the method to test research questions of interest to you. If you want to play around with the method, you should, but I encourage you to work with someone who is more knowledgeable and experienced. If you plan to use SEM on a regular basis, further reading is needed. I hope this section will guide you in that process. Be vigilant! But don't let the need for vigilance deter you from exploring further. SEM is a fascinating and powerful methodology. Experiment with it!

I once made a t-shirt with the caption “Happiness Is a Latent Variable”; an updated version of the accompanying model is shown in Figure 23.11. I trust by now you understand the various meanings of this statement. At the most basic level, in the model the variable Happiness is, in fact, a latent variable. More broadly, in the real world, happiness is a latent variable: it's not something we can measure exactly, but we do get indicators of it from many different behaviors. Finally, the statement is meant to say something about latent variable SEM. It is challenging, humbling, fascinating, and satisfying. I hope you experience some of the same enjoyment I have from learning and applying the method!

Appendix A

Data Files

My Web site that accompanies this book (www.tzkeith.com) includes the data used as research examples throughout the book. These are in the “Data Files” tab on the window. There you will find a zip file for each chapter in the book that uses data. The data sets that are used for a single chapter are included in the file that corresponds to the chapter in which they appear. Most files appear in several formats. Raw data files are available as SPSS “.sav” files, and are also generally available in some other format, as well, including Excel format (“.xls” or “.xlsx” files), and as plain text files (usually with the extension “.txt” or “.dat”). If you can use the SPSS files, they generally have the most information (e.g., value labels, missing values). I believe most general statistics programs can read SPSS files. My second choice for these raw data files would be to use the Excel files (except for the large NELS dataset).

Research examples in Part 2 of the text use a mix of raw data and matrix data (means, *SDs*, and correlations). Suggestions for raw data, above, apply to the raw data files for Part 2 chapters. Matrix files, analyzable by all SEM programs, generally appear both as Excel files and as SPSS files. Plain text versions (useful, for example, by Mplus) of all data sets are also available. These usually use the extension “.dat” or “.txt.”

NELS Data

There are several zip files corresponding to the NELS dataset used throughout the book. There is the primary NELS dataset, including 1,000 cases and over a thousand variables. This is labeled “NELS data.zip” There is also a shorter version of this file all of the variables that includes only those variables used in various analyses throughout the book, or used to create variables used in those analyses (around 100 variables). This zip file is listed under the heading “Shorter version of the NELS data.” Finally, there is a data dictionary that includes detailed information about all the variables in the larger file. The data zip files include the data in multiple formats. Your first choice for analysis, if you can use it, should be the spss .sav files (e.g., “n=1000, stud & par_3.sav”).

For the primary (large) data, I have converted the original SPSS file into several different formats: SYSTAT, SAS Transport, and plain text. The conversions were done using the program DBMS/COPY. If you can use the SPSS file, I recommend doing so because it is the original form of the data. The file is also saved as an SPSS portable file (extension .por). The SYSTAT and SAS files are also clean and easily usable (although users of both programs

should be able to use the SPSS file as well). This larger version of the dataset allow you to explore your own research questions of interest.

For the smaller dataset, there is an SPSS .sav file,, Excel files, and raw data files. One version of the raw data has all blanks set to -999, and the word file includes syntax to use these data in Mplus.

For more information about the NELS data, including how to obtain the full dataset, visit the National Center for Education Statistics Web site (nces.ed.gov/surveys/nels88/). While you are there, check out the other data sets you can get access to, including more recent data collection efforts than NELS. The ECLS data include Kindergarteners first surveyed in 2010 and (as of this writing) followed through fifth grade. My one criticism of some of these more recent data collection efforts is that they do not always include a full range of test scores. Nevertheless, you'll be amazed at all the data available to you!

The variable labels for all the variables in the primary NELS file are listed in the searchable word file “nels by ffu vars.docx.” The table of variable names, positions, and variable labels was created using the DISPLAY LABELS command in SPSS. A quick perusal of these labels should give you an idea of the power and scope of these amazing data. Variable names that start with BY are from the base year, when the students were in the 8th grade. The prefix BYS means the variable is from the student file; BYP means the variable is from the Parent File. Variables that start with F1 are from the first follow-up, when students were in the 10th grade. Composite variables created by NELS generally do not have the S or P designation. Composites that I created for various purposes start with variable 1379, ParentEd.

The abbreviation R in the variable labels refers to the respondent. Thus, the variable BYS8A, labeled R LIVES IN HOUSEHOLD WITH FATHER, means that the respondent lives in a household with his or her father. The name shows that this is a base year (BY, 8th grade) student (S) variable. The variables are listed in the order in which they appear in the dataset. If you want, once you get the data in your statistics program, you should be able to arrange the variables in alphabetic order.

Appendix B

Review of Basic Statistics Concepts

This appendix is intended as a brief review of some basic statistics concepts that are assumed in this book. It skims the surface of a broad range of material and is intended as a conceptual overview and memory jogger, not an in-depth treatment. If you need more background or review, a number of excellent introductory textbooks are available. One of my favorites is Howell's *Statistical Methods for Psychology* (2013).

Why do we need statistics? You may have wondered about that as you signed up for or sat in a statistics course, but reconsider the question now. Suppose you were to conduct an experiment in which you examined the effect of a specific type of therapy on the depressive symptoms of depressed adolescents (compared to those in a no-treatment control group). Assume that you used random assignment to treatment groups and that the random assignment was effective. After six months of treatment, you collect data on a measure of depression. Why calculate statistics? Why not just eyeball the data to determine whether your treatment worked?

If almost every person in the experimental group performed better on this posttest than did every member of the control group, there is indeed no reason to calculate statistics. You simply graph the data (e.g., Figure B.1) and any reasonable person will agree that you have demonstrated the efficacy of the treatment. Your data will pass the “interocular trauma test”; the data will hit you between the eyes.

Social science research is rarely this clear, however. What is more common is considerable overlap between the two groups so that reasonable people eyeballing the data will likely disagree as to whether the treatment was effective or not (e.g., Figure B.2). That’s why we need statistics: to help us determine whether the difference between groups is big enough, unusual enough, so that we can say with assurance that the treatment worked or, more generally, that the relation between two variables (in this case, treatment and outcome) is large enough so that we can assume it did not happen by chance.

This sort of reasoning is related to the notion of null hypothesis significance testing. A little more formally, when we test to determine whether two groups are statistically significantly different, the underlying logic goes something like this. First, we assume that in the population the groups are not, in fact, different. We then calculate the statistic of interest (e.g., the difference between the two means) and ask this question: if, in fact, the two groups are not different in the population, what is the probability of getting a difference this large by chance alone, given the size of the sample? If the chance of getting a difference that large is, say, 25%, few researchers will be willing to say the groups differed; that is, they will say you could not

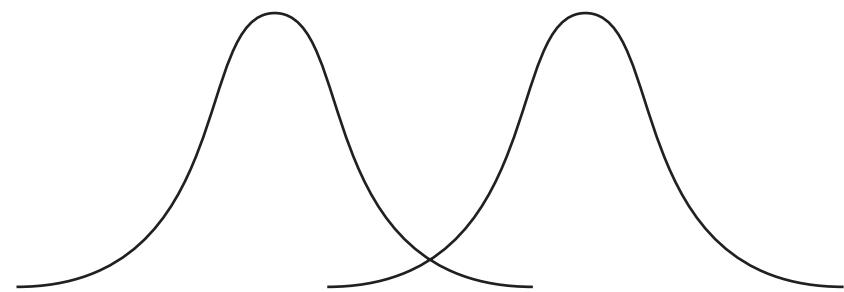


Figure B1 A large difference between groups. If data always looked like this, we'd rarely need statistical significance tests.

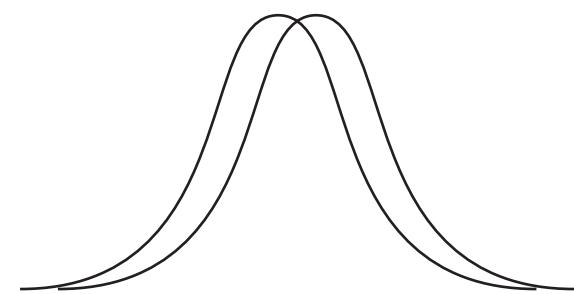


Figure B2 Data more commonly look like this. Without statistical tests, it's difficult to tell whether the two groups are really different.

reject chance as the cause of the differences between groups. Many researchers require that we obtain a difference large enough so that it occurs less than 5% of the time by chance. Others require that the difference be large enough so that it occurs only 1% of the time by chance. These rules of thumb are, of course, the values of $p < .05$ and $p < .01$, commonly used as benchmarks for deciding that something is statistically significant. When we say something is statistically significant, what we are saying is “it would be very unusual to get this kind of difference if chance variation were the only thing going on. We'd only get a difference this large five times (or one time) out of a hundred. Therefore, there must be something other than chance operating. Because I used random assignment to form groups, that other thing can only be the treatment; therefore, I can conclude that the treatment probably worked.” The same logic applies to tests of correlation coefficients, regression coefficients, F values, and so on.

This logic is not always pretty and has been criticized for years (e.g., Cohen, 1994), but it works fairly well and has served the social sciences well. It definitely should be augmented, however, with a focus on *effect sizes* and *confidence intervals*, as discussed later in this appendix and throughout this book.

BASIC STATISTICS

Mean

How would you describe a set of scores? Suppose your professor tells you that you got a score of 123 on an exam for which the total possible points were 140. Would you be happy or upset? Unless your professor always uses a 93%-equals-an-A type scale, you'd probably want more information. You'd want to know what the average score was on the test. In statistics, we generally define “the average” as the mean. The mean of a set of measures is the simple

arithmetic average: Sum the scores and divide by the number of scores to get the mean. Here, I will symbolize the mean using the symbol M .

Variance and Standard Deviation

Suppose the mean of scores on this test was 110. Now are you happy or disappointed? Probably happy; your score of 123 was above the mean, but how much above the mean? Is your score just above the average or well above the average? In addition to needing some idea of what the average is, you need some idea as to what the variability is in a set of scores. If 99% of people in the class scored between 100 and 120, you'd probably be pretty happy; you scored well above the mean. The range of scores is indeed one measure of variability, but in statistics we more commonly use the variance as the measure of variability in the set of scores.

Conceptually, the variance (V) is the average, squared, variation in a set of scores. Subtract the mean score from every score in the set. If you sum this number [$\sum(X - M)$], you get a value of zero because the negative values for those who scored below the mean cancel out the values for those who scored above the mean. To get around this problem, we can square each deviation prior to summing and then, to get the average, divide by the number of scores:

$$V = \frac{\sum(X - M)^2}{N}$$

In fact, with variance, as with many statistics, we generally divide by the number of scores *minus 1* ($N-1$), rather than N . The reason is that we are generally calculating a *sample* variance, and using $N-1$ gives us a better estimate of the population variance than we get using N . The new formula is

$$V = \frac{\sum(X - M)^2}{N - 1}$$

The variance, although useful, is not in the original unit of measurement, because we had to square the deviations from the mean to calculate the variance. To convert back to the original metric, it is easy to take the square root of the variance; this new measure of variability is referred to as the standard deviation ($SD = \sqrt{V}$). The standard deviation is useful in measurement because it is a measure of variability that is in the original units of measurement. If you know that your score on the statistics test was 123 and that the mean and SD of the test were 110 and 5, respectively, you now know that you scored more than 2 standard deviations above the mean.

z scores are scores transformed into standard deviation units. If my score was 2 standard deviation units below the mean, my z score will be -2 ; a z score of 1.5 means a score $1\frac{1}{2}$ SDs above the mean; a z score of zero corresponds to a score exactly at the mean. z scores are the parent of all other types of standard scores, and you can easily convert from z scores to other types of standard scores.

Distributions

As you know, many natural and social phenomena have frequency distributions that conform to a normal, or bell, curve. Figure B.3, for example, shows the frequency distribution for students' scores on the base year Science test in the NELS data on the accompanying Web site (www.tzkeith.com). Each bar on the histogram represents a 2.5 point range of scores on the test, and the height of the bars represents the number of students with scores in this range. As you can see, the data conform fairly closely to the normal curve superimposed over the histogram.

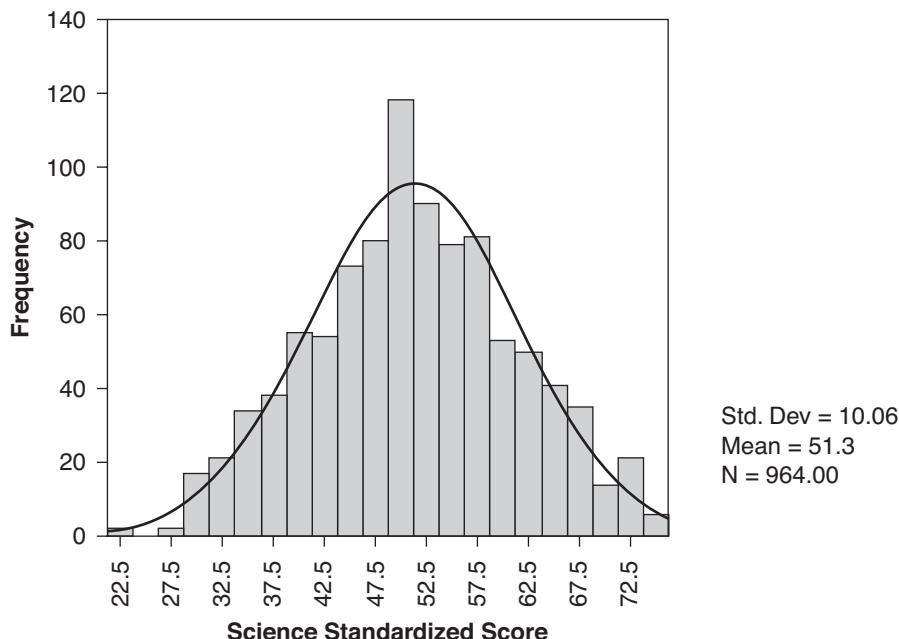


Figure B3 Frequency distribution of scores on the Base Year Science test from the NELS data. The data conform well to a normal curve.

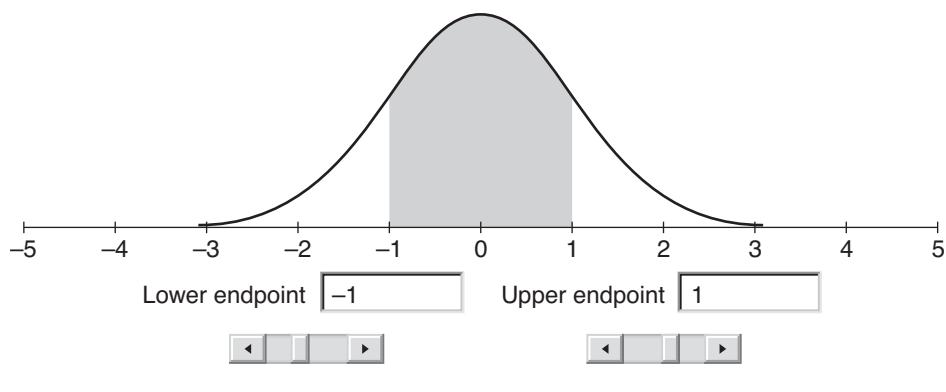


Figure B4 Sixty-eight percent of cases are between negative and positive 1 SD around the mean in a normal curve.

When data conform to a normal curve, this curve can be described fairly accurately using the mean and standard deviation of the data. [You can improve this description further by focusing on skew (whether the distribution has an extended tail in one direction or the other) and kurtosis (the flatness or peakedness of the distribution.)] When data conform to a normal curve, there are also well-defined relations between the distribution and statistics that describe the distribution. So, for example, approximately 68% of people will score between 1 SD above the mean and 1 SD below the mean (as shown in Figure B.4); approximately 95% will score between 2 SDs below and 2 SDs above the mean (Figure B.5), and so on. You can also use this information to determine the percentile rank of a particular score.

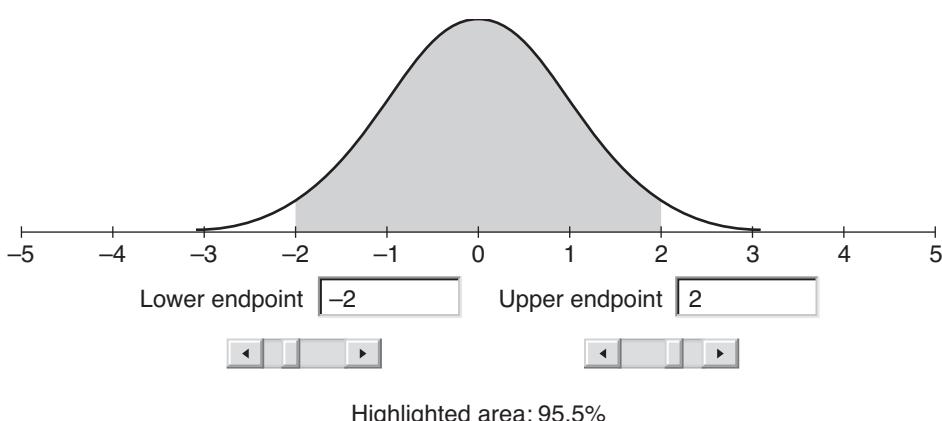


Figure B5 ± 2 SDs around the mean encompass approximately 96% of cases in a normal curve.

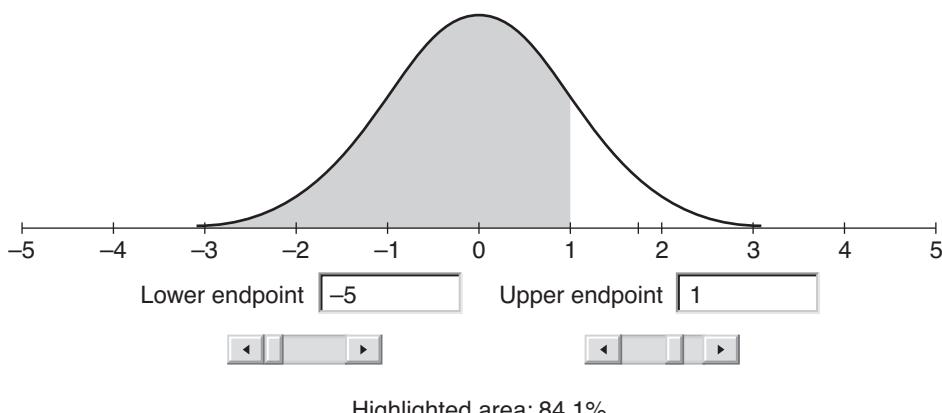


Figure B6 One SD above the mean corresponds to the 84th percentile. That is, 84% of people score at or below a standard deviation above the mean.

Say you scored 1 SD above the mean on another test. Fifty percent of people score below the mean, and 34% score between the mean and 1 SD above the mean (68% divided by 2). You therefore scored higher than 84% of people on this test ($50 + 34$). This information is summarized in Figure B.6. (These normal curves were drawn using Phillip. B. Stark's SticiGui tools, at stat www.stat.berkeley.edu/~stark/SticiGui/index.htm, using the normal probability tool.) Thus it appears that your score on the statistics test—more than 2 SDs above the mean—was very good!

This is nice, but you may want to know the percentage corresponding to, say, 1.75 SDs above the mean. Alternatively, you may want to know where on the normal curve (in standard deviation units) a score at the 98th percentile will be. For this purpose, you can turn to the z distribution or a z table (or a variety of tools available on the Web). There you can look up a z of 1.75, and you will find a value of .9599, meaning that a z of 1.75 is higher than 95.99% of other scores. In this book, I have encouraged you to use electronic versions of such tables. For example, click on a cell in Excel and then click on “Formulas” and “Insert Function.” Find the function called “NORMSDIST” (for normal, standard distribution) and use

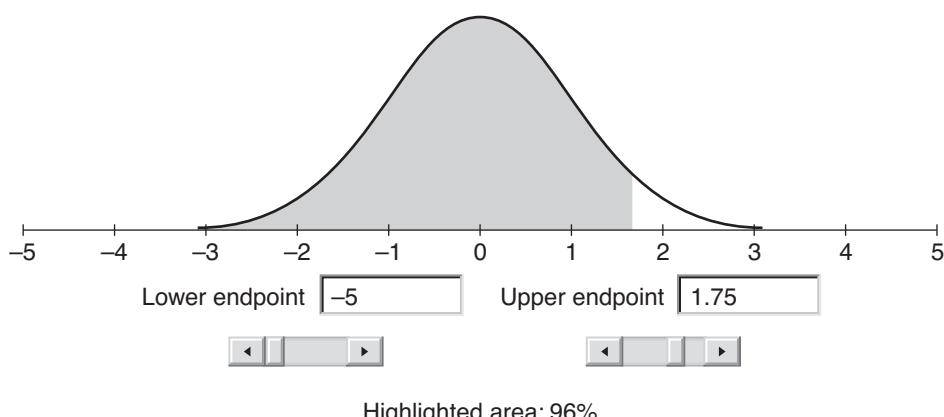


Figure B7 The 96th percentile corresponds to 1.75 *SDs* above the mean.

it. Type in 1.75 and Excel will return a value of .9599. I also recommend the SticiGui tools mentioned above (see Figure B.7). Return to your score of 123 on the statistics exam, which was 2.6 *SD* above the mean. What is the corresponding percentile rank? According to Excel, this value corresponds to a percentile rank of 99.53. Nice work!

Standard Error

Suppose you take a random sample of five cases from NELS for the base year Science test scores and compute the mean of these five scores. Will the mean be identical to the mean for the full sample of 1000? No, it will vary to a certain extent, because we took a small sample from a larger group. The first time I did this the mean of the five cases was 53.50, the second sample of 5 had a mean on the Science test of 51.47, and the third sample had a mean of 47.55. If I do this over and over, what do you think we will find? If we plot these means in a frequency distribution, what will it look like? If you answered “a normal curve,” good for you! Yes, we will get a normal curve of means. The frequency distribution and normal curve for 200 such samples is shown in Figure B.8. It’s more narrow than the normal curve of original scores because means are more stable than and vary less than individual scores. The *SD* of the normal curve of means was 4.59 versus 10.6 for the distribution of individual scores.

The reason for this exercise is that when we select a sample from a population we are assuming that the sample information (in this case, the sample mean) reflects the population. You can see from the histogram, however, that this assumption is sometimes more accurate and sometimes less accurate. Not all of our five-person-sample means were close to the overall mean. What is interesting is that the *SD* of this distribution of means provides useful information about the amount of variability (the amount of error) in the distribution of means. A narrow curve with a small *SD* tells us that most of our samples provide fairly accurate estimates of the real mean. A wide curve with a large *SD* tells us that many of our estimates will be error laden. Because this standard deviation reflects the error likely inherent in any estimate of the mean, it has a special name: the standard error of the mean.

In practice, we don’t repeatedly take smaller samples from a larger population. Instead, we can estimate the standard error from the characteristics of a single sample and the size of the sample. Other things being equal, the larger the *n* for each subsample, the more narrow the normal curve of means. Thus, as sample size increases, the standard error decreases. In addition, we can estimate the standard error of many different statistics, regression coefficients, for example, and use this information to test these parameters for statistical significance.

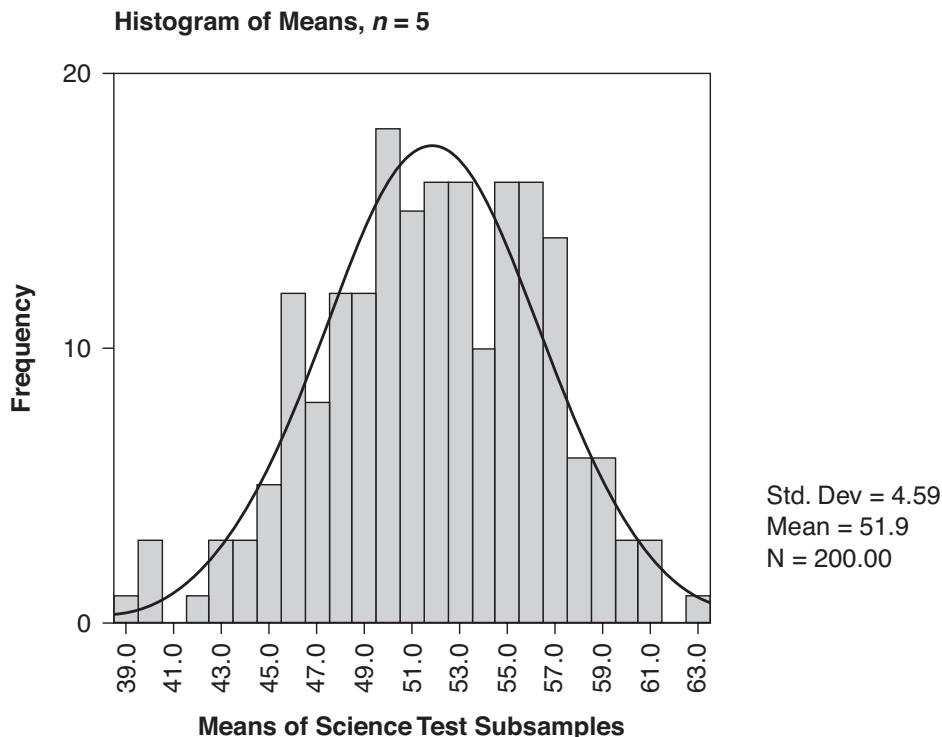


Figure B8 Means of 200 random samples of five cases each, Science test, NELS data.

Confidence Intervals and Statistical Significance

Because our normal curve of means has the same properties as other normal curves, we can apply our knowledge of normal curves to this one. Because 68% of cases are between -1 and $+1$ SD around the mean, we know that 68% of the means in our sample of means are between 47.31 and 56.49 (the overall mean of means \pm the SE, or 51.9 ± 4.59). Now, if we were to sample a single mean only, we could use this information in reverse. Our first mean that we sampled was 53.50. I could add and subtract the SE from this value and make a statement about the plausible values of the overall mean, something like “we can be 68% confident that the range 48.91 to 58.09 ($M \pm SE$, or 53.50 ± 4.59) includes the actual (population) value of the mean.” This use of the SE is called the confidence interval (CI), and in this case the 68% confidence interval. Another possible interpretation is “if we were to collect samples from this population and calculate confidence intervals repeatedly, around two-thirds of those CIs would include the true value of the mean” (both interpretations based on Cumming & Finch, 2005; see also Cumming et al., 2012). (In reality, if we sample a single mean, we will likely get a slightly different estimate of the SE of the mean, but we will continue to use the value 4.59 for this illustration.)

Sixty-eight percent isn’t the most convenient number to use. It would be more convenient to talk about 90% confidence (or 95%). Because we know the properties of a normal curve, however, it is easy to make this transformation. To encompass 90% of the curve, we multiply the SE by 1.645; to encompass 95% of the normal curve around the mean, we multiply the SE by 1.96. As already noted, we call these bands of error around our means (or any other statistic) *confidence intervals*, the 90% confidence interval (CI), or the 95% CI. CIs are extremely useful for giving you an estimate of the plausible range of a parameter and how error laden our estimate of the parameter likely is.

The standard error of a statistic can also be used to test its statistical significance in a variation of the *t* test. You may be most familiar with the *t* test as a test of differences between group means (discussed later in this appendix), but the *t* test is also a general statistical formula, in which a statistic (e.g., a regression coefficient) is divided by its standard error: $t = \text{statistic}/SE_{\text{statistic}}$. We can test all sorts of questions with the *t* test. The *t* in the *t* test is actually a series of distributions depending on sample size (with large sample sizes, the *t* distribution mirrors the *z* distribution), and we can look up in a table (or Excel or a probability calculator) the probability of obtaining a given *t* with a certain sample size. If the probability of obtaining a *t* by chance is small (say less than a 5% chance), we say that the parameter is statistically significant.

Degrees of Freedom

Most statistics that we use are accompanied by degrees of freedom (*df*). Conceptually, degrees of freedom are what the name suggests, the degree to which a given parameter is free to vary. Return to the example where we drew five cases from the NELS data (the Science test) and calculated a mean. The values of the five cases were 45.23, 47.66, 47.38, 60.39, and 66.84, and the mean was 53.50. Given the value of this mean, how many of the five cases could have different values? Say the first value was 44.23 instead of 45.23. Could we still get the same mean? Yes, we could if, for example, the final value is 65.84 instead of 66.84. In fact, four of the five scores ($N-1$) could change and we could still get the same mean (by adjusting the final score). This is the essence of degrees of freedom—how much maneuvering room you have in your data or the number of independent pieces of information in your data—and the reason we often use $N-1$ in formulas instead of N . As you will see in Part 2, SEM is an exception to the use of N for calculating degrees of freedom.

Correlations

Correlation coefficients describe the degree to which two variables are related, that is, the degree to which they are co-related. Correlation is one of the most fundamental concepts in statistics, and it underlies everything presented in this text. But think for a minute, if the correlation coefficient did not exist, how could you come up with such an index, a single number that accurately describes the degree to which two variables are related?

You'd probably start by graphing the two variables together. Figure B.9 shows a scatterplot, a graph of a group of high school students' scores on two variables: scores on an intelligence test and scores on an achievement test. The intelligence test has a mean of 100 and a *SD* of 15; for the achievement test $M=50$, $SD=10$ (these are common scales for such measures; the first is known as a deviation IQ scale, the second is a *T* scale.) Note the data point in the upper-right corner of the plot. This point belongs to the 24th individual in the dataset; that person obtained a score of 145 on the Intelligence test (the horizontal or *X*-axis), and a score of 86 on the Achievement test (the vertical or *Y*-axis). Each other data point represents one person's scores on the two measures. Would you say that these two variables are fairly highly co-related? Yes; it is apparent that people who obtain a high score on the Intelligence test also generally earn a high score on the Achievement test, and those who score at a low level on one test generally score at a low level on the other test. This, then, is one aspect we might look for in a correlation coefficient: it should tell us the degree to which the rank order stays the same for the two variables, whether high scores on one variable are matched with high scores on the other, and so on.

We could get a little more sophisticated and a little clearer, however, by making sure our two variables are on the same scale. Figure B.10 shows a scatterplot of the same two variables, after converting them to *z* scores. Now the two scales are directly comparable. And we can now ask the degree to which the *z* scores stay the same on the two tests. This reasoning is

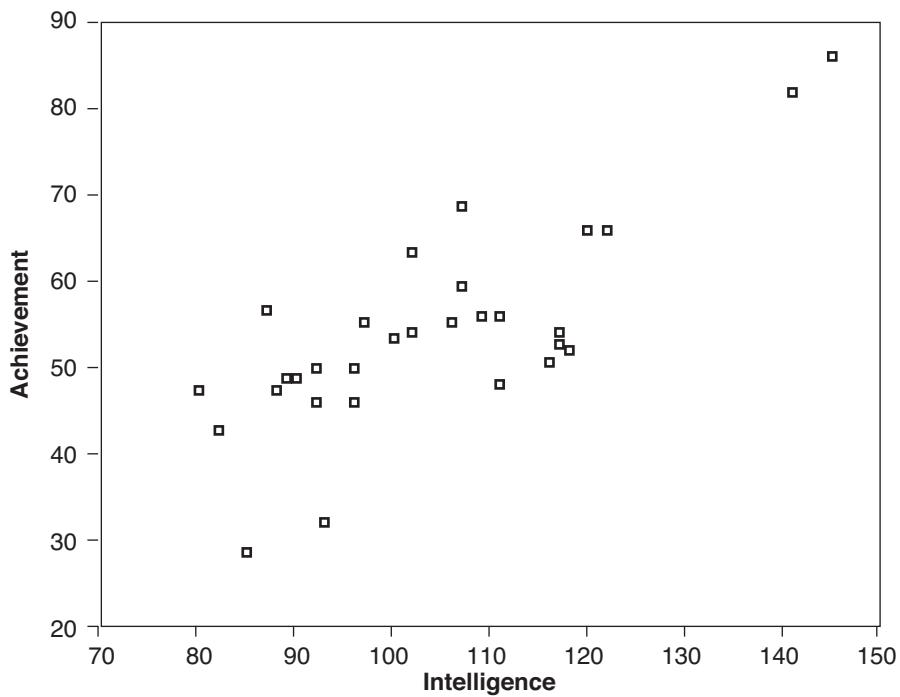


Figure B9 Plot of 30 people's scores on an Intelligence and an Achievement test. The scatterplot shows that the two tests are closely related.

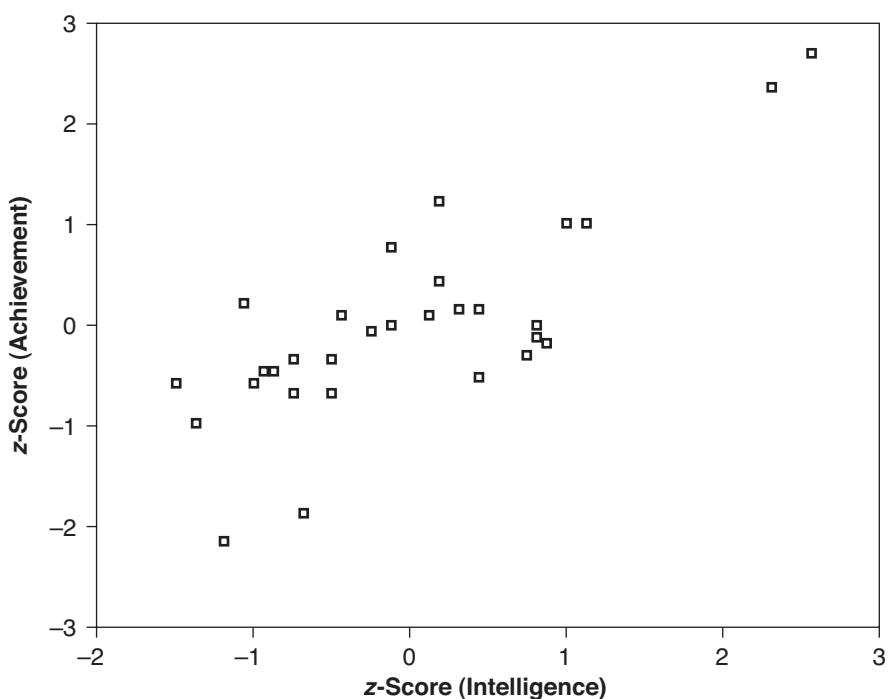


Figure B10 Plot of the Intelligence–Achievement data in z score format.

likely similar to that of Karl Pearson when he invented what we now know as the Pearson product moment correlation: to what extent do the z scores for two different measures stay the same versus the extent to which they differ?

Given this description of the correlation coefficient, let's develop a formula. The kind of formula we're talking about would index the degree to which two sets of z scores stay the same versus differ. In other words, we are interested in the *average* difference in these scores. We could subtract each person's z score on one instrument from his or her z score on the other to get an idea of this degree of difference. Then square these values (because if we were simply to sum them, the negatives would cancel out the positives), sum them, and divide by $N-1$ (see the discussion of variance for the use of $N-1$ instead of N). Our formula is now

$$\sum \frac{(z_x - z_y)^2}{N-1}$$

(Cohen et al., 2003). If you calculate this coefficient for the data shown in the scatterplot, you obtain a value of .44 (the data for this example are contained in the files "IQ Achieve.sav" and "IQ Achieve.xls" on the Web site www.tzkeith.com).

What does this value mean? If the two measures are perfectly related, that is, there is no difference at all in the z scores, our formula returns a value of zero. If, in contrast, the two scales are perfectly inverted so that every person who scored high on the first test scored low on the second test (and vice-versa), we obtain a value close to 4. Finally, if the two tests are unrelated, with scores on the Intelligence test providing no information whatsoever for scores on the Achievement test, then our formula produces a value around 2. Our scale ranges from 0 to 2 to 4. Although we are getting close, this is not a very logical scale, so let's make a few adjustments.

We can easily transform the scale into one that makes more sense:

$$r = 1 - \frac{1}{2} \left(\sum \frac{(z_x - z_y)^2}{N-1} \right)$$

We divide the previously obtained coefficient by 2 and subtract that value from 1. With these changes, our correlation coefficient is .778 ($r = .778$). Furthermore, our new correlation coefficient is much more logical. It ranges from 0, meaning that the two variables are unrelated, to 1, meaning that the two variables have the *exact same* z scores. In addition, the scale tells the direction of the relation. If it is positive, this means that high scores on one scale are paired with high scores on the other scale. If it is negative, high scores on one scale go with negative scores on another scale.

Another formula for r that also makes it obvious that we are comparing z scores is

$$r = \frac{\sum z_x z_y}{n-1}$$

We will not normally calculate r using either of these formulas (there is a formula that allows some computational short-cuts), but they make it obvious that we are looking for similarities and differences in the z scores of the two scales (for more detail, see Cohen et al., 2003, chap 2). Even better, we can calculate the correlation coefficient using a statistical program. Let's check the value above against the value calculated by SPSS. This value is also .778, as shown in Figure B.11.

The Pearson correlation coefficient can range from -1.0 , suggesting a perfect relation, but with high scores on one scale paired with low scores on the other scale, to $+1.0$, suggesting

Correlations

		IQ	ACHIEVE
IQ	Pearson Correlation	1	.778**
	Sig. (2-tailed)	.	.000
	N	30	30
ACHIEVE	Pearson Correlation	.778**	1
	Sig. (2-tailed)	.000	.
	N	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

Figure B11 Correlation (and its statistical significance) of Intelligence and Achievement scores.

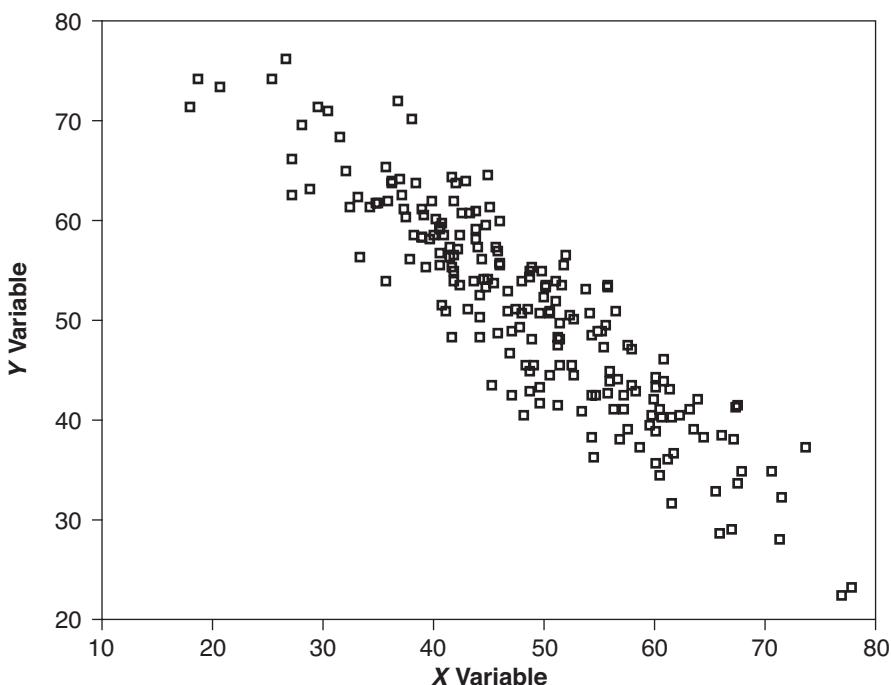


Figure B12 Scatterplot of a high negative correlation ($r = -.905$).

a perfect positive relation. A correlation of zero between the two scales would suggest no relation between the z scores; that is, the z scores between the two tests are unrelated. These relations are illustrated in Figures B.12 through B.14, which show scatterplots of large negative (B.12, $r = -.905$) near zero (B.13, $r = -.067$), and large positive (B.14, $r = .910$) correlations.

Statistical Significance of r

Correlation coefficients can be tested for statistical significance using the formula $t = r\sqrt{N-2} / \sqrt{1-r^2}$. You can then look up the value of t with $N-2$ df to determine how likely it is to get a value of this size given a “true” population value of zero. Using the current

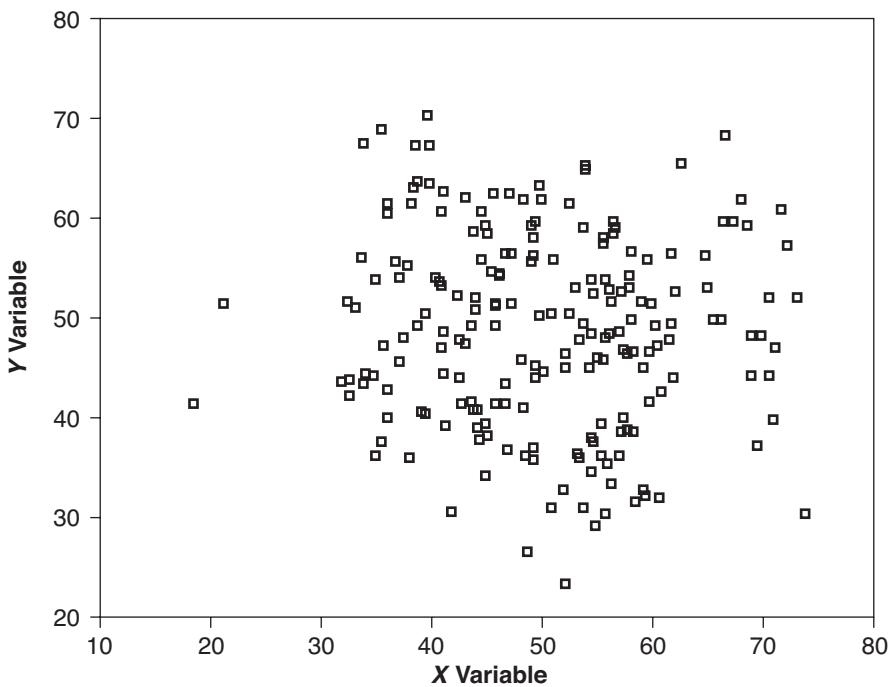


Figure B13 Scatterplot of a near zero correlation ($r = -.067$). The two scales are virtually unrelated.

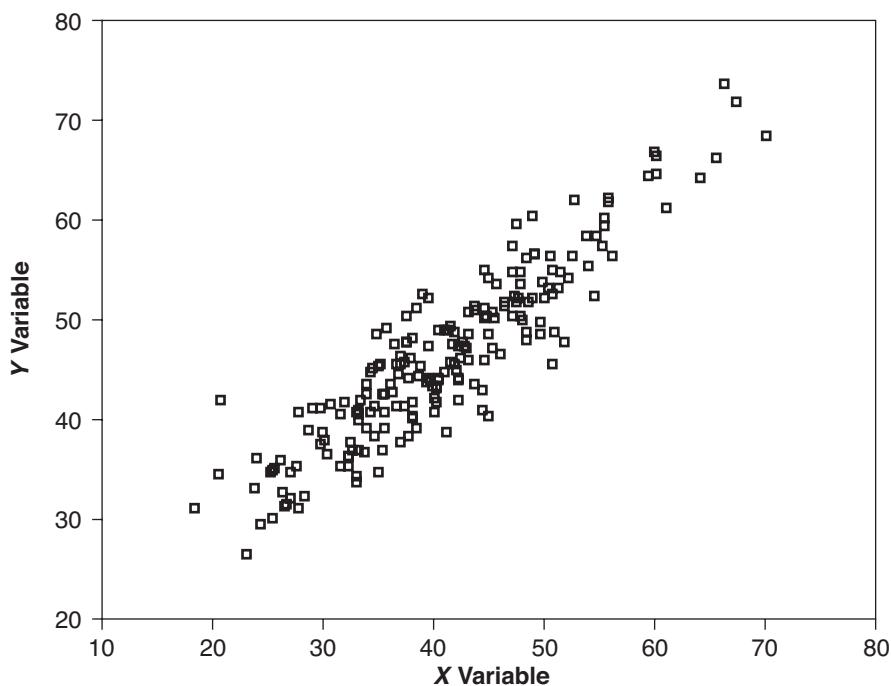


Figure B14 Scatterplot of a high positive correlation ($r = .910$).

example, the correlation between an Intelligence and an Achievement test, we obtain a t value of 6.56:

$$\begin{aligned} t &= \frac{\sqrt{N-2}}{\sqrt{1-r^2}} \\ &= \frac{.778\sqrt{30-2}}{\sqrt{1-.778^2}} \\ &= \frac{4.117}{.628} \\ &= 6.56 \end{aligned}$$

with 28 df . The correlation is indeed statistically significantly different from zero ($p < .001$).

In your reading, you will come across other varieties of correlation coefficients, such as Spearman's rho and the point–biserial correlation. Spearman's rho is appropriate when the variables are rankings; the point–biserial correlation is appropriate when one variable is continuous and the other dichotomous. How are they calculated? In fact, these two types of correlations (along with phi, a correlation with two dichotomous variables) are simply short-cuts for the correlation coefficient we derived above, Pearson's r . They are holdovers from the days when we calculated statistics by hand and, given the nature of the data (e.g., dichotomous), one could take a few short-cuts in the calculation. In this era of computers, you can calculate these three types of correlations just as easily using the standard r . Stated differently, these three types of correlations are really no different than Pearson r , but with different types of data. For more information, see Howell (2013).

T-TESTS

If you come to this book with a background in psychology or education, you may well have more experience with t tests and analyses of variance than with regression. It is likely, for example, that much of the research you read uses these methods. It is easy to think that these methods that are so appropriate for experimental research do something fundamentally different than does regression. This is not the case; the t test and ANOVA are simply subsets of multiple regression, illustrated early in this text. Here I will briefly review the use of these methods and illustrate them with computer output.

These statistical analyses (t tests and ANOVA) are especially useful in experimental research, in which participants are assigned to one group or another and given different experimental treatments. The primary difference between the t test and ANOVA is that a t test is appropriate when there is only one independent variable and there are only two groups, whereas ANOVA can be used with more than two groups and more than one IV.

As an example, suppose you were interested in the effects of cognitive behavior therapy (CBT) on the depressive symptoms of adolescent girls. Perhaps you set up an experiment in which each girl from a sample of 40 depressed girls is assigned, at random, to a CBT group or to a waiting list (members of which will receive treatment following the experiment if the treatment proves effective). The simulated data are included in the dataset “t test.sav” and “t test.xls” on the Web site. Some of the cases are shown in Table B.1. The first column shows the group (1 = experimental, or CBT, and 0 = control, or wait list), and the second column shows the girls' scores on a measure of depressive symptoms following treatment (a high score represents more depressive symptomology and thus is bad).

Table B.1 Portion of the Data from the *t* Test Example

<i>Group</i>	<i>Depress</i>
0	66
0	63
0	44
0	56
0	62
0	65
0	35
0	62
0	76
.	.
.	.
.	.
1	60
1	56
1	59
1	47
1	47
1	49
1	45
1	63
1	57
1	30

Group Statistics

	GROUP Treatment group	N	Mean	Std. Deviation	Std. Error Mean
DEPRESS symptoms	0 Control	20	59.65	9.599	2.146
	1 CBT, Experimental	20	49.25	10.915	2.441

Independent Samples Test

	t-test for Equality of Means						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
DEPRESS symptoms	3.200	38	.003	10.40	3.250	3.820	16.980

Figure B15 *t*-test results for the simulated CBT therapy experiment.

Figure B.15 shows a portion of the results of a *t* test conducted on these data using SPSS. After treatment, the average score for the control group was 59.65 on the measure of depressive symptoms versus an average score of 49.25 for the experimental group. Given that a high score on the measure represents greater depression, the experimental group indeed showed less depression than the control group. Is this difference between groups statistically significant? The second table in the figure shows the results of the *t* test. The *t* associated

with the difference between groups was 3.20. With 38 ($N-2$) degrees of freedom, the probability of obtaining a t this large by chance is .003, or 3 out of 1000. Using common rules of thumb for statistical significance, the difference between groups was indeed statistically significant. Because girls were assigned at random to the two groups, we have effectively ruled out other plausible explanations for the difference (e.g., that girls who received treatment were less depressed to start with) and can conclude that the CBT treatment was probably indeed effective. (In this book we will come to call such ruling out of alternative explanations by a different name: ensuring that there are no common causes of the presumed cause and the presumed effect.)

The general formula for a t test is $t = (M_e - M_c)/SE_{e-c}$ or the difference in means between the experimental and control group divided by the standard error of that difference. It really doesn't matter which group is subtracted from which, because you are primarily interested in the absolute value of t . The df are $N-2$.

Although this process of comparing means seems very different from the process of correlating two variables, they are, in fact, the same process. As I will show in the main text, you will get the same essential results if you correlate the two variables (group and depressive symptoms scores) that we got with the t test.

Effect Sizes

It is interesting to know that the two groups are statistically significantly different, but is the difference large, small, or somewhere in between? This is, in my opinion, one advantage of regression approaches: with multiple correlation and standardized regression coefficients, we automatically get an index of the magnitude of the effect. There are a number of measures of effect size that are common for two-group experimental research; the most common is likely d . The formula for d is $d = (M_e - M_c)/SD$ or the difference between the two groups divided by the overall standard deviation (think of d as somewhat like a z score). For the present example, d is .910 [$d = (49.25 - 59.65)/11.431$]; ignore the sign). According to common rules of thumb, d 's above .80 are considered large (small = .20, medium = .50, large = .80; Cohen, 1988), although it is possible and desirable to have different rules of thumb for specific areas of research. According to these generic rules of thumb, CBT therapy in our simulated data had a large effect on depressive symptoms.

ANOVA

Analysis of variance is appropriate when there are more than two groups in an experiment or when there is more than one independent variable. It can also be used to analyze data from experiments with one IV and only two groups and will give the same results as the t test.

Consistency with the t Test

Figure B.16 shows the results of an ANOVA for the therapy–depression example above. The lower table shows that for the ANOVA, like the t test, the difference between the two groups on the posttest was statistically significant. The F statistic was 10.239 with 1 and 38 degrees of freedom; such an F is unlikely to occur if there are no real differences between groups ($p = .003$). The F is equal to t^2 and shows the same level of statistical significance as does the results of the t test.

The general formula for F is $F = \frac{V_{\text{between group}}}{V_{\text{within group}}}$, the variation between groups divided by the average variation within groups. You can actually calculate the variance of the group means

Between-Subjects Factors

		Value Label	N
GROUP Treatment group	0	Control	20
	1	CBT, Experimental	20

Descriptive Statistics

Dependent Variable: DEPRESS Depressive symptoms

GROUP Treatment group	Mean	Std. Deviation	N
0 Control	59.65	9.599	20
1 CBT, Experimental	49.25	10.915	20
Total	54.45	11.431	40

Tests of Between-Subjects Effects

Dependent Variable: DEPRESS Depressive symptoms

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
GROUP	1081.600	1	1081.600	10.239	.003	.212
Error	4014.300	38	105.639			
Corrected Total	5095.900	39				

Figure B16 ANOVA results for the simulated CBT therapy experiment. The results are the same as for the t -test, although $F = t^2$.

(times the n in each group) to get the $V_{between}$ and take a weighted average of the variances of the groups to obtain the V_{within} . F statistics require two values for degrees of freedom, generally corresponding to the treatment and error (within group). The total df for the ANOVA is equal to $N-1$. The df for the treatment is the number of groups minus 1, and the df for the error term is equal to the $df_{total} - df_{group}$.

It may seem that we are doing something different with a t test, which compares group means, versus ANOVA, which analyzes variances. But the general formula for F above shows that the variance in the numerator of the equation is the variance of *group means*. Yes, the processes are essentially the same. I will demonstrate in the text that ANOVA can be accomplished through multiple regression. As you read the text itself, you should note the general similarity of the formula for F in ANOVA and that for F for regression. Both divide the variance explained by the independent variable by the variance left unexplained.

Effect Sizes, η^2 and f^2

A number of measures of effect size are available for ANOVA. Shown in Figure B.16 is eta-squared ($\eta^2 = .212$). η^2 is a great measure of effect size for our purposes, because as we will see, it is equal to R^2 from a regression solution to the same problem. Common rules of thumb for η^2 are small = .01, medium = .10, and large = .25. Another common measure of effect size, Cohen's f (or f^2) may be calculated from η^2 using the formula $f^2 = \frac{\eta^2}{(1-\eta^2)}$. As noted in Chapter 4, a common rule of thumb for f^2 is that .02 represents a small effect, .15 a medium effect, and .35 a large effect (Cohen et al., 2003, p. 95).

Between-Subjects Factors

	Value Label	N
GROUP CBT .00	Control	40
vs Control 1.00	CBT	40
SEX Girls vs .00	Girls	40
Boys 1.00	Boys	40

Descriptive Statistics

Dependent Variable: DEPRESS Depressive Symptoms

GROUP CBT vs Control	SEX Girls vs Boys	Mean	Std. Deviation	N
.00 Control	.00 Girls	60.500	11.4455	20
	1.00 Boys	54.650	13.0557	20
	Total	57.575	12.4754	40
1.00 CBT	.00 Girls	48.850	8.2798	20
	1.00 Boys	45.800	9.7257	20
	Total	47.325	9.0480	40
Total	.00 Girls	54.675	11.4900	40
	1.00 Boys	50.225	12.2149	40
	Total	52.450	11.9936	80

Tests of Between-Subjects Effects

Dependent Variable: DEPRESS Depressive Symptoms

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^a
GROUP	2101.250	1	2101.250	18.091	.000	.192	18.091	.987
SEX	396.050	1	396.050	3.410	.069	.043	3.410	.446
GROUP * SEX	39.200	1	39.200	.337	.563	.004	.337	.088
Error	8827.300	76	116.149					
Corrected Total	11363.800	79						

a. Computed using alpha = .05

Figure B17 Results of a factorial ANOVA to compare the effects of CBT versus no therapy on the depressive symptoms for boys and girls.

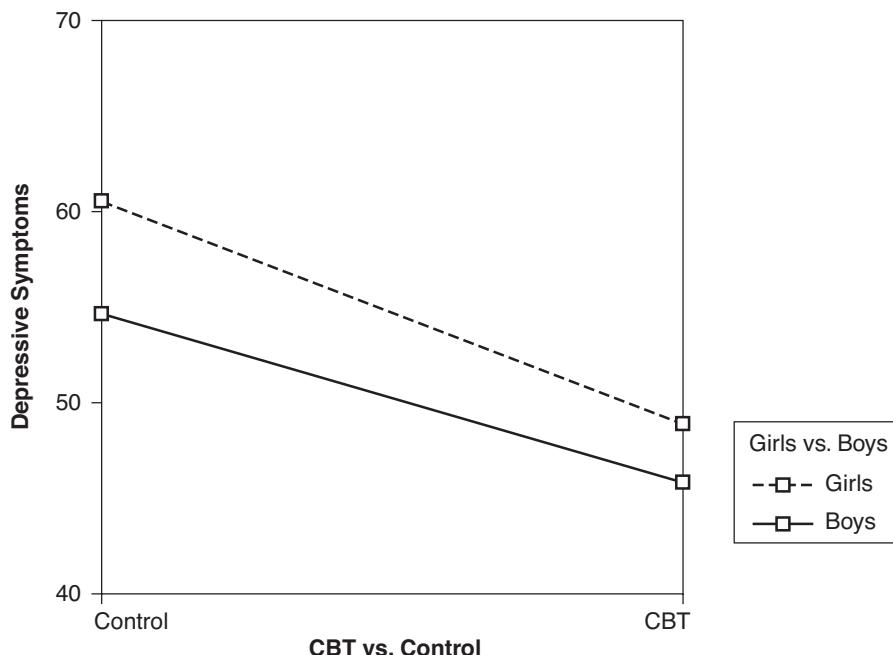


Figure B18 Graph of means for the factorial ANOVA example.

Factorial ANOVA

To take our therapy–depression example a little further, suppose you were interested in whether CBT had positive effects for depressed adolescent boys as well as girls. You could conduct a new experiment using both boys and girls. You are not sure, however, whether CBT will have the same effect for both sexes, so you add Sex as a second independent variable. In the parlance of ANOVA, you now have a design appropriate for analysis via a 2×2 factorial ANOVA. The analysis will determine the effect of CBT, the effect of Sex, and whether there are differential effects of CBT for boys and girls (the interaction between Group and Sex in their effects on Depressive Symptoms).

Figure B.17 shows some of the output from the 2×2 ANOVA. The data are in the files “cbt 2way.sav” and “cbt 2way.xls” on the Web site. As shown in the bottom table of the figure, Group (CBT versus Control) had a medium to large effect on Depressive Symptoms in these simulated data and this effect was statistically significant ($\eta^2 = .192$, $F = 18.091$ [1, 76], $p < .001$). An examination of the means shows that adolescents in the experimental (CBT) group had fewer depressive symptoms at posttest than did the control group. Sex had a small effect that was not statistically significant ($p = .069$); nor was the interaction statistically significant ($p = .563$).

The data are graphed in Figure B.18, an excellent way to summarize data from this type of experiment. It is clear that both boys and girls in the experimental group benefited from the CBT therapy. It appears that boys in both groups show somewhat fewer symptoms than do girls, but the ANOVA tells us that this difference is not statistically significant. The fact that the two lines are basically parallel reaffirms the nonsignificant interaction term from the ANOVA and shows that CBT had similar effects for both boys and girls.

I hope this quick review has gotten your mind back into statistics so that you are ready to begin exploring multiple regression and structural equation modeling. If you need additional review, Howell (2013) and Pituch, Whitaker, and Stevens (in press) are excellent. For an even more gentle review, Kranzler (2018) is an excellent resource.

Appendix C

Partial and Semipartial Correlation

In earlier chapters we touched on the topic of semipartial correlations and noted how they are related to ΔR^2 and to t . I also mentioned partial and semipartial correlation when we first raised the issue of the meaning of “controlling for” in Chapter 2. In this appendix, we will focus in more detail on the topics of partial and semipartial correlation. I have placed this topic in an appendix for several reasons. It does not really fit in with the flow of the other chapters, and it is a topic that will not be of interest to all readers of the text. In addition, although the topic fits better in Part 1 of the text, it will be more understandable following an introduction to path and SEM models.

PARTIAL CORRELATIONS

Partial correlations are correlations between two variables, with other variables taken into account. You may also hear partial correlations described as the correlation between two variables with the effects of other variables removed or other variables controlled. Let’s use an example to illustrate.

Example: Optimism and Locus of Control

Figure C.1 shows the correlations among several variables from the NELS data. Optimism is a composite I created from a series of 11 questions about students’ outlook toward the future (F1S64A through F1S64K; note that F1S64L was not used): “Think about how you see the future. What are the chances that:

- You will graduate from high school?
- You will go to college?
- You will have a job that pays well?
- You will be able to own your own home?
- You will have a job that you enjoy doing?
- You will have a happy family life?
- You will stay in good health most of the time?
- You will be able to live wherever you want in the country?
- You will be respected in your community?
- You will have good friends you can count on?
- Life will turn out better for you than it has for your parents?”

		Correlations				
Control Variables		optimism Level of optimism, 10th grade	f1locus2 LOCUS OF CONTROL 2	par_inv Percent involvement	byses SOCIO- ECONOMIC STATUS COMPOSITE	bygrads GRADES COMPOSITE
-none-a	optimism Level of optimism, 10th grade	Correlation Significance (2-tailed) df	.1000 .0 799	.364 .000 799	.315 .000 799	.204 .000 799
	f1locus2 LOCUS OF CONTROL 2	Correlation Significance (2-tailed) df	.364 .000 799	1.000 .0 0	.243 .000 799	.203 .000 799
	par_inv Percent involvement	Correlation Significance (2-tailed) df	.315 .000 799	.243 .000 799	1.000 .0 0	.419 .000 799
	byses SOCIO-ECONOMIC STATUS COMPOSITE	Correlation Significance (2-tailed) df	.204 .000 799	.203 .000 799	.419 .000 799	1.000 .0 0
	bygrads GRADES COMPOSITE	Correlation Significance (2-tailed) df	.299 .000 799	.253 .000 799	.391 .000 799	.342 .000 0
par_inv Parent Involvement & byses SOCIO-ECONOMIC STATUS COMPOSITE & bygrads GRADES COMPOSITE	optimism Level of optimism, 10th grade	Correlation Significance (2-tailed) df	1.000 .0 796	.284 .000 796		
	f1locus2 LOCUS OF CONTROL 2	Correlation Significance (2-tailed) df	.284 .000 796	1.000 .0 0		

a. Cells contain zero-order (Pearson) correlations.

Figure C.1 Zero-order correlations among Optimism, Locus of Control, Parent Involvement, SES, and Grades. The lower half of the table shows the partial correlation of Optimism and Locus of Control with the other variables controlled, or “partialed out.”

Students with high scores on the composite had a fairly optimistic view of the future, whereas those with low scores were more pessimistic about the future. A dichotomized version of this variable was used in Chapter 11 to illustrate logistic regression. F1Locus2 is a locus of control scale; students with an internal locus of control had high scores, whereas those with an external locus had low scores. Par_Inv is a measure of parent involvement in education, defined as the educational aspirations parents have for their children along with the extent that they communicate with their children about school and education. BySES and ByGrads are the SES (Family Background) and GPA composites we have used previously.

You could easily create this composite as the mean of the 11 items, but here I have used a subset of the NELS data that just includes the relevant variables. The reason for doing so was to create a dataset with no missing data so that all of the different methods below would treat the missing data in the same way (listwise deletion of missing data). With different methods used some of the estimates of partial and semipartial correlations would differ across methods. The data are in the file “nels optimism partial 11 item.sav” on the web site (www.tzkeith.com).

The output shown in Figure C.1 is from the SPSS Partial Correlation procedure. The top half of the table shows the correlations among the variables used in the analysis, without controlling for any other variables. Thus the first column, under “Control Variables” says “none” for this portion of the Table. Previously in this text, I noted that simple Pearson correlation coefficients are sometimes called zero-order correlations. Thus, the note at the bottom of this table labels these correlations as zero-order (Pearson) correlations. This simply means that these are correlations with no other variables controlled. The primary correlation of interest is between the variables Optimism and Locus of Control: .36. Adolescents who have a more internal locus of control are also more optimistic. Note also, however, that these primary variables of interest also show small to moderate correlations with the other variables, most in the .2 to .3 range. Thus, it is likely that once we control for, or remove the effects of, these variables the correlation between Optimism and Locus will decrease.

The lower portion of Figure C.1 shows the partial correlation between Optimism and Locus, controlling for Parent Involvement, SES, and base year GPA. As expected, once these background variables are controlled, the partial correlation is lower than the zero-order correlation (.284, or .28 rounded). I'll symbolize this partial correlation as $pr_{Optimism-Locus-Parent,SES,Grades} = .28$ with the pr symbolizing partial correlation and the dot symbolizing "controlling for ..." If this sounds a lot like regression coefficients, it should. We spoke of the coefficients in MR as representing the effect of one variable on another, controlling for one or more background variables. These regression coefficients from multiple regression are sometimes also referred to as *partial* regression coefficients. The difference is that partial correlations are correlations; that is, they have no directional quality, no implication of cause and effect or the prediction of one variable from another. In the introduction to path analysis, I referred to an agnostic model; partial correlations are like agnostic regression coefficients. Again, this partial correlation may be considered as the correlation between Optimism and Locus of Control, with the effects of Parent Involvement, SES, and Grades controlled or removed.

Understanding Partial Correlations

If you recall our initial discussions of multiple regression, the phrase "with the effects of ... removed" should also sound familiar. Recall that in Chapter 3 we used this phrase to describe the residuals. There we described the residual from the regression of Grades on Homework and Parent Education as representing Grades with the effects of Homework and Parent Education removed. Are partial correlations, then, related to the residuals in some way? Yes. One way of calculating partial correlations is to regress each variable of interest (Optimism and Locus) on the control variables (Parent Involvement, SES, and Grades) and to save the residuals. These residuals then represent Optimism and Locus of Control with the effects of the control variables removed. The correlation between these two residuals is then equivalent to the partial correlation of Optimism with Locus, with the effects of Parent Involvement, SES, and Grades removed. The correlation between the Optimism and Locus residuals (with the effects of SES, Parent Involvement, and Grades removed from each) is shown in Figure C.2; the value (.28) is the same as the partial correlations in Figure C.1.

Correlations

		Opt_Res Optimism Unstandardized Residual	Loc_Res Locus Unstandardized Residual
Opt_Res Optimism Unstandardized Residual	Pearson Correlation	1	.284**
	Sig. (2-tailed)		.000
	N		801
Loc_Res Locus Unstandardized Residual	Pearson Correlation	.284**	1
	Sig. (2-tailed)	.000	.
	N	801	801

**. Correlation is significant at the 0.01 level (2-tailed).

Figure C.2 The correlation of Optimism and Locus residuals (controlling for SES, Grades, and Parent Involvement) is equal to the partial correlation between Optimism and Locus, controlling for these background variables.

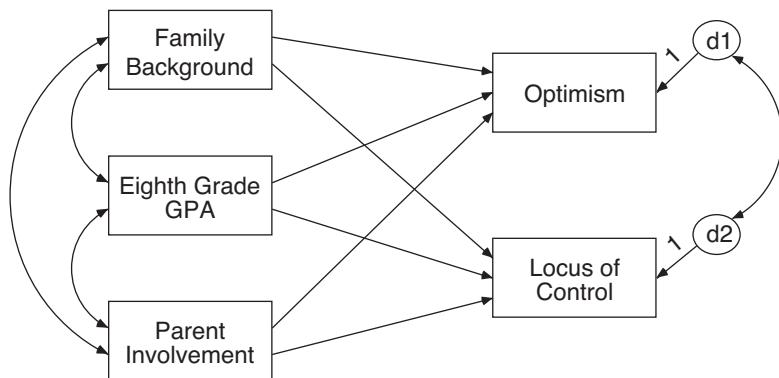


Figure C.3 Partial correlation in path analytic form. The correlation between the disturbances (residuals) is a partial correlation.

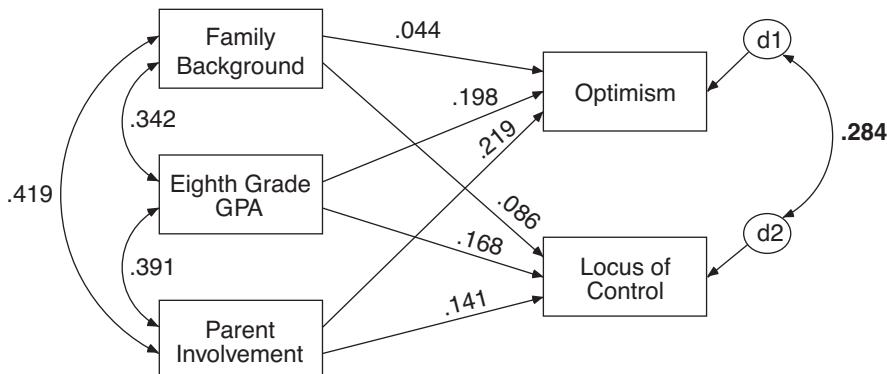


Figure C.4 Solved path model shows the equivalence of the partial correlation with the estimates in the previous figures.

Figure C.3 demonstrates the relation between residuals and partial correlations using path analysis. Recall that the disturbances in path analysis are the same as the residuals in MR. Thus the disturbance d_1 represents Optimism with SES, Parent Involvement, and GPA controlled; d_2 represents Locus with these three background variables controlled. The correlation between these disturbances, then, is the partial correlation of Optimism with Locus of Control, with SES, Grades, and Parent Involvement taken into account, or controlled. Figure C.4, the solved path model, shows that the value .28 is again the same as the partial correlations from the partial correlation procedure and from the correlation between residuals.

Uses of Partial Correlations

Why would you use partial correlations? One potential reason is when you want to take obvious control variables into account without making causal statements about the two variables of interest. Alternatively, you may be interested in whether the correlation between two variables is spurious, the product of each being affected by one or more common causes. To use the present example, you may be interested in whether the correlation between

students' levels of optimism and their locus of control is nonspurious or the extent to which the correlation remains after taking into account the background variables (potential common causes) of SES, Parent Involvement, and Grades. From an explanatory standpoint, you might be interested in the effects of all these variables on each other, but be unable to decide whether optimism affects locus of control or the reverse. Although the model shown in Figures C.3 and C.4 will not help you decide which variable was the cause and which the effect, it will allow you to determine that the variables are still related in some way, after controlling for other relevant variables (assuming these are *the* relevant variables). Another possible meaning of partial correlations in the path models is that we recognize that there may also be other common causes of these two variables not taken into account in the model.

Partial correlations are also sometimes used in research on mediation (cf., Baron & Kenny, 1986). As discussed in both Parts 1 and 2 of this text, I think most questions of mediation are more easily tested via multiple regression or the indirect effects in path analysis and structural equation modeling, however.

Semipartial Correlations

With semipartial correlations (also known as *part* correlations), the effects of the background or control variables have been removed from only *one* variable of interest. An example is shown in Figure C.5, which illustrates the semipartial correlation of Locus with Optimism, with the effects of SES, Parent Involvement, and GPA removed from Optimism (but not removed from Locus of Control). The correlation between the disturbance/residual of Optimism and the variable Locus of Control is equivalent to this semipartial correlation: $sr_{Locus-(Optimism \cdot Parent, SES, Grades)} = .271$. (In this method of representation, the parentheses around both Optimism and the control variables illustrates that the control variables are partialed from Optimism, but not Locus of Control; the *sr* stands for semipartial correlation.)

Given this description of semipartial correlations, it should also be possible to compute them using the residuals from multiple regression. It is. Again, our interest is in the correlation between Locus of Control and Optimism, with the background variable effects removed from Optimism. This means that we should correlate the Optimism residuals (SES, Grades, and Parent Involvement controlled) with the original Locus of Control variable. The value .271 is shown in Figure C.6. This value is the same as that shown in the path model.

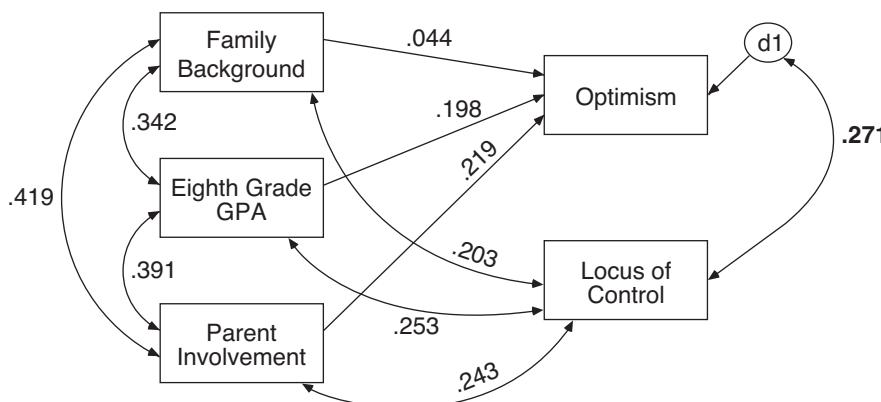


Figure C.5 Semipartial, or part, correlations in path analytic form. The effects of the control variables have been removed only from Optimism, not Locus of Control. The semipartial correlation is equal to the correlation between the Optimism disturbance and the Locus variable.

Correlations

		Opt_Res Optimism Unstandardized Residual	f1Locus2 LOCUS OF CONTROL 2
Opt_Res Optimism Unstandardized Residual	Pearson Correlation	1	.271**
	Sig. (2-tailed)		.000
	N	801	801
f1Locus2 LOCUS OF CONTROL 2	Pearson Correlation	.271**	1
	Sig. (2-tailed)	.000	
	N	801	801

**. Correlation is significant at the 0.01 level (2-tailed).

Figure C.6 Calculating semipartial correlations via residuals. The Locus of Control variable is correlated with the Optimism residual.

Many statistics programs do not compute semipartial correlations directly; there is, for example, no semipartial correlation procedure in SPSS. In Chapter 5, however, you saw several methods of getting semipartial correlations as a result of MR output. It is possible in most programs, for example, to request semipartial correlations as a part of the MR output. Recall, also, that *squared* semipartial correlations are equivalent to the unique variance of a variable entered last in a regression equation, that is, ΔR^2 .

The only tricky part about using MR to calculate semipartial correlations is understanding which of the two variables being correlated has background variables controlled and which does not. Unlike partial correlations, semipartial correlations are not symmetric. That is, $sr_{\text{Locus}-(\text{Optimism}\cdot\text{Parent,SES,Grades})} \neq sr_{\text{Optimism}-(\text{Locus}\cdot\text{Parent,SES,Grades})}$. When using MR to calculate semipartial correlations, the outcome or dependent variable is uncontrolled (or outside the parentheses), whereas the variable controlled is considered one of the predictor variables. From Chapter 5: “Conceptually, a semipartial correlation is the correlation of Y with X_1 , with the effects of X_2, X_3 , and so on, removed from X_1 . It may be symbolized as $sr_{y-(1\cdot 23)}$, with the parentheses showing that the effects of X_2 and X_3 are removed from X_1 , but not from Y ” (p. 88). Thus, to calculate $sr_{\text{Locus}-(\text{Optimism} \cdot \text{Parent, SES, Grades})}$, we would need to regress Locus of Control on SES, Parent Involvement, Grades, and Optimism.

I regressed Locus of Control on SES, Parent Involvement, and Grades in a simultaneous regression and then sequentially added Optimism to the regression. Figure C.7 shows that the change in R^2 for the addition of Optimism was .073. Recall that the semipartial correlation is equivalent to $\sqrt{\Delta R^2}$, or $\sqrt{.073} = .270$, again consistent with other estimates within errors of rounding. Think about what this means: The semipartial correlation squared is equal to the unique variance that Optimism explains in Locus of Control, after the other variables have been taken into account. This should make sense when you focus on Figure C.5 as well. We’ve already removed any effects that SES, Grades, and Parent Involvement have on Optimism; what then is the unique aspect that Optimism can explain in Locus of Control?

Figure C.8 shows the table of coefficients from the second part of this same regression. The final three columns of the table list the original correlation between each of the four variables (SES, Grades, Parent Involvement, and Optimism) and Locus; the partial correlation between each variable with Locus, with the other three variables partialled out of both the dependent and respective independent variables, and the semipartial (part) correlation of each variable with Locus, with the other three variables removed only from the independent variable side of the equation. Thus, the first part correlation (Optimism, .271, bolded),

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.307 ^a	.094	.091	.59007	.094	27.664	3	797	.000
2	.409 ^b	.167	.163	.56608	.073	69.982	1	796	.000

- a. Predictors: (Constant), par_inv Parent Involvement, bygrads GRADES COMPOSITE, byses SOCIO-ECONOMIC STATUS COMPOSITE
- b. Predictors: (Constant), par_inv Parent Involvement, bygrads GRADES COMPOSITE, byses SOCIO-ECONOMIC STATUS COMPOSITE, optimism Level of optimism, 10th grade

Figure C.7 Calculating semipartial correlations using multiple regression. The square root of the change in R^2 when Optimism is entered last in a regression is equal to its semipartial correlation with Locus (the outcome), with the effects of the background variables removed from Optimism.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	Correlations		
	B	Std. Error				Zero-order	Partial	Part
1	(Constant)	-1.530	.167	-9.186	.000			
	optimism Level of optimism, 10th grade	.310	.037	.291	8.366	.000	.364	.284
	byses SOCIO-ECONOMIC STATUS COMPOSITE	.060	.030	.073	2.007	.045	.203	.071
	bygrads GRADES COMPOSITE	.096	.032	.111	3.018	.003	.253	.106
	par_inv Parent Involvement	.061	.030	.078	2.042	.041	.243	.072
								.066

a. Dependent Variable: f1locus2 LOCUS OF CONTROL 2

Figure C.8 Additional output from the multiple regression. Some programs (e.g., SPSS) will produce semipartial correlations on request. It is also possible to calculate semipartial correlations from the t values.

shows the semipartial correlation of Optimism with Locus of Control, with the effects of SES, Grades, and Parent Involvement removed from Optimism. The coefficient for Parent Involvement, in turn, shows the semipartial correlation of Parent Involvement with Locus of Control, with SES, Grades, and Optimism removed from Parent Involvement, and so on.

As noted in Chapter 5 (p. 107, note 1), it is possible to calculate the semipartial correlations from the values of t given in the output for each coefficient:

$$sr_{y(1-234)} = t \sqrt{\frac{1-R^2}{N-k-1}}$$

For the Locus of Control–Optimism semipartial correlation, the value of t (from Figure C.8) is 8.366 (also bolded), and the equation is

$$\begin{aligned} sr_{\text{Locus-(Optimism-SES,Grades,Parent)}} &= t \sqrt{\frac{1-R^2}{N-k-1}} \\ &= 8.366 \sqrt{\frac{1-.167}{801-4-1}} \\ &= .271 \end{aligned}$$

(with R^2 and df from Figure C.7).

Finally, as pointed out to me by my colleague Matt Reynolds, it is possible to calculate the semipartial correlation from the b . Recall the formula for calculating β from b : $\beta = b \frac{SD_x}{SD_y}$.

If you substitute the value for the SD for the Optimism *residual* in the formula for the SD of Optimism (that is, in place of SD_x), you will get a value of .270. This is equal, within errors of rounding, to the semipartial correlation shown in the various figures. This method is probably not one you will use for calculation, but it helps us understand how semipartial correlations are related to b and to β .

Uses of Semipartial Correlations

In my experience, the most common use of semipartial correlations is in attempts to describe the unique variance of a predictor in accounting for some outcome. Given the adequacy of the variables in the model, the squared semipartial correlations provide estimates of the unique variance of each independent variable in explaining the outcome. The s^2 values are equal to the ΔR^2 values obtained when each variable is added last in the regression equation.

Semipartial correlations (not squared) can also be used to describe the relative importance of the variables in a regression. In such usage (and, again, given the adequacy of the regression model), they are interpreted in much the same way as β 's, as representing the *relative direct effects* of each variable on the outcome. Indeed, some authors recommend the semipartial correlations over regression coefficients for this purpose (e.g., Darlington & Hayes, 2017).

Conclusion

Partial and semipartial correlations are useful adjuncts to multiple regression analysis and can be useful procedures by themselves. Although the primary focus of this book has been on using multiple regression and related methods in an explanatory fashion, research questions do not always fit this mold. We sometimes are interested in the extent to which a set of background variables explains the existing correlation between two variables, that is, the extent to which the relation may be spurious. Alternatively, we may be interested in demonstrating that a correlation still exists after controlling for such background variables or that a key variable predicts an outcome after controlling for background effects. Partial and semipartial correlations are useful in these cases. In this short appendix we have approached these concepts from several different orientations; it is not necessary that you understand all these different methods of obtaining and explaining partial and semipartial correlations. One or two of them should resonate so that you feel comfortable with and understand these concepts.

Appendix D

Symbols Used in This Book

Symbol Definition

<i>a</i>	Intercept in a regression equation
aBIC	Sample-size adjusted BIC, a measure of fit in SEM
AIC	Akaike information criterion, a measure of fit in SEM
<i>b</i>	Unstandardized regression coefficient
BIC	Bayes information criterion, a measure of fit in SEM
CFI	Comparative fit index, a measure of fit in SEM
<i>CI</i>	Confidence interval
<i>CoV_{xy}</i>	Covariance of X and Y
<i>d</i>	Disturbance in SEM, same as the residual in MR; <i>d</i> is also used to symbolize a measure of effect size when comparing two groups in experimental research
<i>df</i>	Degrees of freedom
<i>e</i>	Error
<i>f</i> ²	A common measure of effect size, used in both ANOVA and MR. Calculable from R^2
<i>F</i>	The product of ANOVA, used to test the statistical significance of R^2 , or the difference between groups in an experiment
<i>g</i>	Number of groups in a categorical variable. <i>g</i> is also commonly used to represent a general intelligence factor
<i>k</i>	Number of independent variables in a regression
<i>M</i>	Mean
<i>MI</i>	Modification index in SEM
<i>N</i>	Number of participants, sample size
<i>p</i>	Probability
<i>r</i>	Correlation coefficient; Pearson product moment correlation; zero-order correlation. <i>r</i> is also sometimes used to represent unique and error variances (or residuals) in SEM and CFA
<i>r_{tt}</i>	Reliability coefficient
<i>R</i>	Multiple correlation coefficient

R^2	Squared multiple correlation. The variance explained in a dependent variable by a set of independent variables.
RMSEA	Root mean square error of approximation, a measure of fit in SEM
SD	Standard deviation
SE	Standard error, as in SE of a regression coefficients (SE_b)
sr	Semipartial correlation, equal to $\sqrt{\Delta R^2}$ when a variable is added last to a regression equation
SRMR	Standardized root mean square residual, a measure of fit in SEM
ss	Sums of squares, a measure of variation, used to calculate R and determine the statistical significance of a regression equation
t	As in t test. t tests are used to test the statistical significance of regression coefficients, means, and many other parameters.
T scores	Standardized scores with a $M = 50$ and $SD = 10$
TLI	Tucker-Lewis index, also known as the NNFI, the non-normed fit index, a measure of fit in SEM
u	Unique and error variance in SEM and CFA
V	Variance
X	An independent variable
Y	A dependent variable
Y'	The predicted Y
z	As in z scores. Standardized scores with $M = 0$, $SD = 1$. The basis for all other types of standard scores.
α	Alpha, the probability level
β	Beta, the standardized regression coefficient
Δ	Delta, used to symbolize change, as in ΔR^2
η^2	Eta-squared, a measure of effect size in ANOVA that is equivalent to R^2
χ^2	Chi-square, a common measure of fit in SEM models

Appendix E

Useful Formulae

Formula	Purpose
$F = \frac{ss_{\text{regression}} / df_{\text{regression}}}{ss_{\text{residual}} / df_{\text{residual}}}$	Test the statistical significance of a regression
$F = \frac{R^2 / k}{(1 - R^2) / (N - k - 1)}$	Test the statistical significance of a regression
$R^2 = \frac{ss_{\text{regression}}}{ss_{\text{total}}}$	Calculate R^2
$b = \frac{\text{rise}}{\text{run}} = \frac{M_y - a}{M_x}$	The unstandardized regression coefficient, or the slope of the regression line
$t = \frac{b}{SE_b}$	Test the statistical significance of a regression coefficient
$Y = a + b_1 X_1 + b_2 X_2 \dots + e$	General form of a regression equation
$\beta = b \frac{SD_x}{SD_y}, b = \beta \frac{SD_y}{SD_x}$	Converting from standardized to unstandardized regression coefficients, and vice-versa
$SD = \sqrt{V}, V = SD^2$	Converting from standard deviation to variance
$r_{xy} = \frac{Cov_{xy}}{SD_x SD_y}$	Calculate a correlation from a covariance
$\beta_1 = \frac{r_{y1} - r_{y2} r_{12}}{1 - r_{12}^2}$	Formula for calculating β in a regression with two independent variables
$R^2_{y12} = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}$	Formula for calculating R^2 in a regression with two independent variables

$$F = \frac{R_{12}^2 - R_1^2 / k_{12} - k_1}{1 - R_{12}^2 / (N - k_{12} - 1)}$$

Statistical significance for change in R^2 (ΔR^2), used to test the statistical significance of variables added sequentially to the regression equation

$$f^2 = \frac{R^2}{1 - R^2}$$

Cohen's f^2 , a common measure of effect size, calculated from R^2

$$f^2 = \frac{R_{y,12}^2 - R_{y,1}^2}{1 - R_{y,12}^2}$$

Cohen's f^2 , a common measure of effect size, calculated from change in R^2

$$V_e = (1 - r_{tt})V$$

Formula for calculating the error variance from the reliability of a measure and its total variance. Used for single-indicator latent variables in SEM.

$$r = \frac{\sum z_x z_y}{n - 1}$$

Formula for the Pearson correlation coefficient

References

- Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York, NY: Routledge.
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Alexander, K. W., Quas, J. A., Goodman, G. S., Ghetti, S., Edelstein, R. S., Redlich, A. D., . . . & Jones, D. P. H. (2005). Traumatic impact predicts long-term memory for documented child sexual abuse. *Psychological Science*, 16, 33–40. doi:10.1111/j.0956-7976.2005.00777.x
- Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, 115, 308–314.
- Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge.
- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3–25. doi:10.1037/amp0000191
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–278). Mahwah, NJ: Erlbaum.
- Arbuckle, J. L. (2017). *IBM SPSS Amos 25 user's guide*. Crawfordville, FL: Amos Development Corporation.
- Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4), 495–508. doi:10.1080/10705511.2014.919210
- 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.
- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York, NY: Routledge.
- Belli, R. F. (2016). Toward reconciliation of the true and false recovered memory debate. In R. Burnett (Ed.), *Wrongful allegations of sexual and child abuse* (pp. 255–270). New York, NY: Oxford University Press.
- Bennett, W. J. (1987). *James Madison high school: A curriculum for American students*. Washington, DC: U.S. Department of Education.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Woodward, J. A. (1978). A Head Start reevaluation: Positive effects are not yet demonstrable. *Evaluation Quarterly*, 2, 493–510.
- Berry, W. D. (1993). *Understanding regression assumptions*. Thousand Oaks, CA: Sage.
- Birnbaum, M. H. (1979). Procedures for the detection and correction of salary inequities. In T. H. Pezzullo & B. E. Brittingham (Eds.), *Salary equity* (pp. 121–144). Lexington, MA: Lexington Books.
- Blalock, H. M. (1972). *Social statistics* (2nd ed.). New York, NY: McGraw-Hill.

- Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., . . . Brandmaier, A. (2012). *OpenMx user's guide*. Charlottesville, VA: University of Virginia.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.
- Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NY: Wiley.
- Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, 52(4), 465–484. doi:10.1080/00273171.2017.1309262
- Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, 50, 229–242.
- Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, 7, 461–483.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71, 425–440.
- Brady, H. V., & Richman, L. C. (1994). Visual versus verbal mnemonic training effects on memory-deficient and language-deficient subgroups of children with reading disability. *Developmental Neuropsychology*, 10, 335–347.
- Bremner, J. D., Shobe, K. K., & Kihlstrom, J. F. (2000). False memories in women with self-reported childhood sexual abuse: An empirical study. *Psychological Science*, 11, 333–337.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Thousand Oaks, CA: Sage.
- Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as an antecedent of young children's school adjustment: An examination of mediating processes. *Developmental Psychology*, 37, 550–560.
- Butler, J. K. (2001). Reciprocity of dyadic trust in close male-female relationships. *Journal of Social Psychology*, 126, 579–591.
- Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). New York, NY: Routledge.
- Byrne, B. M. (2010). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.
- Byrne, B. M. (2016). *Structural equation modeling with Amos: Basic concepts, applications, and programming* (3rd ed.). New York, NY: Routledge.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for equivalence of factor covariance and means structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466.
- Caemmerer, J. M. (2017). *Beyond individual tests: The effects of children's and adolescents' cognitive abilities on their achievement*. Doctoral dissertation, University of Texas, Austin, TX.
- Caemmerer, J. M., & Keith, T. Z. (2015). Longitudinal, reciprocal effects of social skills and achievement from kindergarten to eighth grade. *Journal of School Psychology*, 53(4), 265–281. doi:10.1016/j.jsp.2015.05.001
- Caemmerer, J. M., Maddocks, D. L. S., Keith, T. Z., & Reynolds, M. R. (2018). Effects of cognitive abilities on child and youth academic achievement: Evidence from the WISC-V and WIAT-III. *Intelligence*, 68, 6–20.
- Carroll, J. B. (1963). A model for school learning. *Teachers College Record*, 64, 723–733.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.
- Carter, S. P., Greenberg, K., & Walker, M. S. (2017). The impact of computer usage on academic performance: Evidence from a randomized trial at the United States Military Academy. *Economics of Education Review*, 56, 118–132. doi:10.1016/j.econedurev.2016.12.005
- 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.
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J. R., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36, 462–494.
- Chen, H.-Y., Keith, T. Z., Chen, Y.-H., & Chang, B.-S. (2009). What does the WISC-IV measure for Chinese students? Validation of the scoring and CHC-based interpretive approaches in Taiwan. *Journal of Research in Education Sciences*, 54(3), 85–108.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.

- Christenson, S. L., Rounds, T., & Gorney, D. (1992). Family factors and student achievement: An avenue to increase students' success. *School Psychology Quarterly, 7*, 178–206.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124.
- Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 18*, 115–126.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–443.
- Cohen, J. (1978). Partial products are interactions; partialled powers are curve components. *Psychological Bulletin, 85*, 114–128.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.
- Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114*, 174–184.
- Coleman, J. S., Hoffer, T., & Kilgore, S. (1981). *Public and private schools*. Washington, DC: U.S. Department of Education.
- Cooper, H. (1989). *Homework*. New York, NY: Longman.
- Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research, 76*, 1–62. doi:10.3102/00346543076001001
- Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.
- Cudek, R. (1989). Analysis of correlation matrices using covariance structure models. *Multivariate Behavioral Research, 27*, 269–300.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170–180. doi:10.1037/0003-066X.60.2.170
- Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*(3), 138–146. doi:10.1111/j.1742-9536.2011.00037.x
- Curran, P. J., Stice, E., & Chassin, L. (1997). The relation between adolescent alcohol use and peer alcohol use: A longitudinal random coefficients model. *Journal of Consulting and Clinical Psychology, 65*(1), 130–140. doi:10.1037/0022-006X.65.1.130
- Darlington, R. B. (1990). *Regression and linear models*. New York, NY: McGraw-Hill.
- Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. New York, NY: Guilford.
- DiPerna, J. C., Lei, P.-W., & Reid, E. E. (2007). Kindergarten predictors of mathematical growth in the primary grades: An investigation using the Early Childhood Longitudinal Study—Kindergarten cohort. *Journal of Educational Psychology, 99*, 369–379. doi:10.1037/022-0663.99.2.369
- Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16*, 939–944.
- Duncan, O. D. (1975). *Introduction to structural equation models*. New York, NY: Academic Press.
- Duncan, O. D., Haller, A. O., & Portes, A. (1971). Peer influences on aspirations: A reinterpretation. In H. M. Blalock (Ed.), *Causal models in the social sciences* (pp. 219–244). New York, NY: Aldine.
- Duncan, S. C., Duncan, T. E., Biglan, A., & Ary, D. (1998). Contributions of the social context to the development of adolescent substance use: A multivariate latent growth modeling approach. *Drug and Alcohol Dependence, 50*, 57–71.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and application* (2nd ed.). Mahwah, NJ: Erlbaum.
- Eberhart, S. W., & Keith, T. Z. (1989). Self-concept and locus of control: Are they causally related in secondary students? *Journal of Psychoeducational Assessment, 7*, 14–30.
- Eisenberg, N., Gershoff, E. T., Fabes, R. A., Shepard, S. A., Cumberland, A. J., Losoya, S. H., . . . Murphy, B. C. (2001). Mothers' emotional expressivity and children's behavior problems and social competence: Mediation through children's regulation. *Developmental Psychology, 37*, 475–490.
- Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07–096). Newbury Park, CA: Sage.

- Elkins, G., Marcus, J., Stearns, V., Perfect, M., Rajab, M. H., Ruud, C., . . . Keith, T. Z. (2008). Randomized trial of a hypnosis intervention for treatment of hot flashes among breast cancer survivors. *Journal of Clinical Oncology*, 26, 5022–5026.
- Elliott, C. D. (2007). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Harcourt Assessment.
- Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 313–342). Greenwich, CT: Information Age.
- Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8, 430–457.
- Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve accuracy of inferences. *Structural Equation Modeling*, 11, 1–19.
- Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indexes to model misspecification and model types. *Multivariate Behavioral Research*, 42, 509–529.
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6, 56–83.
- FAQ: How do I interpret odds ratios in logistic regression? Retrieved May 31, 2018, from <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(1149–1160).
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, 40, 935–952.
- Ferrer, E., & McArdle, J. J. (2010). Longitudinal modeling of developmental changes in psychological research. *Current Directions in Psychological Science*, 19, 149–154. doi:10.1177/0963721410370300
- Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: Information Age.
- Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling*, 13, 456–486.
- Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Thousand Oaks, CA: Sage.
- Fredrick, W. C., & Walberg, H. J. (1980). Learning as a function of time. *Journal of Educational Research*, 73, 183–204.
- Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics*, 12, 101–128.
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18, 233–239.
- Gage, N. L. (1978). *The scientific basis of the art of teaching*. New York, NY: Teachers College Press.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549–576.
- Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer-Verlag.
- Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26, 499–510.
- Green, S. B., & Thompson, M. S. (2006). Structural equation modeling for conducting tests of differences in multiple means. *Psychosomatic Medicine*, 68, 706–717.
- 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, S78–S94.
- Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, 30, 91–105.
- 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.). Charlotte, NC: Information Age.
- Hancock, G. R., & Mueller, R. O. (Eds.). (2013). *Structural equation modeling: A second course*. Charlotte, NC: Information Age.

- Hancock, G. R., Harring, J.R., & Lawrence, F. R. (2013). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 171–196). Charlotte, NC: Information Age.
- Hansen, C. P. (1989). A causal model of the relationship among accidents, biodata, personality, and cognitive factors. *Journal of Applied Psychology*, 74, 81–90.
- Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: Johns Hopkins University Press.
- Hayduk, L. A. (1996). *LISREL issues, debates, and strategies*. Baltimore, MD: Johns Hopkins University Press.
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). New York, NY: Guilford.
- Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus* (3rd ed.). New York, NY: Routledge.
- Heck, R. H., Thomas, S. L., & Tabata, L. N. (2014). *Multilevel and longitudinal modeling with IBM SPSS* (2nd ed.). New York, NY: Routledge.
- Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, 19(1), 36–50. doi:10.1080/10705511.2012.634710
- Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2001). Longitudinal family and peer group effects on violence and nonviolent delinquency. *Journal of Clinical Child Psychology*, 30, 172–186.
- Hershberger, S. L. (2006). The problem of equivalent structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course*. Greenwich, CT: Information Age.
- Hintze, J. M., Callahan, J. E., III, Matthews, W. J., Williams, S. A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review*, 31, 540–553.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. S. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.
- Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth.
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, NY: Routledge.
- Hoyle, R. H. (Ed.). (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.
- Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. New York, NY: Guilford.
- Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage.
- 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.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, 3, 518–531.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409–426.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631–639.
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Erlbaum.
- Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Lincolnwood, IL: Scientific Software.
- Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5, 602–619.
- Kaplan, D. (1995). Statistical power in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100–117). Thousand Oaks, CA: Sage.
- Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Los Angeles, CA: Sage.
- Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (2nd ed.: Technical manual. Circle Pines, MN: American Guidance Service.
- Keith, T. Z. (1993). Causal influences on school learning. In H. J. Walberg (Ed.), *Analytic methods for educational productivity* (pp. 21–47). Greenwich, CT: JAI Press.

- Keith, T. Z., & Benson, M. J. (1992). Effects of manipulable influences on high school grades across five ethnic groups. *Journal of Educational Research*, 86, 85–93.
- Keith, T. Z., & Cool, V. A. (1992). Testing models of school learning: Effects of quality of instruction, motivation, academic coursework, and homework on academic achievement. *School Psychology Quarterly*, 7(3), 207–226.
- Keith, T. Z., & et al. (1986). Parental involvement, homework, and TV time: Direct and indirect effects on high school achievement. *Journal of Educational Psychology*, 78(5), 373–380.
- Keith, T. Z., & Reynolds, M. R. (2018). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed.). New York: Guilford.
- Keith, T. Z., Caemmerer, J. M., & Reynolds, M. R. (2016). Comparison of methods for factor extraction for cognitive test-like data: Which overfactor, which underfactor? *Intelligence*, 54, 37–54. doi:10.1016/j.intell.2015.11.003
- Keith, T. Z., Diamond-Hallam, C., & Fine, J. G. (2004). Longitudinal effects of in-school and out-of-school homework on high school grades. *School Psychology Quarterly*, 19, 187–211.
- Keith, T. Z., Keith, P. B., Troutman, G. C., Bickley, P. G., Trivette, P. S., & Singh, K. (1993). Does parental involvement affect eighth-grade student achievement? Structural analysis of national data. *School Psychology Review*, 22, 474–496.
- Keith, T. Z., Kranzler, J. H., & Flanagan, D. P. (2001). What does the Cognitive Assessment System (CAS) measure? Joint confirmatory factor analysis of the CAS and the Woodcock-Johnson Tests of Cognitive Ability (3rd edition). *School Psychology Review*, 30, 89–119.
- Keith, T. Z., Low, J. A., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2010). Higher-order factor structure of the Differential Ability Scales—II: Consistency across ages 4 to 17. *Psychology in the Schools*, 47, 676–697. doi:10.1002/pits.20498
- Keith, T. Z., Reimers, T. M., Fehrmann, P. G., Pottebaum, S. M., & Aubey, L. W. (1986). Parental involvement, homework, and TV time: Direct and indirect effects on high school achievement. *Journal of Educational Psychology*, 78, 373–380.
- Kenny, D. A. (1979). *Correlation and causality*. New York, NY: Wiley.
- Kenny, D. A. (2008). Reflections on mediation. *Organizational Research Methods*, 11, 353–358.
- Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, 96, 201–210.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2011). *The performance of RMSEA in models with small degrees of freedom*. Unpublished manuscript. University of Connecticut.
- Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY: Holt, Rinehart and Winston.
- Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.
- Klecka, W. R. (1980). *Discriminant analysis*. Thousand Oaks, CA: Sage.
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65(4), 457–474. doi:10.1007/bf02296338
- Kline, R. B. (1998). *Principles and practices of structural equation modeling*. New York, NY: Guilford.
- Kline, R. B. (2006). Reverse arrow dynamics: Formative measurement and feedback loops. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 43–68). Greenwich, CT: Information Age.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.
- Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin*, 125, 470–500.
- Kohn, A. (2006). *The homework myth: Why our kids get too much of a bad thing*. Cambridge, MA: Da Capo Press.
- Kranzler, J. H. (2018). *Statistics for the terrified* (6th ed.). Lanham, MD: Rowman & Littlefield.
- Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly*, 14, 327–342.
- Krivo, L. J., & Peterson, R. D. (2000). The structural context of homicide: Accounting for racial differences in process. *American Sociological Review*, 65, 547–559.
- Lee, S., & Hershberger, S. L. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research*, 25, 313–334.
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.
- Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.

- 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.
- Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural analysis* (4th ed.). Mahwah, NJ: Erlbaum.
- Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural analysis* (5th ed.). New York, NY: Routledge.
- Lott, J. R. (2010). *More guns, less crime: Understanding crime and gun-control laws* (3rd ed.). Chicago, IL: University of Chicago Press.
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100, 107–120.
- 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, NY: Psychology Press.
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effects. *Prevention Science*, 1(4), 173–181.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Magidson, J., & Sörbom, D. (1982). Adjusting for confounding factors in quasi-experiments: Another reanalysis of the Westinghouse Head Start evaluation. *Educational Evaluation and Policy Analysis*, 4, 321–329.
- Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence*. doi:10.1016/j.intell.2017.01.012
- Marcoulides, G. A., & Schumacker, R. E. (2001). *New developments and techniques in structural equation modeling*. Mahwah, NY: Erlbaum.
- Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal*, 30, 841–860.
- Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2017). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis Comparison of latent means across many groups. *Psychological Methods*. doi:10.1037/met0000113
- 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., 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., & Nagengast, B. (2013). Structural equation models of latent interaction and quadratic effects. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 267–308). Greenwich, CT: Information Age.
- Maruyama, G. M. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage.
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12(1), 23–44. doi:10.1037/1082-989X.12.1.23
- Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, 11, 344–362. doi:10.1037/1082-989X.11.4.344
- McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, 29, 409–454.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60(1), 577–605. doi:10.1146/annurev.psych.60.110707.163612
- McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the rectilinear action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, 37, 234–251.
- McArdle, J. J., Hamagami, F., Meredith, W., & Bradway, K. P. (2000). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning & Individual Differences*, 12, 53–79.
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82.
- McManus, I. C., Winder, B. C., & Gordon, D. (2002). The causal links between stress and burnout in a longitudinal study of UK doctors. *Lancet*, 359, 2089–2090.

- 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.
- Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, 5(1), 23–43. doi:10.1037/1082-989x.5.1.23
- Mels, G. (2006). *LISREL for Windows: Getting started guide*. Lincolnwood, IL: Scientific Software International, Inc.
- Menard, S. (1997). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.
- Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.
- Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44(11, Suppl 3), S69-S77.
- 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). Invariance in measurement and prediction revisited. *Psychometrika*, 72, 461–473. doi:10.1007/s11336-007-9039-7
- Morris, W. (Ed.). (1996). *The American heritage dictionary of the English language*. Boston, MA: Houghton Mifflin.
- Mueller, R. O. (1995). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. New York, NY: Springer-Verlag.
- Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman & Hall/CRC.
- Mulaik, S. A., & Millsap, R. E. (2000). Doing the four-step right. *Structural Equation Modeling*, 7, 36–73.
- Mulaik, S. A., & Quartetti, D. A. (1997). First-order or higher-order general factor? *Structural Equation Modeling*, 4, 193–211.
- Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41, 407–422. doi:10.1016/j.intell.2013.06.004
- Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). New York, NY: Routledge.
- Muthén, B. O., & Asparouhov, T. (2015). *Latent variable interactions*. Retrieved from www.statmodel.com.
- Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431–462.
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (2002). *How to use a Monte Carlo study to decide on sample size and determine power*. Retrieved March 26, 2003, from <http://www.statmodel.com/index2.html>
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: U.S. Government Printing Office.
- Neyt, B., Omey, E., Baert, S., & Verhaest, D. (in press). Does student work really affect educational outcomes? A review of the literature. *Journal of Economic Surveys*. doi:10.1111/joes.12301
- Nurss, J. R., & McGauvran, M. E. (1986). *The Metropolitan Readiness Tests*. New York, NY: Psychological Corporation.
- Page, E. B., & Keith, T. Z. (1981). Effects of U.S. private schools: A technical analysis of two recent claims. *Educational Researcher*, 10(7), 7–17.
- Park, J. S., & Grow, J. M. (2008). The social reality of depression: DTC advertising of antidepressants and perceptions of the prevalence and lifetime risk of depression. *Journal of Business Ethics*, 97, 379–393. doi:10.1007/s10551-007-9403-7
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York, NY: Cambridge University Press.
- Pearl, J., & MacKenzie. (2018). *The book of why: The new science of cause and effect*. New York: NY: Basic Books.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. New York, NY: Wiley.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Prediction and explanation* (3rd ed.). New York, NY: Holt, Rinehart & Winston.
- Peugh, J. L., & Enders, C. K. (2005). Using the SPSS Mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement*, 65, 717–741.

- Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. New York, NY: Routledge.
- Pituch, K. A., Whittaker, T. A., & Stevens, J. P. (in press). *Intermediate statistics: A modern approach* (4th ed.). New York, NY: Routledge.
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66(1), 825–852. doi:10.1146/annurev-psych-010814-015258
- Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA. Retrieved from <http://quantpsy.org/>
- Quirk, K. J., Keith, T. Z., & Quirk, J. T. (2001). Employment during high school and student achievement: Longitudinal analysis of national data. *Journal of Educational Research*, 95, 4–10.
- Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, 23, 99–115.
- Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2017). *A User's Guide to MLwiN, v3.01*. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.
- Rasberry, W. (1987, December 30). Learn what the smart kids learn? *Washington Post*, p. A23.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & du Toit, M. (2011). *HLM 7: Linear and nonlinear modeling*. Skokie, IL: Scientific Software International, Inc.
- Reibstein, D. J., Lovelock, C. H., & Dobson, R. DeP. (1980). The direction of causality between perceptions, affect, and behavior: An application to travel behavior. *Journal of Consumer Research*, 6, 370–376.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47, 667–696. doi:10.1080/00273171.2012.715555
- Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, 51, 818–838.
- Rentzsch, K., Wenzler, M. P., & Schütz, A. (2016). The structure of multidimensional self-esteem across age and gender. *Personality and Individual Differences*, 88, 139–147. doi:10.1016/j.paid.2015.09.012
- Reynolds, M. R., & Keith, T. Z. (2013). Measurement and statistical issues in child psychological assessment. In D. H. Soklofske, V. L. Schwean & C. R. Reynolds (Eds.), *Oxford handbook of child and adolescent assessment* (pp. 48–83). New York, NY: Oxford University Press.
- Reynolds, M. R., & Keith, T. Z. (2017). Multi-group and hierarchical confirmatory factor analysis of the Wechsler Intelligence Scale for Children—Fifth Edition: What does it measure? *Intelligence*. doi:10.1016/j.intell.2017.02.005
- Reynolds, M. R., & Turek, J. J. (2012). A dynamic developmental link between verbal comprehension-knowledge (Gc) and reading comprehension: Verbal comprehension-knowledge drives positive change in reading comprehension. *Journal of School Psychology*, 50(6), 841–863. doi:10.1016/j.jsp.2012.07.002
- Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2012). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology*, 51(4), 535–555.
- Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, 36, 236–260.
- Rhemtulla, M., & Little, T. D. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development*, 13, 425–438. doi:10.1080/15248372.2012.717340
- Rigdon, E. E. (1994). Demonstrating the effects of unmodeled random measurement error. *Structural Equation Modeling*, 1, 375–380.
- Rigdon, E. E. (1995). A necessary and sufficient identification rule for structural models estimated in practice. *Multivariate Behavioral Research*, 30, 359–383.
- Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, 23, 51–67.
- Rosenthal, R., & Rubin, D. B. (1979). A note on percent variance explained as a measure of the importance of effects. *Journal of Applied Social Psychology*, 9, 395–396.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Rubin, D. B. (1976). Inference and missing data. *Psychometrika*, 63, 581–592.

- Salzinger, S., Feldman, R. S., Ng-Mak, D. S., Mojica, E., & Stockhammer, T. F. (2001). The effect of physical abuse on children's social and affective status: A model of cognitive and behavioral processes explaining the association. *Development and Psychopathology*, 13, 805–825.
- Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 477–497. doi:10.1080/10705510903008238
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall.
- Schafer, J. L. (1999). *NORM user's guide: Multiple imputation of incomplete multivariate data under a normal model*. University Park, PA: The Methodology Center, Penn State. Retrieved from <http://methodology.psu.edu>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schmidt, F. L., & Hunter, J. E.. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.
- Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York, NY: Routledge.
- Schumacker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interactive and nonlinear effects in structural equation modeling*. Mahwah, NJ: Erlbaum.
- Sethi, S., & Seligman, M. E. P. (1993). Optimism and fundamentalism. *Psychological Science*, 4, 256–259.
- Shipley, B. (2000). *Cause and correlation in biology*. Cambridge, UK: Cambridge University Press.
- Simon, H. A. (1954). Spurious correlation: A causal interpretation. *Journal of the American Statistical Association*, 48, 467–479.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24, 323–355.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Method for studying change and event occurrence*. New York, NY: Oxford University Press.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13, 290–312.
- Sorjonen, K., Hemmingsson, T., Lundin, A., Falkstedt, D., & Melin, B. (2012). Intelligence, socio-economic background, emotional capacity, and level of education as predictors of attained socioeconomic position in a cohort of Swedish men. *Intelligence*, 40, 269–277. doi:10.1016/j.intell.2012.02.009
- Stapleton, L. M. (2013). Multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 521–562). Charlotte, NC: Information Age.
- Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling*, 5, 411–419.
- 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.
- Stelzl, I. (1986). Changing the causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, 21, 309–331.
- Stone, B. J. (1992). Joint confirmatory factor analyses of the DAS and WISC-R. *Journal of School Psychology*, 30, 185–195.
- Stoolmiller, M. (1994). Antisocial behavior, delinquent peer association, and unsupervised wandering for boys: Growth and change from childhood early adolescence. *Multivariate Behavioral Research*, 29, 263–288.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. S. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.
- Teigen, K. H. (1995). Yerkes-Dodson: A law for all seasons. *Theory and Psychology*, 4, 525–547.
- Thompson, B. (1998, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Invited address presented at the annual meeting of the American Educational Research Association, San Diego.
- Thompson, B. (1999, April). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap*. Invited address presented at the annual meeting of the American Educational Research Association, Montreal.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 25–32.
- Thompson, B. (2006). *Foundations of behavioral statistics*. New York, NY: Guilford.

- Tiggeman, M., & Lynch, J. E. (2001). Body image across the life span in adult women: The role of self-objectification. *Developmental Psychology, 37*, 243–253.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- Vandenberg, R. L., & 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.
- VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for Kindergarten students. *School Psychology Review, 30*, 363–382.
- Walberg, H. J. (1981). A psychological theory of educational productivity. In F. H. Farley & N. Gordon (Eds.), *Psychology and education* (pp. 81–110). Berkeley, CA: McCutchan.
- Walberg, H. J. (1986). Synthesis of research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 214–229). New York, NY: MacMillan.
- Wallis, C. (2006, August 29). The myth about homework. *Time*.
- Wampold, B. E. (1987). Covariance structures analysis: Seduced by sophistication? *Counseling Psychologist, 15*, 311–315.
- Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children* (Revised ed.). New York, NY: Psychological Corporation.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation.
- Wicherts, J. M., & Millsap, R. E. (2009). The absence of underprediction does not imply the absence of measurement bias. *American Psychologist, 64*, 281–283. doi:10.1037/a0014992
- Widaman, K. F. (2006). Missing data: What to do with or without them. In K. McCartney, M. R. Burchinal & K. L. Bub (Eds.), *Best practices in quantitative methods for developmentalists* (pp. 42–64). Monographs of the Society for Research in Child Development, 71 (3, Serial No. 285).
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.
- 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, P. A., Haertel, E. H., Haertel, G. D., & Walberg, H. J. (1982). The impact of leisure-time television on school learning: A research synthesis. *American Educational Research Journal, 19*, 19–50.
- Wolfe, L. M. (1979). Unmeasured variables in path analysis. *Multiple Linear Regression Viewpoints, 9*(5), 20–56.
- Wolfe, L. M. (1980). Strategies of path analysis. *American Educational Research Journal, 17*, 183–209.
- 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.
- Woithke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. Mahwah, NJ: Erlbaum.
- Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113–128.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Author Index

- Aiken, L. S. 133, 136–137, 140, 148, 155, 165, 167
Alexander, K. W. 110, 159
Alexander, R. A. 137
Alfonso, V. C. 578
Allison, P. D. 202
American Psychological Association 62, 196
Arbuckle, J. L. 297, 328, 460, 551, 574
Ary, D. 529
Aubey, L. W. 137
Bandalos, D. L. 453
Baron, R. M. 133, 180–181, 183–184, 289, 609
Beaujean, A. A. 297, 311, 314–315, 323, 326, 567, 574, 578
Bennett, W. J. 434
Benson, M. J. 426, 435
Bentler, P. M. 296, 308, 311, 326, 406, 458, 576
Berry, W. D. 202
Bickley, P. G. 288
Biglan, A. 529
Bilgic, R. 100
Birnbaum, M. H. 150
Blalock, H. M. 583
Boker, S. M. 297
Bollen, K. A. 262, 266, 326, 338, 574, 582
Boomsma, A. 578, 583
Borsboom, D. 150
Bradway, K. P. 529
Brady, H. V. 151
Brandmaier, A. 297
Bremner, J. D. 110, 127–128
Brick, T. R. 297
Brown, T. A. 479, 482, 488, 490, 495, 582
Browne, M. W. 311, 329, 578
Browne, W. J. 245
Bryk, A. S. 243, 245, 252
Buchner, A. 216
Buhner, M. 327
Buhs, E. S. 391–392, 395, 397, 398, 404–405
Buswell, B. N. 130
Butler, J. K. 322–323
Byrne, B. M. 482, 489–490, 582
Callahan, J. E. III 149
Carothers, A. 492
Carroll, J. B. 18, 351, 369, 372
Chassin, L. 532
Chen, F. F. 214, 326, 378–379
Chen, H. Y. 499
Cheung, G. W. 315, 483, 489, 492
Christenson, S. L. 284
Cleary, T. A. 160
Cliff, N. 583
Coffman, D. L. 378, 578
Cohen, J. 15, 18, 40, 56, 62, 63, 101, 111, 117, 126,
133–134, 136, 155, 165, 167, 174, 202–203, 204, 205,
206, 208, 215, 216, 221, 239, 241, 588, 596, 601–602
Cohen, P. 40, 56, 63, 101, 117, 126, 133–134, 155, 165, 167,
174, 202–203, 204, 205, 206, 208, 215, 216, 221, 241,
596, 602
Coleman, J. S. 286
Congdon, R. 245
Cool, V. A. 222, 268
Cooper, H. 4, 169
Cronbach, L. J. 153, 155
Cudeck, R. 306, 311, 329
Cumberland, A. J. 440–441
Curran, P. J. 326, 532, 582
Darlington, R. B. 40, 43, 54, 88, 89, 100, 107, 126, 128, 133,
137, 155, 201, 204, 208–209, 213, 215–216, 235, 241,
286, 612
Deary, I. J. 492

- DeShon, R. P. 137
 Diamond-Hallam, C. 10, 57, 62
 DiPerna, J. C. 513, 517, 527, 529
 DiStefano, C. 580
 Dobson, R. DeP. 323
 Duckworth, A. L. 107
 Duncan, O. D. 292, 323, 405
 Duncan, S. C. 529–530, 582
 Duncan, T. E. 529–530, 582
- Eberhart, S. W. 85
 Edelstein, R. S. 110, 159
 Edwards, J. E. 100, 107
 Eisenberg, N. 440–441
 Eliason, S. R. 574
 Elkins, G. 456, 460
 Elliott, C. D. 349, 379
 Enders, C. K. 252, 453, 576–577, 578
 Erdfelder, E. 216
- Fabes, R. A. 440–441
 Fabrigar, L. R. 320
 Falkstedt, D. 407
 Fan, X. 311, 326
 FAQ 241
 Faul, F. 216
 Fehrman, P. G. 137, 162
 Feldman, R. S. 275
 Ferrer, E. 528, 579–580
 Fine, J. G. 10, 57, 62
 Finney, S. J. 580
 Flanagan, D. P. 54, 578
 Fleer, P. F. 100, 107
 Fox, J. 213
 Fredrick, W. C. 169
 Freedman, D. A. 583
 Freudenthaler, H. H. 327
- Gage, N. L. 88
 Gershoff, E. T. 440–441
 Ghetti, S. 110, 159
 Goldstein, H. 245
 Goodman, G. S. 110, 159
 Gordon, D. 323
 Gorman-Smith, D. 330–331
 Gorney, D. 284
 Graham, J. W. 453, 576–577
 Green, S. B. 225, 460
 Gregorich, S. B. 489–490
 Grow, J. M. 75
- Haertel, E. H. 162
 Haertel, G. D. 162
 Haller, A. O. 323
 Hamagami, F. 529
- Hancock, G. R. 448, 509, 529, 578
 Hansen, C. P. 278–279
 Hau, K. T. 308, 326, 536
 Hayduk, L. A. 311, 323, 380, 382, 579
 Heck, R. H. 252
 Heene, M. 327
 Hemmingsson, T. 407
 Henry, D. B. 330–331
 Hershberger, S. 320
 Hilbert, S. 327
 Hintze, J. M. 149
 Ho, M.-H. R. 583
 Hoffer, T. 286
 Hoffman, J. M. 181
 Hollis, M. 453
 Hosmer, D. W. 241
 Howell, D. C. 15, 107, 114, 216, 221, 587, 599, 604
 Hoyle, R. H. 308, 581–582
 Hox, J. J. 242–243, 248, 252
 Hu, L. 308, 311, 326, 458
 Hyde, J. S. 130
- Jackson, D. L. 578
 Jensen, A. R. 267
 Johnson, W. 378–379, 492
 Jones, D.P.H. 110, 159
 Jordan, L. 144
 Jöreskog, K. G. 296, 369, 428, 495
- Kaplan, D. 453, 578, 582
 Kaufman, A. S. 476
 Kaufman, N. L. 476
 Keith, P. B. 288
 Keith, T. Z. 10, 54, 57, 62, 85, 137, 161–162, 165, 222, 259, 268, 286, 288, 328, 353, 369, 372, 379, 384, 425, 426, 435, 475, 489–490, 499, 541, 578
 Kenny, D. A. 15, 71, 72, 133, 180–181, 192, 201, 259, 262, 263, 265, 266, 280, 282, 289, 308, 315, 328, 334, 388, 406, 458–459, 536, 609
 Kerlinger, F. N. 12, 193
 Kihlstrom, J. F. 110
 Kilgore, S. 286
 Kirby, J. 326
 Kirk, R. E. 114
 Klecka, W. R. 241
 Kline, R. B. 187, 200, 201, 263, 266, 307, 320, 323, 326, 351, 354, 368, 384, 438, 495, 517, 530, 544, 578, 580, 581, 583
 Kling, K. C. 130
 Kohn, A. 4
 Kranzler, J. H. 54, 144–145, 159, 604
 Krivo, L. J. 141
- Ladd, G. W. 391–392, 395, 397–398, 404–405, 568
 Lance, C. E. 489, 490, 492, 512

- Lang, A.-G. 216
 Lawrence, F. R. 529
 Lee, S. 316, 320
 Lei, P.-W. 513, 517
 Li, Y. 499
 Little, T. D. 324, 425, 492–493, 499, 578
 Lockwood, C. M. 181
 Loechl, J. C. 311, 314, 315, 323, 326, 567, 574, 578, 581
 Lomax, R. G. 581
 Long, J. S. 582
 Losoya, S. H. 440–441
 Lott, J. R. 118
 Lovelock, C. H. 323
 Low, J. A. 353, 372
 Lundin, A. 407
 Lynch, J. E. 74
 McArdle, J. J. 447, 528–530, 578, 579–580, 582
 MacCallum, R. C. 311, 320, 369, 578, 583
 McDonald, R. P. 447, 583
 McGauvrani, M. E. 392
 McLeod, L. D. 378
 MacKinnon, D. P. 181, 289
 McManus, I. C. 323–324
 Maes, H. H. 297
 Magidson, J. 406
 Marcoulides, G. A. 435, 544
 Marcus, J. 456, 460
 Marsh, H. W. 308, 326, 433, 489, 536, 537, 544
 Maruyama, G. M. 581
 Matthews, W. J. 149
 Maydeu-Olivares, A. 378
 Mehta, P. D. 529
 Melin, B. 407
 Mels, G. 296
 Menard, S. 241
 Meredith, W. 481, 489, 490, 529
 Miller, M. D. 144
 Millsap, R. E. 150, 357, 390
 Mojica, E. 275
 Morris, W. 266
 Mueller, R. O. 448, 582
 Mulaik, S. A. 328, 379, 390, 582
 Murphy, B. C. 440–441
 Murray, A. 378–379
 Muthén, B. O. 246, 296, 453, 482, 489
 Muthén, L. K. 246, 296
 National Commission on Excellence in Education 434
 Neale, M. C. 297
 Ng-Mak, D. S. 275
 Nurss, J. R. 392
 Page, E. B. 286
 Panter, A. T. 583
 Park, J. S. 75
 Patall, E. A. 4
 Patel, P. G. 353, 372, 475–476
 Paxton, P. 326
 Pearl, J. 263, 289, 580
 Pedhazur, E. 15, 24, 56, 89, 118, 126, 148, 155, 209,
 213, 216
 Perfect, M. 456, 460
 Peterson, R. D. 141
 Peugh, J. L. 252, 576
 Portes, A. 323
 Pottebaum, S. M. 137, 161–162
 Preacher, K. C. 181, 578
 Quas, J. A. 110, 159
 Quirk, J. T. 173
 Quirk, K. J. 173
 Rajab, M. H. 456, 460
 Raju, N. S. 100, 107
 Rasbash, J. 245
 Rasberry, W. 434
 Raudenbush, S. W. 243, 245, 252
 Redlich, A. D. 110, 159
 Reibstein, D. J. 323
 Reid, E. E. 513, 517
 Reimers, T. M. 137, 162
 Reise, S. P. 378, 379, 490, 493, 512
 Rensvold, R. B. 315, 483, 489, 492
 Reynolds, M. R. 328, 353, 369, 379, 384, 475–476,
 489–490, 499, 500, 509, 530, 541, 578
 Rhemtulla, M. 578
 Richman, L. C. 151
 Ridley, K. H. 353, 372, 475–476
 Rigdon, E. E. 323, 338
 Rindskopf, D. 372
 Robinson, J. C. 4
 Rose, T. 372
 Rosenthal, R. 88
 Rosseel, Y. 297
 Rounds, T. 284
 Rubin, D. B. 88, 574
 Ruud, C. 456, 460
 Salzinger, S. 275
 Savalei, V. 576
 Sayer, A. G. 529
 Schafer, J. L. 453, 577
 Schumacker, R. E. 435, 544, 581–582
 Seligman, M. E. P. 106–107
 Sethi, S. 106
 Shepard, S. A. 440–441
 Shipley, B. 580
 Singer, J. D. 252, 530
 Sivo, S.A. 326

632 • AUTHOR INDEX

- Shobe, K. K. 110
Showers, C. J. 130
Simon, H. A. 283
Singh, K. 288
Snow, R. E. 153, 155
Sobel, M. E. 181
Sörbom, D. 296, 363, 369, 406, 428
Sorjonen, K. 407–408
Sousa, K. H. 378–379
Spiegel, M. 297
Stapleton, L. M. 204, 545, 546, 547, 549, 553,
 554, 557
Stearns, B. 456, 460
Steele, F. 245
Steiger, J. H. 306, 429, 479, 583
Stelzl, I. 316
Stice, E. 532
Stockhammer, T. F. 275
Stone, B. J. 384
Stoolmiller, M. 530
Strycker, L. A. 530, 582
Sugawara, H. M. 311, 578

Tabata, L. N 252
Tanaka, J. S. 311
Teigen, K. H. 168
Teresi, J. A. 481, 490
Thissen, D. 378
Thomas, S. L. 252
Thompson, B. 13, 15, 54, 56, 62, 100, 107, 200, 235, 239,
 241, 311, 326, 460
Tiggeman, M. 74
Tobin, K. G. 149

Tolan, P. H. 330–331
Trivette, P. S. 288
Troutman, G. C. 288
Tufte, E. R. 68

Uchino, B. N. 320

Vandenberg, R. L. 489, 490, 492, 512

Walberg, H. J. 18, 65, 162, 269, 262, 267
Wallis, C. 4
Wampold, B. E. 583
Wang, L. 311, 326
Wechsler, D. 384
Wegener, D. T. 320
Weiss, L. 499
Wen, Z. 308, 326, 536
West, S. G. 133, 136–137, 140, 148, 155, 165, 167, 181, 289,
 378–378, 529
Wichert, J. M. 150
Widaman, K. F. 490, 493, 512, 536
Wilde, M. J. 297
Willett, J. B. 529–530
Williams, P. A. 162, 165
Williams, S. A. S. 149
Winder, B. C. 323
Wolfle, L. M. 98, 199, 338, 363, 581
Woodward, J. A. 406
Wothke, W. 453

Yung, Y. F. 378–379

Zhu, J. 499

Subject Index

Note: Page numbers in italics indicate figures and in bold indicate tables on the corresponding pages.

- Adjustment latent variable model 399, 399–402, **400**, 401
Akaike Information Criterion (AIC) **314**, 315, **327**
Amos (Analysis of Moment Structures) 297, 580; estimating parent involvement model via 299–302, 299–303
analysis of covariance (ANCOVA) 153–154, **154**, 195
analysis of variance (ANOVA) 3, 4; analyzing growth data using 530; categorical variables and 111, **112**; cognitive behavioral therapy (CBT) effects 600–603, **602**, 603; consistency with the *t* test 601–602; factorial 604
Aptitude-Treatment Interactions (ATIs) 150–153, **151**–**152**; steps for testing 152–153; verbal skills and memory strategies 150–151, **151**–**152**
assumptions and regression diagnostics 54, 201–216; assumptions underlying regression 201–202; diagnosing data problems **208**, 208–216, **210**, **211**, 213–215; diagnosing violations of 202–208, **203**, 205–207
Attribute-Treatment Interactions (ATIs) *see* Aptitude-Treatment Interactions (ATIs)
- b* versus β 36–38, **38**, *76n2*
Baron and Kenny causal steps 180, **180**, **184**
best fitting 52
 β : *b* versus 36–38, **38**, *76n2*; direct calculation of R^2 and 41–42
bias: predictive 142–150, 160n4; test 142–146, 142–148, 148; statistical 53, 201, 282; construct, 483, 488
bifactor model: justification and setup 373, 374; results 374–376, **375**
block entry using sequential regression 90–92, 91
bootstrapping 181–182, 186–187
categorical variables 108, 126–127, **127**, 129–130, 158, 199, 580; ANOVA and follow-up 111, **112**; caveats and additional information about 154–158, **156**–**157**; criterion scaling 118–119, **119**; dummy variables 109–114, 109–116, **111**, **116**, 128n2; effect coding **117**, **117**, 117–118; effects of 154–155; false memory and sexual abuse 110–111, **111**; family structure and substance use **120**–**123**, 120–125, **123**–**124**, 125, 128n7; interactions with continuous variables (*see* interactions between continuous and categorical variables); more complex 110, **111**; post hoc probing 114–115; regression analysis with dummy variables 112–113, **112**–**114**; regression using sex, achievement, and self-esteem 130–131, **130**–**132**; simple 109–110, **109**–**110**; unequal group sizes and **120**–**123**, 120–125, **123**–**124**, 125
causal inference 263
causality 19–20; in path analysis 263–264; and veracity of models 580–581
causal language 192–193
causal steps mediation 180, **180**, **184**
centering and cross products: interactions between continuous and categorical variables 133–134, **134**, 158, 160n2; interactions between continuous variables 162, 162–163
cognitive behavior therapy (CBT) 599–601, **600**, **602**, **603**
Cohen's f^2 63, 89, **602**, 616
combined methods in multiple regression 105
common causes 187–192, **188**–**192**, 200; dangers of, in path analysis 282–289, **283**–**285**, 287, 295n2–3, 563–564; indirect effects and 68–70, 69; not including all causes 286, 287; omitted, in latent variable models 435–436, **436**–**437**; true experiments and 287, 287–288
comparative fit index (CFI) 309, 312; comparing competing models 312–315, **313**, **314**, **327**

- comparison across samples 38–40, 39
 comparison of competing models in SEM programs 312–315, **313, 314**
 conditional growth model 524–528, 525–527
 confidence intervals (CI) 4, 13–14, 14, 196, 593–594
 configural invariance 478–479, 478–481, **480**
 confirmatory factor analysis (CFA) 258, 343–345, 344, 347n2, 348–388, 384–385, 566, 566–567; additional uses of model constraints in 379–384, 380–383; correlated factors in 387–388n2; Differential Ability Scales, Second Edition (DAS-II) 349–358; hierarchical models 369–379, 370–371, 373–375, 377; higher-order models 500–505, **501–504, 502**; invariance testing with means 475–505; invariance testing without means 499–500; with latent means and invariance (*see* invariance testing with means); as measurement model 348–349; model fit and model modification 363–369, 364, **366, 368**; single-group, mimic models 505–509, 506–508; testing competing models 358–363, 359–360, **361, 362**; testing plausible cross-loadings 358–360, 359–360; three-factor combined nonverbal model 360–363, **361, 362**
- constraints, confirmatory factor analysis model 368–369, 379–384, 380–383
 continuous variables 199–200; categorizing 239–240, 240; curvilinear regression 168–174, 168–175; interactions between 161–168, 162–166, 534–544, 535–544; interactions between categorical and (*See* interactions between continuous and categorical variables)
 controlling for . . . multiple regression 35–36
 correlated disturbances 317, 324, 333n5
 correlated errors 412–413
 correlations 594–597, 595, 597; causality and 19–20; covariance and 20–21, **21**, 306, 306–307; inter- 78, **78, 79**; partial 36, 605–609, 606–608; in path analysis 306, 306–307; residual 367–368, **369**; semipartial 36, 107n1, 609–611, 609–612
 covariance and correlation 20–21, **21**; in multilevel modeling (MLM) 248, 248; in path analysis 306, 306–307
 criterion scaling 118–119, **119**
 cross-loading model, DAS-II 358–360, 359–360
 cross products and interactions 155, 160n2; multilevel modeling (MLM) **250, 250–251, 251**; *See also* centering and cross products
 cross-validation in stepwise regression 99, 107n2
 curriculum-based measurement (CBM) 144–146, 144–148, **148**
 curvilinear regression 168–174, 168–175; latent variable interactions 539–544, 540–544
- data files 585–586
 data problems, diagnosing 208, 208–209; distance 209; influence 212; leverage 209–212, **210, 211**; multicollinearity 213–215, 213–216; uses 212–213
- degrees of freedom 8, 75–76n1, 594; latent variable SEM 408n1; model fit and, SEM programs **307, 307–309**; in stepwise regression 100
 dependent variables 24n1; in logistic regression 232–235, 233–234
 diagnostics, regression 54, 202–216
 Differential Ability Scales, Second Edition (DAS-II) 349, 387n1; initial model 351–353, **352**; standardized and unstandardized results 353, 353–354; standardized model testing 354–358, 355–357; structure of 350, 350–351; testing plausible cross-loadings 358–360, 359–360; three-factor combined nonverbal model 360–363, **361, 362**
- direct effects, path analysis of 269–271, 270
 directionality in path analysis 320–321, 320–321
 discriminant analysis versus logistic regression 241
 distance data problem 209
 distributions 4, 589–592, 590–592
 dummy coding 109, 109–114, 109–116, **111, 116, 126–127, 127**
 dummy variables 126–127, **127, 128n2**; ANOVA and follow-up 111, 112; criterion scaling 118–119, **119**; effect coding **117, 117, 117–118**; g-1 dummy 115–116, **116**; post hoc probing 114–115; regression analysis with 112–113, **112–114**; simple categorical variables 109–110, 109–110; unequal group sizes **120–123, 120–125, 123–124, 125**
- Dunnett's test 114–115
 dynamic modeling 579–580
- effect coding **117, 117, 117–118**
 effects, magnitude of 62–63
 effect sizes 602, **602**
 equivalent models 315–320, **316–319**
 error(s) 334–335, 346–347, 565–566; effects of invalidity 339–343, 340–343; effects of unreliability 335–339, 335–339; of measurement and latent variable SEM 343–345, 344, 346; nonindependence of 204, 205; single indicators and correlated 409–412, 410, **411**
- exogenous and endogenous variables 266
 expectation-maximization (EM) algorithms 576–577
 explanation: conditional growth model 524–528, 525–527; versus prediction 19, 72–73, 104–105, 197; stepwise regression as inappropriate for 97–98
- exploratory factor analysis 349
- factor analysis: confirmatory (*see* confirmatory factor analysis (CFA)); exploratory 349
 factor covariances 493
 factorial ANOVA 604
 factorial invariance 489–490
 factor loading invariance 481, 481–483
 false memory and sexual abuse 110–111, **111**
 family structure and substance use **120–123, 120–125, 123–124, 125, 128n7**

- figural representation of multiple regression 34, 35
 Fisher least significant difference (LSD) post hoc procedure 115, 128n4
 forced entry regression 79
 formative measures 580
 formulae, useful 615–616
 F test 7–9
 full information maximum likelihood estimation (FIML) 574–578, 575
- g -1 dummy variables 115–116, 116
 G^* Power 216–217
 grades on homework and parent education multiple regression: assumptions of regression and regression diagnostics 54, 201–216; b versus β 36–38, 38, 76n2; cautions 40–41; common causes and indirect effects 68–70, 69; comparison across samples 38–40, 39; controlling for . . . 35–36; data on 27, 28–30; direct calculation of β and R^2 41–42; figural representation 34, 35; importance of R^2 with 70–72; interpretations 33–34; least squares 52–53, 52–53; multiple R 31, 31; partial and semipartial correlations 36; predicted scores and residuals 47–50, 48–49; regression calculation 27, 30; regression coefficients 32, 32–33; regression equation 54; regression line 50–51, 51; testing the difference between two regression coefficients 63–64, 63–64; three predictor variables 57–62, 58–61; why $R^2 \neq r^2 + r^2$ in 44–47, 45–47, 56n1
- hierarchical models: bifactor model justification and setup 373, 374; bifactor model results 374–376, 375; comparing the 376–379, 377; higher-order model justification and setup 369–370, 370; higher-order model results 370–373, 371, 373; total effects 372–373, 373
- higher-order models: justification and setup of 369–370, 370; KABC-II model 500–505, 501–504, 502; results of 370–373, 371, 373; total effects 372–373, 373
- homework and math achievement multiple regression: common causes and indirect effects 68–70, 69; four independent variables 64–68
- homework and math achievement simple bivariate regression: confidence intervals 13–14, 14; data in 4–6, 5; interpretation 10; regression analysis 6–8, 6–9; regression equation 9–10, 10; regression line 10–12, 11–12; standardized regression coefficient 14–15; statistical significance of regression coefficients in 12–13
- homework model, latent variable 409–412, 410, 411; multigroup 426–435, 427–429, 430, 431–434
- homoscedasticity 205, 206
- hypnosis for hot flashes latent means SEM: MG-MACS approach 460–466, 461, 463–465, 466; single group/dummy variable approach 456–460, 457–459, 460
- identification in path analysis 265–266
 independent variables 24n1; in causality 19; importance of R^2 with 70–72; multicollinearity and 213–215, 213–216; in multiple regression 18–19, 23, 64–68; multiple regression with four 64–68, 66, 67; negative correlation between 56n1; in simple bivariate regression 4, 7–11, 13, 16
- indirect effects: common causes and 68–70, 69; in latent variable SEM 398, 398–399; path analysis of total and 271–275, 272, 273–274; single indicators and correlated errors model 418, 418–419; using sequential regression to estimate total and 272–275, 273–274
- influence data problem 212
 initial rejection model, latent variable SEM 402, 402–404, 403
- interactions between continuous and categorical variables 132, 132–137, 134–136, 136; analysis of covariance (ANCOVA) analysis of 153–154, 154; aptitude-treatment 150–153, 151–152; centering and cross products 133–134, 134; cross products and 155, 158, 160n2; interpretation of 135–137, 136; MR analysis 134–135; predictive bias and 142–143, 142–144; specific types of 141–153, 142–146, 148, 151–152; statistically significant 137–140, 138–139, 140, 155–158, 156–157; test (and other) bias 142–146, 142–148, 148
- interactions between continuous variables 161–168, 162–166, 534–544, 535–544; curvilinear regression 168–174, 168–175; PROCESS 168; *see also* moderation
- intercept invariance 483–489, 484–487
- invalidity, effects of 339–343, 340–343; accounting for invalidity and 340–343, 341–343; and meaning and importance of validity 339–340, 340
- invariance testing with means 475–477, 476, 477, 572–573; configural invariance 478–479, 478–481, 480; intercept invariance 483–489, 484–487; measurement invariance steps 478–494; metric invariance 481, 481–483; residual invariance 489–490, 491; single-group, mimic models 505–509, 506–508; structural invariance 491–493, 492; variance/covariance matrix of measured variables 495–499, 496, 496–498
- invariance testing without means 499–500
- jargon and notation in path analysis 264–265, 264–266
- joint significance 181
- just-identified models 260, 265
- Kaufman Assessment Battery for Children-Second Edition (KABC-II) 476, 476–477, 476–477, 477, 480, 483, 490; higher-order model 500–505, 501–504, 502; single-group, mimic models 505–509, 506–508
- latent change score (LCS) modeling 579–580
- latent growth models (LGM) 513–517, 514–516, 515, 531, 531–532, 573, 573–574; conditional growth model

- 524–528, 525–527; data requirements 528–529; and other methods of analyzing growth data 530–531; unconditional, simple growth model 517–518, 517–523, 520–524; variations in model specifications 529–530
- latent means in SEM 444–445, 471–472, 570–572, 571–572; analyzing matrices versus raw data in 470, 470–471; calculating degrees of freedom in 454; comparing two methods in 466–470, 467–469; displaying means and intercepts in SEM and 445–448, 446–447; estimation of means and intercepts in single group SEM models and 448–452, 448–452; hypnosis for hot flashes example 456–471; MG-MACS approach 460–466, 461, 463–465, 466; missing values 453; overview of 454–456, 454–456; preparatory work 445–454, 446–452; single group/dummy variable approach, latent means in SEM 456–460, 457–459, 460
- latent variable interactions 559; between continuous variables 534–544, 535–544; testing curvilinear effects 539–544, 540–544
- latent variable models 438–439; latent variable homework model 409–412, 410, 411; multigroup models 426–435, 456, 460–471; omitted common causes in 435–436; panel models 424, 424–425, 426; path in the wrong direction in 437–438, 438; single indicators and correlated errors 409–425, 426
- latent variable SEM 343–345, 344, 346, 404–405, 567–570, 568–570; competing models 399, 399–402, 400, 401; on effects of peer rejection 391, 391–399, 392, 393–398; indirect and total effects 398, 398–399; initial model 395–398, 395–399; measurement model 393, 394; mediation 398; model modifications 402, 402–404, 403; overview, data, and model 391, 391–393, 392, 393; putting the pieces together in 389–391, 390; standardized results 396–397; structural model 393–395, 394; unstandardized findings 397, 397–398; *see also* confirmatory factor analysis (CFA)
- latent variables multigroup models 426–435, 427–429, 430, 431–434
- least squares 52–53, 52–53
- leverage data problem 209–212, 210, 211
- linearity assumption in regression 201
- locus of control *See* socioeconomic status (SES), previous grades, self-esteem, and locus of control
- logistic regression: appropriate uses of 240–241; categorizing continuous variables and 239–240, 240; conducting and understanding output of 235–239, 236–238; versus discriminant analysis 241; multiple regression analysis 228; predicting optimism versus pessimism 227, 227–228; transforming the dependent variable to log odds in 232–235, 233–234
- longitudinal models 323–324, 324–325, 579–580; latent variable panel model 424, 424–425, 426
- magnitude of effects 62–63; order of entry and 83–86, 84–85
- main effects and interactions in latent means in SEM 468–469, 468–470
- maximum likelihood estimation (MLE) 574
- mean(s) 588–589; invariance testing with 475–499, 572–573; invariance testing without 499–500; *see also* latent means in SEM
- measured variables, variance/covariance matrix of 495–499, 496, 496–498
- measurement model 345, 346, 348–349; latent variable SEM 393, 394
- measures of fit in SEM programs 308–315, 326–328, 327
- mediation 179–187, 200; bootstrapping and PROCESS 181–182, 186–187; causal steps, Baron and Kenny 180, 180, 184; example of 182–183, 182–187, 185–186; joint significance 181; latent variable SEM 398; Sobel test 181, 185–186
- metric invariance 481, 481–483
- MG-MACS approach 460–466, 461, 463–465, 466; compared to single group/dummy variable approach 466–470, 467–469
- MIMIC model 505–509, 506–508
- missing completely at random (MCAR), data 574–576, 575
- missing values 453, 574–578, 575
- model fit: and degrees of freedom in path analysis 307, 307–309; and model modification in confirmatory factor analysis 363–369, 364, 366, 368
- moderation 177–179, 178–179, 200; causal language and 192–193
- motivation to achievement, path analysis of 258–262, 258–262; cautions with 262–264; direct effects 269–271, 270; indirect and total effects 271–275, 272, 273–274; recursive and nonrecursive models in 264, 264–265; steps for conducting 266–269, 267, 269; using sequential regression to estimate total and indirect effects in 272–275, 273–274
- mPlus 297, 557–558, 557–559
- multicollinearity 213–215, 213–216
- multigroup models 426–435, 427–429, 430, 431–434; latent means SEM 454–456, 454–456
- multilevel modeling (MLM) 241–242, 252–253; adding a level 1 covariate in 249, 249–250; adding a level 2 covariate in 248, 248; adding the cross-product to test interaction of school-level and individual-level SES in 250, 250–251, 251; analyzing growth data using 530–531; different slopes in 245; of effects of SES on achievement 242–251; multilevel modeling of effects of 242–251; multiple regression analysis 242–243, 243; non-ML SEM alternative 557–558, 557–559; problems with MR for 221–225, 222–223, 224, 225; in SEM 544–559, 545–558, 548; separate regression lines by school 243–245, 244; unconditional model 247, 247

- multiple regression (MR) 26–27; advantages of 18–19; Aptitude-Treatment Interactions (ATIs) test 152–153; assumptions of regression and regression diagnostics 54, 201–216; *b* versus *B* 36–38, 38; categorical variables in (*see* categorical variables); cautions 40–41; combining methods in 105; common cause in 187–192, 188–192, 200; comparison across samples 38–40, 39; continuous variables in (*see* continuous variables); controlling for . . . 35–36; curvilinear 168–174, 168–175; direct calculation of *B* and R^2 41–42; effect of SES on achievement 242–245, 243–246; explanation versus prediction using 19, 72–73, 104–105, 197; figural representation 34, 35; with four independent variables 64–68, 66, 67; grades on homework and parent education 27–41; independent variables in 18–19, 23, 64–68; interactions between continuous variables 163, 163; intercorrelations among variables in 78, 78, 79; interpretations 33–34; least squares 52–53, 52–53; magnitude of effects 62–63; mediation in 179–187, 200; moderation in 177–179, 200; multicollinearity in 213–215, 213–216; multiple *R* 31, 31; partial and semipartial correlations 36; predicted scores and residuals 47–50, 48–49; prediction of optimism versus pessimism 228–232, 229–232; prediction versus explanation in 19; predictive bias 148–150; predictor variables 57–62, 58–61; problems with 221–225, 222–223, 224, 225; purposes of research and selection of type of 102–105, 103; regression coefficients 32, 32–33, 63–64, 63–64; regression equation 54; regression line 50–51, 51; sample size and power in 216–221, 217–221; versus SEM programs 325–326; sequential 81–95, 198; simultaneous 79–81, 198; stepwise 95–102, 198–199; summary of “standard” 195–200; test for interactions 132, 132–137, 134–136, 136; why $R^2 \neq r^2 + r^2$ in 44–47, 45–47, 56n1
- multivariate analysis of covariance (MANCOVA) 456–460
- National Education Longitudinal Study (NELS) 15–16, 16, 78, 128n7; data 21–23, 22, 241, 585–586; homoscedasticity 205, 206; nonindependence of errors and 204, 205; normality of residuals 206–208, 207; three predictor variables and 58, 59–60; unequal group sizes 120–123, 120–125, 123–124, 125, 128n7
- nonexperimental research 19–20, 23, 25n6
- nonindependence of errors 204, 205
- nonlinearity 202–204, 203
- nonnormed fit index (NNFI) 309
- nonrecursive models 264, 264–265; SEM programs and 322–323, 322–323
- normality of residuals 206–208, 207
- null model 309
- odds ratios 232–235, 233–234
- omitted common causes in latent variable models 435–436, 436–437
- Omnibus Tests of Model Coefficients 237
- OpenMx 298
- optimism versus pessimism, predicting 227, 227–228; categorizing continuous variables in 239–240, 240; logistic regression for predicting 235–239, 236–238; using multiple regression 228–232, 229–232
- order of entry and magnitude of effect 83–86, 84–85
- ordinary least squares (OLS) regression 52–53, 52–53
- overidentified models 265, 265, 303–307, 303–312
- panel models 324; latent variable 424, 424–425, 426
- parameters, number of 216–221, 217–221, 578–579
- parent involvement model 298, 298–299; estimated using Amos 299–302, 299–303
- partial correlations 36, 605–609, 606–608
- path analysis 257–258, 275–278, 277, 293–294; assumptions in 281–282; basics of 562–563, 563; cautions with 262–264; danger of common causes in 282–289, 283–285, 287, 295n2–3, 563–564; dealing with danger in 291–292; effects of unreliability on results of 336–339, 336–339; equations 279–280n1; equivalent models 315–320, 316–319; exogenous and endogenous variables in 266; first law of 280n2; identification in 265–266; indirect and total effects 271–275, 272, 273–274; interpretation of direct effects 269–271, 270; introduction to 258–266; jargon and notation 264–265, 264–266; latent variables and 437–438, 438; measured and unmeasured variables in 266; more complex example of 266–275; overidentified models 265, 303, 303–312, 304, 305–306, 307, 310–311; paths in the wrong direction 289–290, 289–291; reciprocal causal relations in 291; recursive and nonrecursive models 264, 264–265; as a simple model 258–262, 258–262; steps for conducting 266–269, 267, 269, 292–293; underrepresented minorities (URM) in 286, 287, 290–291, 295n1; unreliability and invalidity in 291; using SEM programs 564–565 (*see also* structural equation modeling (SEM) programs)
- Pearson correlation coefficient 7, 596–597
- peer rejection effects latent variable SEM: initial model 395–398, 395–399; measurement model 393, 394; overview, data, and model 391, 391–393, 392, 393; structural model 393–395, 394
- planned missingness 577–578
- post hoc probing 114–115
- posttest, hot flash *see* hypnosis for hot flashes latent means SEM
- power and sample size 216–221, 217–221, 578–579
- predicted scores and residuals 47–50, 48–49
- prediction: versus explanation 19, 72–73, 104–105, 197; of optimism versus pessimism 227, 227–228
- predictive approach in stepwise regression 98–99
- predictive bias 142–143, 142–144, 160n4; curriculum-based measurement (CBM) 144–146, 144–148, 148; multiple regression steps 148–150

- predictor variables 57–62, 58–61; interpretation 61–62;
 regression results 60–61, 61
- previous grades *See* socioeconomic status (SES), previous
 grades, self-esteem, and locus of control
- PROCESS 167, 181–182, 186–187
- Quantitative Applications in the Social Science* 241
- r* 597–599, 598
- R, multiple regression 31, 31
- R^2 7; adjusted, in stepwise regression 100; direct
 calculation of β and 41–42; importance of, with three
 independent variables 70–72; as measure of effect in
 sequential multiple regression 87–88, 89; sample size
 and power and 216–221, 217–221; and why $R^2 \neq r^2 + r^2$
 44–47, 45–47, 56n1
- random assignment 18, 25n7
- reading comprehension scores *See* invalidity, effects of
 reciprocal causal relations 291
- recursive models 264, 264–265
- regression, curvilinear 168–174, 168–175
- regression analysis, simple bivariate 6–8, 6–9; relation to
 other statistical methods 15–17, 16–17
- regression analysis with dummy variables 112–113, 112–114
- regression coefficients: comparison across samples 38–40,
 39; interpreted using sequential regression 89, 89, 90; in
 multiple regression 32, 32–33, 38–40, 39; standardized,
 in simple regression 14–15; statistical significance of, in
 simple regression 12–13; testing the difference between
 two 63–64, 63–64
- regression diagnostics: diagnosing data problems 208,
 208–216, 210, 211, 213–215; diagnosing violations of
 assumptions 202–208, 203, 205–207
- regression equation: creating a composite in multiple
 regression 54; simple bivariate 9–10, 10
- regression lines: multiple regression 50–51, 51; separate
 243–245, 244; simple regression 10–12, 11–12
- reliability, importance of 335–336, 335–336
- residual invariance 489–490, 491
- residuals: in confirmatory factor analysis 366, 366–368,
 368; normality of 206–208, 207
- root mean square error of approximation (RMSEA)
 311–312; comparing competing models 312–315, 313,
 314; configural invariance and 479
- sample size and power 216–221, 217–221, 578–579
- SAS: predicted scores and residuals in 49; PROCESS
 for 182
- SAT (scholastic aptitude test) test bias 142–143, 142–144
- scaling, criterion 118–119, 119
- self-esteem *See* socioeconomic status (SES), previous
 grades, self-esteem, and locus of control
- semipartial correlations 36, 88–93, 107n1, 609–611,
 609–612
- separate regression lines 243–245, 244
- sequential multiple regression 198; analysis in 81–83, 82;
 block entry 90–92, 91; Cohen's f^2 as measure of effect
 size in 89; comparison with simultaneous regression
 83; estimated total and indirect effects using 272–275,
 273–274; importance of order of entry in 83–86, 84–85;
 interactions and curves 93; interpretation 93, 93–94;
 interpretation of regression coefficients using 89, 89,
 90; problems with R^2 as measure of effect in 87–88,
 89; summary of 94–95; total effects in 86–87, 86–87,
 272–275, 273–274, 280n4; unique variance and 92,
 92–93
- sequential unique regression 92, 92–93
- sequential variance decomposition 89, 89, 90
- SES *See* socioeconomic status (SES), previous grades, self-
 esteem, and locus of control
- sex, achievement, and self-esteem example, categorical
 variables 130–131, 130–132; centering and cross
 products 133–134, 134; statistically significant
 interactions 137–140, 138–139, 140; using multiple
 regression to test for interactions between continuous
 and categorical variables 130–131, 130–132
- simple bivariate regression 4–15; confidence intervals
 13–14, 14; data in 4–6, 5; interpretation 10; regression
 analysis 6–8, 6–9; regression equation 9–10, 10;
 regression line 10–12, 11–12; standardized regression
 coefficient 14–15; statistical significance of regression
 coefficients in 12–13
- simple categorical variables 109–110, 109–110
- simultaneous multiple regression 79–81, 80, 198; analysis
 in 79, 80; comparison with sequential multiple
 regression 83; purpose 80; strengths and weaknesses
 81; what to interpret in 80–81
- single-group, MIMIC models 505–509, 506–508
- single group/dummy variable approach, latent means in
 SEM 456–460, 457–459, 460; compared to MG-MACS
 approach 466–470, 467–469
- single indicators and correlated errors: competing models
 419–421, 420, 421; indirect and total effects 418,
 418–419; latent variable homework model 409–412,
 410, 411; model modifications 421–424, 422, 423;
 unstandardized coefficients 417, 417–418
- slope differences in MLM 245
- Sobel test 181, 185–186
- socioeconomic status (SES), previous grades, self-
 esteem, and locus of control 78, 78, 79; simultaneous
 regression with 79–81, 80
- SPSS: analyzing path models using SPSS 260–262,
 268–275, 325–326; curvilinear regression in 170–172;
 predicted scores and residuals generated in 48, 48–49;
 PROCESS for 182; sequential regression and 89, 90;
 simple bivariate regression line 11–12; testing the
 difference between two regression coefficients 63–64,
 63–64
- standard deviation (SD) 20, 589
- standard errors 4, 592, 593

- standardized regression coefficient 14–15
 standardized residual covariances 366, 366–368, 368
 standardized root mean square residual (SRMR) 311, 326;
 comparing competing models 312–315, 313, 314
 statistics, basic 587–588, 588; ANOVA (*see* analysis of variance (ANOVA)); confidence intervals 4, 13–14, 593–594; correlations 594–597, 595, 597; degrees of freedom 594; distributions 4, 589–592, 590–592; effect sizes 602, 602; mean in 588–589; standard error 4, 592, 593; statistical significance of r 597–599, 598; t tests 3, 4, 594, 599–601, 600; useful formulae for 615–616; variance and standard deviation in 589
- stepwise multiple regression 95, 198–199; adjusted R^2 in 100; alternatives to 101; analysis in 96, 96–97; cross-validation in 99, 107n2; degrees of freedom in 100; as inappropriate for explanation 97–98; lack of generalizability in 101; not necessarily the best predictors in 100; predictive approach 98–99; summary of 101–102; weaknesses 102
- structural equation modeling (SEM) 257–258; analyzing growth data using 531; confirmatory factor analysis (CFA) (*see* confirmatory factor analysis (CFA)); displaying means and intercepts in 445–448, 446–447; estimation of means and intercepts in single group 448–452, 448–452; latent means in (*see* latent means in SEM); latent variable (*see* latent variable SEM); with mean structures 570–572, 571–572; missing values in 453, 574–578, 575; multilevel modeling in 544–559, 545–558, 548; path analysis in (*see* path analysis); sample size, number of parameters, and power in 216–221, 217–221, 578–579
- structural equation modeling (SEM) programs 296–298, 298, 328–330; advantages of 303, 303–312, 304, 305–306, 307, 310–311; Amos and Mplus 297–298, 298, 557–558, 557–559, 580; basics of 297–298, 298; comparing competing models in 312–315, 313, 314; correlations and covariances in 306, 306–307; directionality and 320–321, 320–321; equivalent models 315–320, 316–319; estimating parent involvement model via 299–302, 299–303; longitudinal models 323–324, 324–325; measures of fit in 326–328, 327; model fit and degrees of freedom in 307, 307–309; more complex models 315–324, 316–325; multiple regression versus 325–326; nonrecursive models 322–323, 322–323; other measures of fit 309–312, 310; overidentified models 303, 303–312, 304, 305–306, 307, 310–311; reanalysis of parent involvement path model using 298, 298–303, 299–302
- structural invariance 491–493, 492
- structural model 345
- symbols 613–614
- test bias 142–146, 142–148, 148; curriculum-based measurement (CBM) 144–146, 144–148, 148
- theory and path analysis 262
- theory trimming 293
- three-factor combined nonverbal model 360–363, 361, 362
- time precedence and path analysis 263
- total effects: in higher-order models 372–373, 373; in latent variable SEM 398, 398–399; in sequential multiple regression 86–87, 86–87, 272–275, 273–274, 280n4; single indicators and correlated errors model 418, 418–419
- Trait-Treatment Interactions (TTIs) *see* Aptitude-Treatment Interactions (ATIs)
- true experiments and common causes 287, 287–288
- t tests 3, 3, 4, 594, 599–601, 600; ANOVA consistency with 601–602; confidence intervals and 13–14, 14
- Tucker-Lewis index (TLI) 309, 312, 327; comparing competing models 312
- unconditional, simple growth model 517–518, 517–523, 520–524
- unconditional model, MLM 247, 247
- underidentified models 264, 265
- underrepresented minority students (URM) 286, 287, 290–291, 295n1
- unequal group sizes 120–123, 120–125, 123–124, 125
- unique variance in sequential regression 92, 92–93
- unreliability: effects of 335–339, 335–339; effects on path results 336–339, 336–339; and the importance of reliability 335–336, 335–336; and invalidity in path analysis 291
- unstandardized coefficients 417, 417–418
- uses of regression diagnostics for data problem 212–213
- validity, meaning and importance of 339–340, 340
- variables: categorical (*see* categorical variables); continuous (*See* continuous variables); dependent (*See* dependent variables); exogenous and endogenous, in SEM 266; independent (*see* independent variables); latent, in SEM (*See* latent variable SEM); measured and unmeasured, in SEM 266; predictor 57–62, 58–61; socioeconomic status (SES), previous grades, self-esteem, and locus of control (*see* socioeconomic status (SES), previous grades, self-esteem, and locus of control); in stepwise multiple regression 97; using multiple regression to test for interactions between continuous and categorical 132, 132–137, 134–136, 136
- variance 20, 589, 614; explaining 17–18; sequential unique 92, 92–93
- variance/covariance matrix of measured variables 495–499, 496, 496–498
- variance partitioning 89, 89, 90
- Venn diagram of shared variance 45, 45–46; on order of entry 85, 85–86
- veracity of models and causality 580–581
- verbal skills and memory strategies 150–151, 151–152
- χ^2 309, 311, 312–315, 326, 327
- z values 368–369